YOU LIVER AND YOU LEARN: MACHINE LEARNING TO RANK WAIT-LISTED LIVER TRANSPLANT CANDIDATES

Kayla Cummings
kaylac@mit.edu

Sam Gilmour
sgilmour@mit.edu

Josh Wilde
jtwilde@mit.edu

1 Introduction

There are almost 13,000 patients on the waitlist for a liver transplant in the United States [1]. The Organ Procurement and Transplantation Network (OPTN) is in charge of allocating livers as they arrive to wait-listed candidates. The ongoing objective is to implement the liver allocation policy that saves the maximum number of lives with the highest possible quality of life over the lifespan of the system.

Consensus among doctors is that a fair policy should prioritize patients whose conditions would deteriorate most quickly without a transplant. Thus, transplant offers are currently made according to a greedy assignment algorithm based on patient scores and patient/liver geographic locations. Patient scores are calculated by the Model for End-Stage Liver Disease (MELD). MELD is a logistic regression model over individual patients' features predicting the likelihood that they will either die or become medically unfit for transplant within 3 months. A higher likelihood translates into a higher priority on the waitlist. The Optimized Prediction of Mortality (OPOM) model, developed by Bertsimas *et al.*, used optimization and machine learning to greatly improve the accuracy of this 3-month mortality prediction. Their optimal classification tree model achieves the improvement with minimal sacrifice to the interpretability of the ranking model, an important factor for the medical community.

Our project explores the impact of machine learning on the process of developing a candidate ranking policy. We first measure the impact of model interpretability on the accuracy of pre-transplant disease severity measurements. Our results establish that OPOM does not sacrifice prediction quality by using an interpretable method, according to the 3-month mortality metric. Thus, we re-orient our objective toward optimizing over system-wide transplant outcomes. This pivot challenges the current paradigm that a fair policy is solely a function of patients' pre-transplant vitality, and explores the potential utility of a ranking policy based on *projected* post-transplant outcomes.

Our core model uses machine learning and optimization to build a priority map based on organ acceptance likelihood and projected post-transplant survival, as well as constraints on fair allocation. We iterate with a model that generates a static priority map toward evaluating the performance of a dynamic policy that changes according to the patients on the waitlist. Finally, we build a model that simulates liver allocation and survival outcomes to benchmark the performance of our policies against existing ranking policies. Simulations show that our policies result in approximately 1% more transplants and consequently 1% fewer pre-transplant deaths, with minimal effect on post-transplant morbidity and with continued fair allocation for targeted demographics. Furthermore, our dynamic policy performs at par with the static policy, indicating that its additional computational overhead may be unnecessary.

Section 2 describes our dataset. Section 3 reviews our machine learning and optimization models, whose results are presented in section 4 and discussed in section 5.

2 Data

The OPTN Standard Transplant Analysis and Research (STAR) liver transplant dataset contains medical information on deceased and living donors, transplant candidate status updates, and opera-

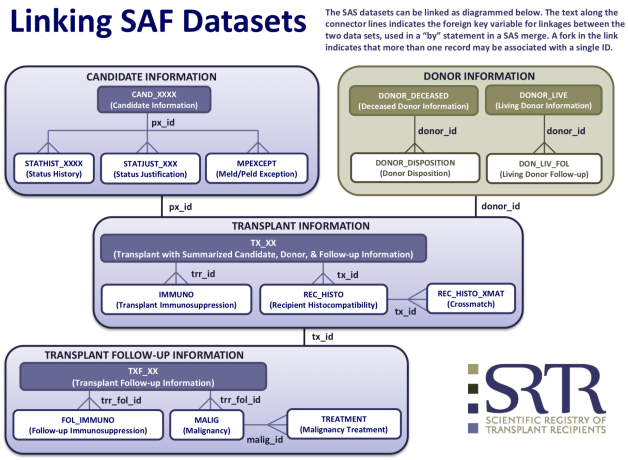


Figure 1: SRTR dataset file linking diagram [2].

tions performed between January 2002 and September 2016. Figure 1 visualizes how the records are linked. The portions of the data that we used can be partitioned into the following sub-datasets.

- **Wait-listed candidates** (267K records). Each record describes an individual's health history, demographics, and location. There are also features that are helpful in determining a patient's compatibility with a potential donor.
- **Wait-listed candidates' status updates** (2.4M records). Each timestamped observation represents a candidate's wellness check, which includes up-to-date health measurements, waitlist status, and MELD score.
- **Organ donors** (334K records). Each record is associated with all organ donors recovered by SRTR, living or deceased. Approximately 144K records correspond to liver donors. Each record includes features on known donor health history, demographics, organ status, compatibility-related information, and organ recovery date.
- **Transplants** (144K records). Rows correspond to realized liver transplants. They specify the candidate, patient, operating hospital, and ultimate patient status.

We also made use of the dataset used in the Liver Simulated Allocation Model (LSAM), which is the simulation tool developed in [2] and used by policy-makers to evaluate allocation policies, when constructing the policies described in Section 3.2. This data has been sampled from a generative model fit to the STAR dataset, but also contained features that are used in the acceptance and survival models (to be described in Section 3.2) which were not readily obtainable from the STAR dataset.

This dataset contained a total of 69,680 patient arrivals and 34,160 organs becoming available for transplant, and was divided according to a 70%/30% split for training and testing, respectively.

3 Models

As described in Section 1, the tasks to be addressed by our models were split into two categories:

1. Predicting 3-month patient mortality. This is the same task addressed by MELD and OPOM, and our approach is described in Section 3.1.
2. Constructing policies which produce ranking scores for transplant offers that depend on both

patient features, and features of the organ being offered. Our approaches are described in Section 3.2.

3.1 Pre-transplant mortality objective

This task was a simple supervised binary classification problem. Given the STAR dataset of patient status history updates (each of which, recall, correspond to a vector of measurements taken at a specific point in time), the dependent variable was calculated as $y_i = \mathbb{1}\{\text{patient in update } i \text{ removed from waitlist within next 3 mo}$

We tested two different feature maps for the input space:

1. $\phi_1(\mathbf{x})$: feature map used in the development of OPOM, composed of medical readings used in the calculation of MELD in addition to differences between current and most recent readings.
2. $\phi_2(\mathbf{x})$: the above $\phi_1(\mathbf{x})$, augmented with additional features derived from patient history (including maximum and minimum values of chemical readings).

Our aim in this section was to investigate whether OPOM and MELD, with their inherently more interpretable structures, sacrifice any ability to predict 3-month mortality when compared with more complex models. We tested a series of neural network models whose structures are summarised in Table 1.

Property	Description
Connectivity pattern	Dense
Number of layers	Between 1 and 5
Neurons per layer	Between 100 and 500
Activation function	ReLU

Table 1: Summary of neural network architectures tested.

All networks were trained, validated and tested with the Adam optimisation algorithm and learning rate of 0.001 on a 70%/15%/15% split of the STAR dataset of patient status history updates.

3.2 Post-transplant acceptance/survival objective

3.2.1 Fairness-adjusted weight calculation

In any greedy allocation mechanism, one interpretation of the scores used to rank the recipients of the resource in question is as the benefit obtained by the system as a result of making that matching.

Prioritising transplants based on severity of disease is one way of measuring benefit, though this approach is used as a proxy for maximising the number of extra years lived by transplant recipients in a fair way. Given that survival and acceptance models are available, it is natural to consider more directly optimising the objective of improving the aggregate life years gained from transplantation.

This required a policy which takes into account both patient and organ features – and, returning to the interpretation of scores as a measure of benefit to the system, estimates the extra years lived by a transplant recipient due to an offer.

This information is provided by the survival and acceptance models. The obvious issue with taking this approach is that it does not consider any notion of transplant fairness across demographics. Scoring based on disease severity more naturally accounts for this, assuming a relatively uniform distribution over these demographics. If only extra life years from transplant were taken into account, then we could imagine a situation where transplants for young recipients are unfairly prioritised.

In order to obtain fairness-adjusted weights that measure the benefit obtained by the system upon a transplant being offered, we used an approach based on work by Bertsimas *et al.* [3]. This first requires us to estimate the benefit to the system obtained by offering organ j to patient i (denoted c_{ij}) by making use of the survival and acceptance models internal to LSAM.

The survival model is based on the Cox proportional hazards model [2], which is commonly used in medical applications for predicting survival times as explained by a set of factors. The inputs include patient age, medical readings, and life support dependence, as well as donor age and disease history – and the output is a randomly sampled number of years the patient is predicted to survive post-transplant.

The acceptance model was constructed as a simple LASSO regression on binary classification organ acceptance data taken from 2011, to produce an acceptance probability q_{ij} [2]. It includes a 14-term model for patients who are in urgent need of a transplant, and a 50-term model for those who are less urgently in need. Each model makes use of both patient and organ features.

Let S_{ij} be a random variable representing the number of days that patient i survives if they receive liver j. Also let A_{ij} be a random variable representing whether patient i accepts liver j for transplant upon receiving the offer.

Then a reasonable estimate of the benefit obtained by the system if organ j is offered to patient i is $c_{ij} = \mathbb{E}\left[A_{ij} \cdot S_{ij}\right]$. We assume that organ acceptance and post-transplant survival are independent processes. Also note that $A_{ij} \sim \text{Bern}(q_{ij})$, and so $c_{ij} = q_{ij} \cdot \mathbb{E}[S_{ij}]$. The expectation over S_{ij} was computed from an empirical mean taken over 100 samples for each patient i and organ j.

Given these estimates, we construct a matching problem with linear fairness constraints through which we can bound the proportions of matches made within certain demographics. The binary constraints on variables $\mathbf{x}$ are relaxed so that the problem is linear, and we can form a meaningful dual:

$$
\begin{aligned}
\max_{\mathbf{x}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} \\
\text{s.t.} \quad & \sum_{j=1}^{n} x_{ij} \leq 1 \qquad \forall i \in [m] \\
& \sum_{i=1}^{m} x_{ij} \leq 1 \qquad \forall j \in [n] \\
& \mathbf{A}\mathbf{x} \leq \mathbf{b} \\
& \mathbf{x} \geq 0
\end{aligned}
$$

Solving this problem provides us with optimal $(\mathbf{x}^*, \mathbf{p}^*)$, where $\mathbf{p}^*$ correspond to the optimal dual variables associated with the fairness constraints. The problem is then equivalent to the following pure matching problem:

$$
\begin{aligned}
\max_{\mathbf{x}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} (c_{ij} - A_{ij}^{\intercal}\mathbf{p}^*)x_{ij} \\
\text{s.t.} \quad & \sum_{j=1}^{n} x_{ij} \leq 1 \qquad \forall i \in [m] \\
& \sum_{i=1}^{m} x_{ij} \leq 1 \qquad \forall j \in [n] \\
& \mathbf{x} \geq 0
\end{aligned}
$$

which allows us to interpret each $\bar{c}_{ij} \coloneqq c_{ij} - A_{ij}^{\intercal}\mathbf{p}^*$ as fairness-adjusted weights.

The linear fairness constraints used in our model were taken from the key demographic groups reported in [4], and bounds were based on the prevalence of wait-listed patients with those characteristics in the LSAM training dataset. A selection are provided in Table 2, where percentages indicate bounds on the proportions of matches made, and values indicate bounds on the mean of that property across all matches made.

	Demographic	**Lower Bound**	**Upper Bound**
Gender	Male	52%	100%
	Male	28%	100%
Race	White	55%	100%
	Black	9.1%	100%
	Hispanic	11%	100%
Blood Type	O	36%	100%
	A	28%	100%
General	Age	35	100
	Days wait-listed	0	200

Table 2: Selection of linear fairness constraints included in matching model.

3.2.2 Static priority map

Having provided a method for estimating weights which represent the benefit to the system obtained by making a transplant offer (adjusted for fairness), a static priority map requires a method for which these benefits can be estimated for pairs of patient and organ which are newly observed.

These were obtained as the solution to a standard OLS regression problem, with the datasets defined as follows from the output of the previously-described matching problem:

$$\mathbf{X} = \begin{bmatrix} \mathbf{p}_1^{\intercal} & \mathbf{o}_1^{\intercal} \\ \mathbf{p}_2^{\intercal} & \mathbf{o}_1^{\intercal} \\ \vdots & \vdots \\ \mathbf{p}_1^{\intercal} & \mathbf{o}_2^{\intercal} \\ \vdots & \vdots \\ \mathbf{p}_m^{\intercal} & \mathbf{o}_n^{\intercal} \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} \bar{c}_{11} \\ \bar{c}_{12} \\ \vdots \\ \bar{c}_{21} \\ \vdots \\ \bar{c}_{mn} \end{bmatrix}$$

Note that $\mathbf{p}_i$ and $\mathbf{o}_j$ represent the feature vectors for patient i and organ j, respectively.

3.2.3 Dynamic priority map generator

The static priority map relies on a constant relationship between patient and organ features and the fairness-adjusted weights $\bar{c}$. We hypothesize that $\bar{c}$ depends on the characteristics of the current patient waitlist and the organs available for transplant. For example, consider two patient waitlists $\mathcal{W}_1$ and $\mathcal{W}_2$ with a common patient p. $\mathcal{W}_1$ contains patients such that all feasible patient-organ matches satisfy the fairness constraints, while $\mathcal{W}_2$ contains some patients from a disadvantaged class with relatively mild clinical transplant needs. Patient p's priority score on $\mathcal{W}_1$ will depend only on his or her clinical match with available livers, while on $\mathcal{W}_2$ patient p's priority will depend on the fairness of the optimal matching solution as well.

Our strategy is to model this dependence as $\bar{c}_i = f_i(z)$, where f_i is a function from batch-level organ and patient features $z \in \mathbb{R}^b$ to one fairness-adjusted weight $\bar{c}_i$ in the static priority map. We divide a training set of patients and organs randomly into B batches and assume that each f_i takes

the form of a linear regression parameterized by a vector $\theta^i \in \mathbb{R}^b$. Our model therefore becomes $\bar{c}_i = (\theta^i)^T z$. For simplicity we form z^k for $k \in B$ by taking the mean of all numeric patient and organ features. Collectively, the learned parameters form a dynamic policy map generator that yields fairness-adjusted weights for a given batch of patients and organs.

3.3 SimSAM: Liver allocation simulator

Policymakers typically evaluate new ranking policies using the Liver Simulated Allocation Model (LSAM), software developed by the Scientific Registry of Transplant Recipients (SRTR) to simulate liver allocation outcomes for wait-listed candidates [2]. To simulate our dynamic policy, we needed the ability to change the ranking method based on the current waitlist and offered organ. This feature was not available in the LSAM software. Thus, we developed our own Simple Simulated Allocation Model (SimSAM). SimSAM is an adaptation of code by Theodore Papalexopoulos, who replicated LSAM in Python to improve its performance for a recent paper [5]. The module takes a ranking policy as input and returns the final status of each patient after allocation; each patient either dies pre-transplant on the waitlist, survives a specified number of days post-transplant, or remains in the same condition.

The data used to carry out each simulation was the same data that LSAM uses as input (described in Section 2). Each patient is associated with a removal date, which we interpret as the natural time of death in the absence of a transplant operation. Due to limited modeling capabilities for the progression of liver-related afflictions, SimSAM assumes that patient condition does not evolve until the time of transplant or death. We discuss this limitation in Section 5.

SimSAM generates organ arrival times uniformly at random between January 1, 2007 and December 31, 2011. The livers are offered in sequential order and allocated one at a time. For each organ, SimSAM filters patients who have arrived to the waitlist, have not died of natural causes, have not received a transplant, and are biologically compatible with the liver. Rankings are computed for this subset of patients according to the input policy. Then the organ is offered sequentially according to each qualified patient's priority, and it is assigned to the first candidate that accepts according to the acceptance model. In rare cases, the organ is not accepted by any patient. After a transplant is performed, the survival model is queried to affiliate an outcome with the transplant recipient.

4 Results

4.1 Model performance

4.1.1 Pre-transplant mortality objective

Our neural network models using feature maps ϕ_1 and ϕ_2 both returned test set AUCs of 0.86, identical to the AUCs reported in [4]. We also confirmed that an optimal classification tree run on the augmented feature set ϕ_2 did not improve performance over the original OPOM results.

4.1.2 Post-transplant acceptance/survival objective

For computational reasons, the static priority maps had to be fit on 45 batches of the available LSAM input data, and aggregated (since querying the survival and acceptance models to find the weights for the matching problem could not be reasonably completed for the entire matrix in the training set).

Fitting these maps on each batch revealed that the model performance did not extend to the 'out-of-sample' batches in the training set, with a 95% confidence interval for the mean difference in MSE between 'out-of-sample' and 'in-sample' being $[570, 1368]$. This observation informed our decision to pursue the idea of dynamic priority maps.

To evaluate the fit of the dynamic priority map generator regressions, we set aside 30% of the patients and organs as a test set and train the dynamic priority map regressions on the remainder of the data.

We evaluate the mean squared error of each of the regressions on the held out test set, standardizing the MSE by the variance of the target fairness-adjusted weights for comparability. Figure 2 shows that the average MSE for these models is larger than the variance of the weights by 17%, indicating that there is significant room for improvement in the dynamic priority map models.

	MSE divided by target variance
Minimum	0.75
Mean	1.17
Maximum	1.47

Figure 2: Mean squared errors of dynamic priority weight generator regressions, standardized by the variance of the target fairness-adjusted weight. The dynamic generator fits a regression for each fairness-adjusted weight, so the statistics shown are taken over the set of regressions.

4.2 Simulation outcomes

We focus on the following key metrics to compare the success of the static and dynamic priority maps to MELD and OPOM using SimSAM:

1. Number of accepted transplants
2. Number of pre-transplant patient deaths
3. Three-month post-transplant survival rate
4. Fraction of patients from disadvantaged classes receiving transplants

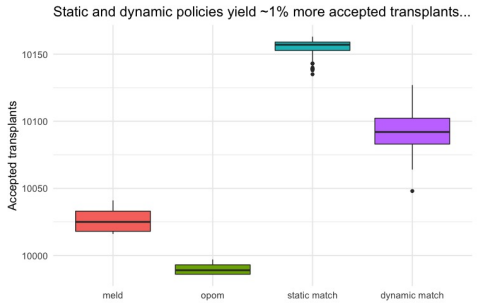
Metrics 1 and 2 measure the ability for the priority maps to generate the maximum number of successful organ transplants. Metric 3 evaluates the post-transplant patient mortality using the standard three-month time frame cited in [4]. Finally, since our policies directly incorporate patient demographics to learn fairness-adjusted weights, there is a risk that these policies introduce unintended bias into the priority maps. We measure the relative fairness of the four policies using metric 4.

Figure 3 shows the relative performance of the four policies based on these metrics. Our results indicate that the static priority map results in approximately 1% more accepted transplants and 1% fewer pre-transplant deaths compared to MELD and OPOM. The dynamic priority map also results in more accepted transplants and fewer pre-transplant waitlist deaths, though the static policy performance is slightly superior. Three-month post-transplant mortality rises slightly for both of our new priority map policies, but difference is smaller than the decrease in waitlist deaths. Importantly, our priority maps are able to achieve these improvements without any negative impact on our selected fairness metrics. Figure 3d indicates that there is no change in the fraction of patients from disadvantaged classes that receive transplants compared to MELD and OPOM.

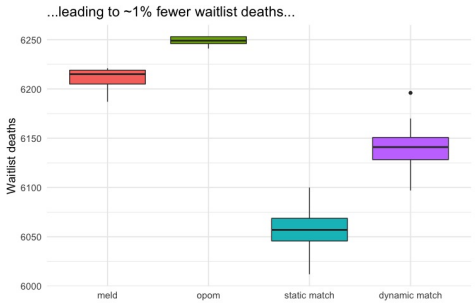
5 Conclusions

Since a less interpretable model with additional feature engineering failed to improve pre-transplant mortality prediction accuracy over OPOM, we conclude that OPOM comes close to approximating the Bayes Error when only considering patient features. Encouragingly, this means that the price of interpetability is near zero for pre-tranplant mortality prediction.

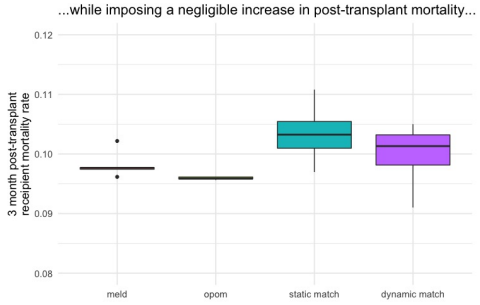
However, our SimSAM results indicate that it is possible to improve patient-liver matches by training policies directly on projected outcomes for organ acceptance and post-transplant survival rather than the intermediate pre-transplant mortality objective. Our priority maps increase simulated accepted transplants and decrease the number of patients that die on the transplant waitlist by approximately



(a)



(b)



(c)

...and continuing to allocate organs fairly

	MELD	OPOM	Static Match	Dynamic Match
Race Non-White	0.30	0.30	0.30	0.30
Female	0.37	0.37	0.37	0.37
Elderly (65+)	0.12	0.12	0.12	0.12
Waitlisted 1+ years	0.12	0.12	0.10	0.13
Acute Hepatic Necrosis	0.05	0.04	0.05	0.05
Cirrhosis	0.08	0.08	0.08	0.08
Malignant Neoplasms	0.09	0.09	0.09	0.09
Liver disease	0.34	0.34	0.34	0.34
Obese (BMI > 30)	0.34	0.34	0.34	0.34
Blood Type O	0.47	0.47	0.47	0.47

(d)

Figure 3: Key simulation results from SimSAM.

1% without meaningfully increasing the rate of post-transplant mortality. In addition, the fairness constraints embedded in the fairness-adjusted weight calculation effectively maintain fair organ allocation within targeted demographics.

We acknowledge the limitations of SimSAM as a policy evaluation tool. Our initial effort to build an allocation simulator that accommodates our new priority maps ignores the dynamics of patient conditions while on the waitlist as well as some geographical constraints. In addition, we omit an explicit tiebreaking mechanism for patients with equal priority scores.

While both the static and dynamic priority maps exhibit fair and efficient outcomes, the dynamic priority map does not demonstrate a meaningful improvement over the simpler static policy. Our modeling results in section 4.1 indicate that aggregating batch-level features through averaging and applying a linear regression does not adequately recover the optimal parameters of the static priority map. Future work could explore whether more complex feature engineering and a nonlinear model structure would improve prediction accuracy. In addition, we have maintained the current greedy nature of organ allocation in our new policies. In a reinforcement learning environment, a policy could conceivably improve outcomes by offering an organ to a healthier patient who is difficult to match rather than a less healthy person who is more likely to match with a future organ. Of course, such a policy would introduce new ethical concerns.

Despite the positive fairness outcomes for our new policies, any policy that changes priority scores for wait-listed patients based on the composition of the waitlist and the features of available organs invites ethical questions. For example, a patient marked high priority for an organ offer could subsequently be marked as lower priority for the next organ due to new patient arrivals who are more likely to achieve a successful transplant. If such a policy resulted in greatly improved mortality outcomes without disproportionately advantaging certain groups over others, policymakers may be willing to absorb the risk of such an ethically grey circumstance. Our proposed policies likely do not accomplish this type of significant improvement, but future work could elicit increased improvement.

6 Acknowledgements

The authors acknowledge Theodore Papalexopoulos and Yuchen Wang, whose code and advice improved the outcomes of this project.

7 Division of labor

Josh trained benchmark models, helped formulate and implement learning tasks, engineered additional features, and visualized results. Sam pre-processed data, helped formulate and implement learning tasks, helped format SimSAM inputs, sped up computation, and led poster design. Kayla pre-processed data, helped formulate the learning tasks, helped format matching model inputs, built SimSAM, generated simulation outcomes, and outlined the report. All contributed to writing the report and the poster.

References

[1] Health Resources and Services Administration, US Department of Health and Human Services, "OPTN Data," December 2019. Accessed 7 Dec. 2019 at `https://optn.transplant.hrsa.gov/data/`.

[2] S. R. of Transplant Recipients, *Liver Simulated Allocation Model*, 2014. Accessed 2 Oct. 2019.

[3] D. Bertsimas, V. F. Farias, and N. Trichakis, "Fairness, efficiency, and flexibility in organ allocation for kidney transplantation," *Operations Research*, vol. 61, no. 1, pp. 73–87, 2013.

[4] 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.

[5] D. Bertsimas, T. Papalexopoulos, N. Trichakis, Y. Wang, R. Hirose, and P. Vagefi, "Balancing efficiency and fairness in liver transplant access: tradeoff curves for the assessment of organ distribution policies," *Transplantation*, vol. Online First, Oct 2019.