# Closing the loop in medical decision support by understanding clinical decision-making: A case study on organ transplantation

**Yuchao Qin**[*]
University of Cambridge
yq257@cam.ac.uk

**Fergus Imrie**[*]
University of California, Los Angeles
imrie@g.ucla.edu

**Alihan Hüyük**
University of Cambridge
ah2075@cam.ac.uk

**Daniel Jarrett**
University of Cambridge
daniel.jarrett@maths.cam.ac.uk

**Alexander Edward Gimson**
University of Cambridge
alexander.gimson@nhs.net

**Mihaela van der Schaar**
University of Cambridge
The Alan Turing Institute
University of California, Los Angeles
mv472@cam.ac.uk

## Abstract

Significant effort has been placed on developing decision support tools to improve patient care. However, drivers of real-world clinical decisions in complex medical scenarios are not yet well-understood, resulting in substantial gaps between these tools and practical applications. In light of this, we highlight that more attention on understanding clinical decision-making is required both to elucidate current clinical practices and to enable effective human-machine interactions. This is imperative in high-stakes scenarios with scarce available resources. Using organ transplantation as a case study, we formalize the desiderata of methods for understanding clinical decision-making. We show that most existing machine learning methods are insufficient to meet these requirements and propose iTransplant, a novel data-driven framework to learn the factors affecting decisions on organ offers in an instance-wise fashion directly from clinical data, as a possible solution. Through experiments on real-world liver transplantation data from OPTN, we demonstrate the use of iTransplant to: (1) discover which criteria are most important to clinicians for organ offer acceptance; (2) identify patient-specific organ preferences of clinicians allowing automatic patient stratification; and (3) explore variations in transplantation practices between different transplant centers. Finally, we emphasize that the insights gained by iTransplant can be used to inform the development of future decision support tools.

## 1 Introduction

The decision that a patient and a clinician jointly make when a donor organ is offered for transplantation is critical and carries serious consequences. Even though the initial offer of a donor organ for a particular recipient is made on the basis of agreed criteria as to optimum allocation

[*]Equal contribution

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

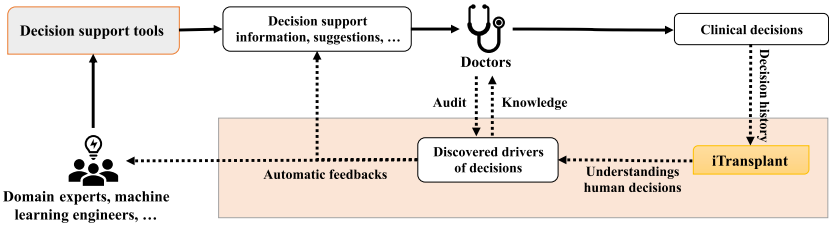



Figure 1: Closing the loop of medical decision support by understanding clinical decision-making. Most existing clinical decision support tools offer generic advice or alerts to clinicians without specificity for the immediate decision about to be taken. In this paper, we highlight that iTransplant, by identifying drivers of that decision, will be able to feedback into future iterations of the decision support tool specific information about which factors impacted the decisions so that in future that specific information can be given to the decision maker or can be taken into account in designing future iterations of the decision support tool itself.

within each transplant jurisdiction, there remains substantial variation in the rate at which organs are accepted. Clinical variation is a well-observed phenomenon, but may have profound consequences and unfavourably impact outcomes. Addressing clinical variation is therefore a major priority in many healthcare systems. Understanding the factors which are associated with variation in clinical decision-making is an important goal as it might be able to inform clinicians of biases which, if rectified, might result in improved clinical outcomes. With a case study on organ transplantation, we explore the potential of inverse decision-making approaches to shed light on such factors.

**Organ transplantation** Transplantation is typically the last life-saving treatment available for patients with end-stage diseases that cause organ failures. However, due to the limited availability of donors, patients often have to wait years before transplantation [14, 23]. Regrettably, waitlists continue to grow despite increases in the number of donors and many patients die while waiting for an organ, with over 7,500 deaths each year in the United States alone [20]. The majority of these patients received at least one organ offer that was declined on their behalf [12], despite these organs often appearing to be suitable for transplantation and yielding good outcomes when eventually transplanted [17]. It is therefore important to understand why donor organs subsequently successfully implanted have been declined for previous patients.

**Clinicians' decision-making is poorly understood** When a donor organ becomes available, it is first offered to a patient on the waitlist on the basis of agreed offering criteria. Once an organ is offered, a clinician must choose whether to accept or decline the organ offer. Although significant effort has been placed into developing organ allocation algorithms [47, 28, 4], a donor offered to the first ranked patient in a waitlist is rejected up to 50% of the time [12, 43].

In real-world organ transplantation systems, the performance of organ allocation algorithms are significantly affected by clinicians' assessments of organ offers. Even for good quality organs, the assigned organ offers could be turned down several times before they are finally accepted [45, 17]. The high ratio of declined organ offers is important as it may impact outcomes for that organ (e.g. due to prolonged cold ischemia time) and the patients involved (e.g. [12] shows that centers with higher acceptance rates experienced significantly lower adjusted estimated waitlist mortality of the highest-ranked patients).

**Substantial variation in clinical practice** Variation in clinical practice is an extensively studied phenomenon across medicine with significant impacts on organ transplantation [42, 2]. Striking discrepancies in organ offer acceptance rates have been observed for different transplant centers, even after accounting for organ quality and the severity of the recipient's illness [12]. However, it has been impossible to disambiguate the causes of this variation, despite its importance for understanding current medical practices and ultimately improving organ allocation policy.

In this paper, we highlight that to address such challenges it is necessary for practical inverse decision-making approaches to provide interpretable and personalized insights into clinical decision-making.

**Human interpretable policies** Interpretability and transparency are crucial for machine learning applications in medicine [39]. As a result, white-box models, such as logistic regression, are widely adopted in the medical literature (e.g. [33, 10]). Numerous studies have demonstrated

improved performance from using black-box models for medical applications [15], albeit at the cost of interpretability. The development of interpretable, yet highly performant, models for clinical decision-making is essential to bridge the gap to black-box models and further medical knowledge. In addition, for such interpretations to be useful for understanding clinical decision-making, models must be counterfactually consistent (i.e. the counterfactual prediction must match the interpretation).

**Precision medicine** Precision, or personalized, medicine seeks to improve medical care by tailoring therapy to the needs of a particular patient [40]. However, in the organ transplantation setting, most existing methods only learn a global decision-making policy for all patients, which fails to address the discrepancies in clinicians' policies for patients from different cohorts (see, e.g., [16, 5, 35]). To both understand clinicians' decision-making and improve precision medicine, models that learn personalized policies are necessary.

**Our contributions** In this paper, we propose *iTransplant (individualized TRANSparent Policy Learning for orgAN Transplantation)*, a novel data-driven framework to learn interpretable organ offer acceptance policies directly from clinical data. Our method learns a patient-wise parametrization of the expert clinician policy that accounts for the differences between patients, a crucial but often overlooked factor in organ transplantation.

We achieve this by training a neural network-based policy selector to identify individualized policies for patients from different cohorts. These policies act on the space of known match criteria using a white-box function, ensuring interpretability with respect to the match criteria. Our method significantly outperforms existing interpretable models, with comparable accuracy to black-box approaches.

We conduct several investigative experiments with real-world liver transplantation data from the Organ Procurement and Transplantation Network (OPTN), covering 190,525 organ offers. The results show that iTransplant can be used to probe clinical decision-making practices in a number of ways. Our investigations allow us to: (1) identify important match criteria for organ offer acceptance; (2) discover patient-wise organ preferences of clinicians via automatic patient stratification in a latent representation space; and (3) examine the transplantation practice variations across transplant centers.

## 2 Problem Formulation

**Notation** We denote the feature space of all possible patients as $\mathcal{X} \subseteq \mathbb{R}^d$ and the feature space of all possible organs available as $\mathcal{O} \subseteq \mathbb{R}^e$. The organ offer that assigns organ $\mathbf{O} \in \mathcal{O}$ to patient $\mathbf{X} \in \mathcal{X}$ is denoted with $\mathbf{s} = (\mathbf{X}, \mathbf{O})$ and the associated decision of clinicians is denoted as $a \in \{0, 1\}$. Here, the event of offer acceptance is denoted as $\{a = 1\}$, and $\{a = 0\}$ means offer rejection. The decision-making policy of clinicians (expert policy) is assumed to be a probability distribution $\pi^*(a|\mathbf{s})$ conditioned on the organ offer $\mathbf{s}$.

Following the discussion on interpretability and precision medicine in Section 1, we propose three key desiderata of practical inverse decision-making approaches: 1) personalized policies, 2) interpretable insights of decisions, and 3) consistent interpretations under perturbation. First, let us introduce some concepts and assumptions related to the above requirements.

**Criteria space** Suppose there are $l$ known match criteria related to decisions on organ offers, denoted by a set of functions $\mathcal{C} = \{c_i(\mathbf{s})\colon \mathcal{X} \times \mathcal{O} \to \mathbb{R}, i = 1, 2, \ldots, l\}$. Each criterion $c_i \in \mathcal{C}$ takes patient and organ features as input and generates a match score as the assessment of organ offer $\mathbf{s} = (\mathbf{X}, \mathbf{O})$. Based on $l$ match criteria in $\mathcal{C}$, we define a transform $\mathcal{T}\colon \mathcal{X} \times \mathcal{O} \to \mathcal{M} \subseteq \mathbb{R}^l$ that maps organ offers to a criteria space $\mathcal{M}$. Given an organ offer $\mathbf{s}$, its representation in space $\mathcal{M}$ can be calculated as $\mathcal{T}(\mathbf{s}) \coloneqq [m_1(\mathbf{s}), m_2(\mathbf{s}), \ldots, m_l(\mathbf{s})]'$. In practice, the criteria space would usually be proposed by experts based on domain knowledge. Note, our method allows clinicians to explore criteria space $\mathcal{M}$ with different sets of criteria based on their own expertise.

In line with our desideratum for a personalized policy, we have Assumption 1 that bridges the criteria space $\mathcal{M}$ with the decision-making policy of clinicians on a per patient basis.

**Assumption 1** (Partial monotonicity) There exists a vector $\mathbf{v} \in \{-1, 1\}^l$ such that for any two organ offers $\mathbf{s}_1 = (\mathbf{X}_1, \mathbf{O}_1), \mathbf{s}_2 = (\mathbf{X}_2, \mathbf{O}_2)$ satisfying the conditions of $\mathbf{X}_1 = \mathbf{X}_2$ and $\mathbf{v} \circ \mathcal{T}(\mathbf{s}_1) \succeq \mathbf{v} \circ \mathcal{T}(\mathbf{s}_2)$, where $\circ$ is the Hadamard product operator, we have $\pi^*(a = 1|\mathbf{s}_1) \geq \pi^*(a = 1|\mathbf{s}_2)$, where $\pi^*$ is clinicians' decision-making policy on organ offers.

**Assumption 2** (Greedy decision policy) Noting that decisions on organ offers by clinicians will not necessarily affect the next organ offer assigned to their patients, we further assume that decisions of clinicians are consistent with a greedy policy, i.e., maximizing the immediate benefit $R(\mathbf{s}, a)$ of decision $a$ on organ offer $\mathbf{s}$ with policy $\pi^*(a|\mathbf{s}) = \arg\max_{a\in\{0,1\}} R(\mathbf{s}, a)$.

Related to the requirement for a personalized policy, clinicians may use different subsets of criteria in $\mathcal{M}$ to evaluate potential outcomes of the offered organ for different cohorts of patients. For instance, [5] reports that the difference in age of donor and recipient has diverse impacts on post-transplant mortality, and young recipients with elderly donors are most affected. Hence, a personalized reward structure is necessary to account for policy variations at the patient level, which leads to the following assumption:

**Assumption 3** (Personalized rewards) Similar to existing literature on inverse decision-making (e.g., [49, 30]), we assume a linear structure of the reward function: $R(\mathbf{s}, a) = \langle \rho_a^*(\mathbf{X}), \mathcal{T}(\mathbf{s})\rangle$, where $\langle\cdot,\cdot\rangle$ is the inner product of two vectors in an Euclidean space, $\rho_a^*(\mathbf{X})$ is the weight vector for different criteria in space $\mathcal{M}$. Note that there exists a family of equivalence reward functions $\mathcal{R}(\mathbf{s}, a) = R(\mathbf{s}, a) - \Phi(\mathbf{s})$, where $\Phi(\mathbf{s}) : \mathcal{X} \times \mathcal{O} \mapsto \mathbb{R}$ can be arbitrary scalar functions of $\mathbf{s}$ (see, e.g., [29]), that are consistent with expert policy $\pi^*$. For the sake of convenience, we assume $R(\mathbf{s}, a = 0) \equiv 0$ and $R(\mathbf{s}, a = 1) = \langle \rho^*(\mathbf{X}), \mathcal{T}(\mathbf{s})\rangle$. It is worth noting that $\rho^*(\mathbf{X})$ is a function of patient features $\mathbf{X}$ and thus is able to represent the patient-specific rewards for clinicians.

To ensure that the decision policy is differentiable, we further adopt the maximum entropy assumption [49] that $\pi^*$ follows a Boltzmann distribution: $\pi^*(a|\mathbf{s}) \propto \exp[R(\mathbf{s}, a)]$, which is a *soft* version of the $\arg\max$ operator. Note that expert policy $\pi^*$ is uniquely determined by the true reward parameter map $\rho^*(\mathbf{X})$, in this paper, we seek to learn a representation $\hat{\pi}$ of expert policy $\pi^*$ with all three desiderata achieved by leveraging the notion of a transparent policy space introduced as follows.

**Transparent policy space** Following the Boltzmann distribution formulation of expert policy $\pi^*$, we can construct a transparent policy space $\Pi$ as

$$\Pi = \{\text{Bernoulli}(p) \colon p = \frac{1}{1 + \exp[-\langle \rho, \mathcal{T}(\mathbf{s})\rangle]}, \rho \in \mathbb{R}^l\}, \tag{1}$$

where $\mathcal{T}(\mathbf{s}) \in \mathcal{M}$. We call the vector $\rho \in \mathbb{R}^l$ a *policy signature* since it uniquely characterizes the behavior of the corresponding policy $\pi_\rho$ in space $\Pi$. Due to the interpretability of criteria space $\mathcal{M}$ and the linear structure in equation (1), policies in space $\Pi$ are human comprehensible by design and are considered *transparent*. Note that for the same patient feature vector $\mathbf{X} \in \mathcal{X}$, there exists a policy signature $\rho_{\mathbf{X}} = \rho^*(\mathbf{X})$ such that $\pi^*(a|\mathbf{X}, \mathbf{O}) = \pi_{\rho_{\mathbf{X}}} \in \Pi, \forall \mathbf{O} \in \mathcal{O}$. Thereby, the expert policy $\pi^*$ can be represented with patient-wise projections in space $\Pi$.

**Personalized policy projection** Given demonstrations from expert policy $\pi^*$ and criteria space $\mathcal{M}$, our target is to find a patient-wise projection $\hat{\pi}^{(\theta)}$ of policy $\pi^*$ in the transparent policy space $\Pi$ such that the distance $\mathrm{d}(\pi^*, \hat{\pi}^{(\theta)})$ is minimized via $\min_\theta \mathrm{d}(\pi^*, \hat{\pi}^{(\theta)})$, where $\hat{\pi}^{(\theta)} = \pi_{\rho(\mathbf{X};\theta)}$ and $\rho(\mathbf{X}; \theta)$ is a function of patient feature vector $\mathbf{X}$ with learnable parameter set $\theta$. The distance $\mathrm{d}(\pi^*, \hat{\pi}^{(\theta)})$ is defined as the accumulated Kullback–Leibler (KL) divergence $\mathrm{d}(\pi^*, \hat{\pi}^{(\theta)}) := \mathbb{E}_{\mathbf{s}\sim\triangle(\mathcal{X}\times\mathcal{O})}[D_{\mathrm{KL}}(\pi^* \| \hat{\pi}^{(\theta)})|\mathbf{s}]$, where $\triangle(\mathcal{X} \times \mathcal{O})$ is the organ offer distribution. The requirement for consistency of $\hat{\pi}^{(\theta)}$ is addressed in Section 3.

## 3 Individualized Transparent Policy Learning

In this paper, we propose a data-driven framework, iTransplant, for learning patient-wise policies that match clinical practice. We utilize neural networks as general function approximators for the policy signature $\rho(\mathbf{X}; \theta)$. The architecture of the proposed model is illustrated in Figure 2.

While the idea in [22] of mapping the input features into a concept space and performing prediction tasks using the concept space is in philosophy similar to our proposed method, the target of our method differs significantly from [22]. [22] aims to predict a set of human-specified concepts as an intermediate step, with each concept having a fixed (albeit learned) importance on the final prediction. Contrastingly, iTransplant aims to predict individualized policies (equivalently a set of weights) over a set of clinician-specified criteria, but not the fixed values of such criteria. The distinct targets lead to significant differences in the problem formulation, methodology and analysis in our paper and in [22].

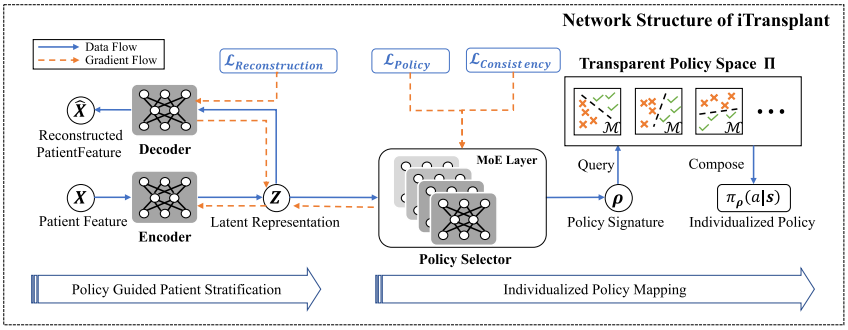



Figure 2: Network structure of the iTransplant framework.

In the proposed iTransplant framework, organ offers $\mathbf{s} = (\mathbf{X}, \mathbf{O})$ are mapped to criteria space $\mathcal{M}$ via a transform $\mathcal{T}$ specified by domain experts. The transparent policies identified by iTransplant are all generated based the match criteria in space $\mathcal{M}$. To achieve individualized policy identification, an auto-encoder structure is utilized to learn the latent representation $\mathbf{Z} \in \mathcal{Z}$ of patient features $\mathbf{X}$. From $\mathbf{Z}$, a specific policy signature $\rho$ is generated by the policy selector network. Finally, the individualized policy $\pi_\rho(a|\mathbf{s})$ is retrieved from the transparent policy space $\Pi$ via the policy signature $\rho$. This design ensures that the policies learned by iTransplant are innately individualized and human comprehensible. Detailed descriptions of the network structure and loss functions can be found in the Appendix.

**Policy guided patient stratification** As shown in Figure 2, the policy selector network is built on top of a Mixture-of-Experts (MoE) layer [36], which contains a gating network and $K$ expert networks. Based on the latent representation $\mathbf{Z}$ of a patient, the gating network will select $k$ experts among the total $K$ candidates and combine their outputs as the policy signature $\rho$. The MoE structure is applied to encourage the encoder network to group patients sharing the same decision policy in the latent space (see the Appendix for illustrations). The MoE layer in the policy selector network passes the gradient from the policy projection loss $\mathcal{L}_{Policy}$ back to the encoder network, enabling it to guide the encoder to map patients share the same policy as neighbours in the latent space $\mathcal{Z}$. In this sense, in iTransplant, the representation learning of patient features in the latent space is guided by the learned policy, which allows the encoder network to learn the implicit stratification of patients from real-world transplantation data.

**Individualized policy mapping** Our main target is to find the patient-wise policy projection $\hat{\pi}^{(\theta)} = \pi_{\rho(\mathbf{X};\theta)} \in \Pi$ of expert policy $\pi^*$ such that the distance $\mathrm{d}(\pi^*, \hat{\pi}^{(\theta)})$ is minimized. Note that the entropy of expert policy $\pi^*$ is a constant, we have $\mathrm{d}(\pi^*, \hat{\pi}^{(\theta)}) = -\mathbb{E}_{\mathbf{s}\sim\triangle(\mathcal{X}\times\mathcal{O})}[\sum_{a\in\{0,1\}} \pi^*(a|\mathbf{s}) \log(\hat{\pi}^{(\theta)}(a|\mathbf{s}))] + \text{Constant}$, where the first term is the accumulated cross-entropy between $\pi^*$ and $\hat{\pi}^{(\theta)}$. Given demonstrations $\mathcal{D} = \{(\mathbf{s}_i, a_i) \colon i = 1, 2, \ldots, N\}$ of expert policy $\pi^*$, the minimization of such accumulated KL-divergence can be achieved by minimizing the following policy projection loss $\mathcal{L}_{Policy} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{a\in\{0,1\}} \mathbb{I}(a_i = a) \log(\hat{\pi}^{(\theta)}(a_i|\mathbf{s}_i))$.

**Enforcing the consistency requirement** Note that policies in space $\Pi$ are innately consistent under perturbations to organ features. To meet the consistency requirement for inverse decision-making methods proposed in Section 2, a regularization term defined in equation (2) is enforced to improve consistency of policy projection in $\Pi$ for patients with similar latent space representations.

$$\mathcal{L}_{Consistency} = \frac{\lambda}{2}\mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2\sim\triangle(\mathcal{X}\times\mathcal{X})}\left[\frac{\|\rho_1 - \rho_2\|_2^2}{\|\mathbf{z}_1 - \mathbf{z}_2\|_2^2}\middle|\mathbf{x}_1, \mathbf{x}_2\right], \tag{2}$$

where $\triangle(\mathcal{X}\times\mathcal{X})$ is the patient pair distribution over space $\mathcal{X}\times\mathcal{X}$, $\rho_1, \rho_2$ and $\mathbf{z}_1, \mathbf{z}_2$ are policy signatures and latent representations for patients $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively, $\lambda > 0$ is a hyperparameter.

## 4 Related Work

**Medical literature on organ offer acceptance** A number of works have attempted to discover variables related to organ offer acceptance or rejection [27, 46] and to improve these decisions [41].

For instance, [33] used logistic regression models to identify donor features related to a higher risk of liver allograft discard and proposed a discard risk index that could be used to improve existing organ allocation policies. This donor discard index describes donor factors but does not address the reasons why a specific center may or may not reject a donor organ.

Table 1: Comparison with Related Work. iTransplant satisfies three key desiderata for understanding clinicians' decision-making for organ transplant offers: (1) individualized policies; (2) interpretability; (3) consistency of interpretation. Organ offers, decisions of clinicians, patient feature vector, criteria space for policy generation, expert policy, recovered policy, feature selection function and learned contributions of different criteria to decisions are denoted as $\mathbf{s}$, $a$, $\mathbf{X}$, $\mathcal{M}$, $\pi^*$, $\hat{\pi}$, $\mathcal{S}$ and $\hat{\rho}$, respectively.

| APPROACH | EXAMPLE REF | INPUT | OUTPUT | (1) | (2) | (3) |
|---|---|---|---|---|---|---|
| IMITATION LEARNING | [38] | $\{\mathbf{s}_t, a_t\} \sim \pi^*$ | $\hat{\pi}$ | ✗ | ✗ | ✗ |
| GLOBAL VARIABLE SELECTION | [31] | $\{\mathbf{s}_t, a_t\} \sim \pi^*$ | $\hat{\pi}, \mathcal{S}$ | ✗ | ✗ | ✗ |
| INSTANCE-WISE VARIABLE SELECTION | [48] | $\{\mathbf{s}_t, a_t\} \sim \pi^*$ | $\hat{\pi}, \mathcal{S}(\mathbf{s})$ | ✓ | ✗ | ✗ |
| INVERSE REINFORCEMENT LEARNING[1] | [49] | $\{\mathbf{s}_t, a_t\} \sim \pi^*, \mathcal{M}$ | $\hat{\pi}, \hat{\rho}$ | ✗ | ✓ | ✗ |
| ITRANSPLANT | (OURS) | $\{\mathbf{s}_t, a_t\} \sim \pi^*, \mathcal{M}$ | $\hat{\pi}, \hat{\rho}(\mathbf{X})$ | ✓ | ✓ | ✓ |

In this paper, we also seek to identify key variables and criteria that inform clinical decisions on organ offers. Therefore, our method is primarily designed as *a tool for understanding* the different factors or the weight of those factors between decision-making of clinicians rather than a model to describe commonly rejected donors. Our method learns clinicians' policies at the patient level rather than a global policy, enabling our method to identify different policies for different patient cohorts.

Because the information derived from iTransplant is specific to a donor offered to a center, it may act as a better prompt to de-bias clinical decision-making. Prior attempts to de-bias clinical decision-making have involved generic prompts relating to common errors in decision-making [1, 7, 8] and have been met with variable success. In contrast, our approach identifies specific factors that are involved in individual donor organ – patient acceptance decisions.

**Interpretation by feature-selection** Both global [13, 31] and instance-wise [48] feature selection methods could be applied to identify important variables involved in organ offer decisions. In comparison, iTransplant is, by design, capable of discovering important variables for decisions on organ offers at both the global and patient level. In addition, unlike methods which represent policies directly by neural networks, our method offers interpretability via the transparent policy space $\Pi$, which is critical for high-stake scenarios such as decision-making for organ offers.

**Imitation learning with post-hoc interpretation** Imitation learning (IL) [18] seeks to replicate expert polices from observational data of decision-making. In the one-step decision-making scenario considered in this paper, behavioral cloning (BC) [3, 38] could be directly applied to imitate the expert policy. However, as a black-box model, the neural networks utilized in BC provide no directly human comprehensible interpretations to the learned policy and have to rely on post-hoc interpretability methods like LIME [34] and SHAP [26] to provide insight for their predictions.

**Inverse reinforcement learning** Compared to imitation learning, inverse reinforcement learning (IRL) techniques [32, 30] seek to infer a reward function that is consistent with observed expert behaviors. The learned reward function can then be used for explanation of expert behaviors. For instance, with feature maps from domain knowledge and the assumption of linear reward structure, maximum entropy IRL [49] can effectively reveal contributions of different features to expert behaviors. Note that in the one-step decision-making setting considered in this paper, maximum entropy IRL [49] degenerates to a global logistic regression model, which means that it is unable to capture the variations in decisions made for different subgroups of patients.

[1] In the setting considered in this paper, IRL methods with assumptions of maximum-entropy policy and linear reward structures degenerate to logistic regression. In this case, their interpretations become consistent under perturbations. However, many IRL methods would be inconsistent in more general settings.

A comparison of iTransplant with alternate methods that could be applied to discover potential drivers of clinicians' decisions in the organ transplantation setting is provided in Table 1.

## 5 Illustrative Examples

To demonstrate the advantage of the proposed individualized policy learning framework and showcase potential applications of iTransplant as a tool for discovering variations in transplantation practices, real-world liver transplantation data from OPTN are utilized to provide several illustrative examples. In all examples included in this section, the criteria space $\mathcal{M}$ is constructed from a set of match criteria manually specified based on domain knowledge, and a detailed explanation of these match criteria can be found in the Appendix, together with full experimental details.

### 5.1 Model Validation

We first sought to validate the ability of iTransplant to predict organ offer acceptance using organ offer data from ten transplant centers with sufficient number (over 1,000) of accepted offers after removal of missing data. We compared iTransplant to a variety of baselines encompassing the approaches outlined in Table 1, including logistic regression, a white-box model frequently used in the medical literature, and neural network-based behavioral cloning (BC). Despite being a black-box model, and thus unsuitable to achieve our goal of *understanding* decision-making, we include BC as an upper bound on predictive performance using the proposed match criteria. We measure performance using three metrics: area under the receiver operating characteristic curve (AUC-ROC), area under the precision recall curve (AUC-PRC) and log-likelihood (LL) of observed samples. Hyperparameters for all methods can be found in the Appendix.

Table 2: Benchmark of different methods on decision prediction.

| METHOD | AUC-ROC | AUC-PRC | LL |
|---|---|---|---|
| LOGISTIC REGRESSION | 0.794±0.054 | 0.341±0.061 | -0.538±0.051 |
| PER-CLUSTER LOGISTIC REGRESSION | 0.803±0.049 | 0.352±0.063 | -0.527±0.048 |
| DECISION TREE | 0.775±0.057 | 0.274±0.069 | -0.552±0.060 |
| PER-CLUSTER DECISION TREE | 0.773±0.052 | 0.281±0.066 | -0.564±0.064 |
| LOCALLY WEIGHTED REGRESSION | 0.865±0.044 | 0.429±0.066 | -0.256±0.089 |
| LASSO | 0.777±0.064 | 0.314±0.068 | -0.570±0.045 |
| RANDOM FOREST | 0.852±0.064 | 0.421±0.106 | -0.271±0.092 |
| INVASE | 0.790±0.062 | 0.341±0.071 | -0.541±0.064 |
| BEHAVIORAL CLONING | 0.899±0.043 | 0.502±0.067 | -0.383±0.067 |
| ITRANSPLANT (OURS) | 0.895±0.045 | 0.502±0.062 | -0.396±0.069 |

As shown in Table 2, iTransplant significantly outperforms all baseline methods and performs similarly to the upper bound provided by BC while maintaining interpretability of policies identified for each patient. Although the policy learned by iTransplant uses logistic regression as the decision function, by learning individualized policies, iTransplant greatly outperformed both global logistic regression and per-cluster logistic regression (see Appendix for further details).

While BC and INVASE are unable to provide direct/sufficient insight into clinical decision-making (our primary goal), post-hoc interpretations could offer an alternate approach. To assess the suitability of such interpretations, we evaluated the consistency of the interpretations of iTransplant and BC/INVASE with post-hoc interpretation via LIME [34]. If a feature has a positive weight, increasing the value of the feature should increase the probability of acceptance. Indeed, this is a requirement for the interpretation to be useful. Thus we performed perturbations to single patient or organ features (25% of a standard deviation for continuous variables and flipped value for binary ones) and measured the consistency of sign between the actual deviation in prediction and the expected deviation from counterfactual interpretation (via importance of perturbed feature) on 200 organ offers.

Table 3: Consistency of interpretations.

| METHOD | CONSISTENCY |
|---|---|
| BC WITH LIME | 46.8% |
| INVASE WITH LIME | 41.7% |
| ITRANSPLANT | 85.0% |

Table 3 shows that iTransplant is highly consistent while BC/INVASE with post-hoc interpretation via LIME [34] barely exceeds random. As a result, iTransplant can be used to understand clinical decision-making, while the low consistency of post-hoc interpretations for BC and INVASE render them inappropriate for this purpose.

iTransplant benefits from the learning capability of neural networks to perform comparably with black-box models while maintaining interpretability of the learned decision-making policies. This makes iTransplant a powerful tool to investigate the decision-making policies in organ transplantation.

### 5.2 Investigative Experiments

As shown above, iTransplant can effectively learn transparent policies from the decision-making history of clinicians while achieving significant performance improvements to white-box models. Here, we demonstrate how iTransplant can be used to investigate clinicians' policies on organ offer acceptance from three perspectives:

1: Which criteria are important for clinicians' decisions on organ offers?

2: How does clinical practice change for patients with different characteristics?

3: What variation in practice exists between transplant centers?

Based on the similarity in organ acceptance rates and total numbers of organ offered, two transplant centers with ID codes 16864 and 19034 are selected for these investigative experiments. For convenience, we denote these centers as center A and center B, respectively. For the first two experiments, transplantation data from center A is used, while for the last one, we use iTransplant to learn separate policies for each center to explore the policy variation between the two centers.

**Discovering important match criteria** The relative importance of different match criteria can be measured by the normalized policy signature $\tilde{\rho} = \rho / \|\rho\|_2$. With the distribution of normalized policy signature $\tilde{\rho}$ plotted in Figure 3, we are able to examine the learned policy at the population level. For most patients, as highlighted in Figure 3, the donor age and weight, presence of hepatitis B (HBV) or C (HBC), non-heart-beating donors, cause of donor death (donation after natural death), and high percentages of macrosteatosis (MaS) have negative contributions to the acceptance of organ offers. These observations are in line with risk factors reported in medical literature [33, 19] and guidelines [6].

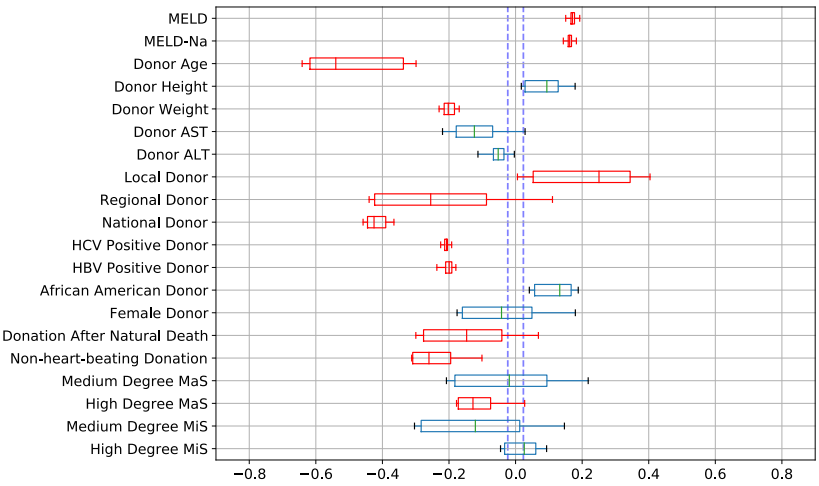
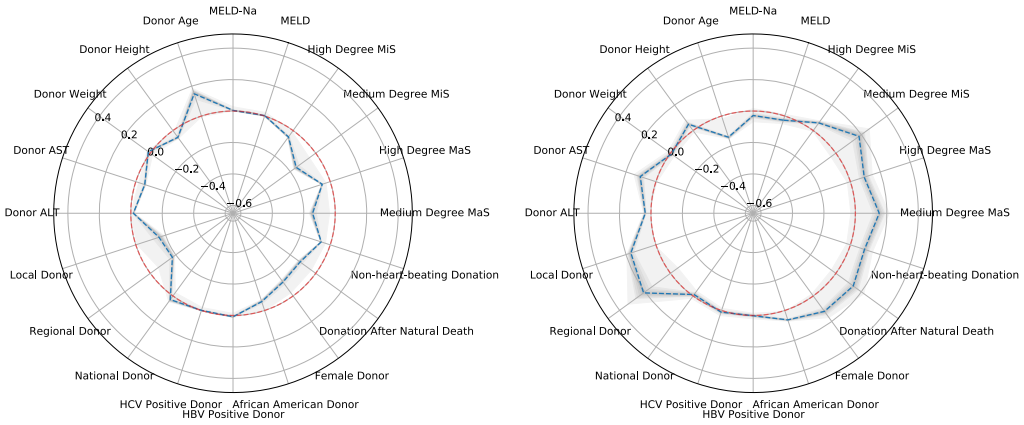



Figure 3: Distribution of normalized policy signature $\tilde{\rho}$ (x-axis) over match criteria (y-axis), see the Appendix for explanation. Criteria with weights between the dotted blue lines could be ignored with no more than 0.1% loss in the average precision score.

On the other hand, organs from local donors are preferred by clinicians while the organs shared at regional or national levels are more likely to be declined. This is supported by research on the impact of cold ischemia time (CIT), which is often a consequence of the organ origin [25, 9, 10]. Also,

Figure 3 shows that, for a given patient, a higher score from the model for end-stage liver disease (MELD) [44] will lead to higher chance of organ offer acceptance. As an indicator for 3-month mortality of hospitalized patients, a high MELD score suggests greater severity of the liver diseases, which may potentially encourage clinicians to accept the offered organ.

**Identifying patient-specific organ preferences of clinicians** Here, we use iTransplant to investigate the patient-specific organ preferences of clinicians and how their decision-making is affected by patient features. We cluster the patients in the test set for center A by applying the KMeans algorithm to the latent space $\mathcal{Z}$. The number of clusters is determined to be six such that the between-cluster variance of policy signatures $\rho$ is maximized. We find that patients in different clusters have significant deviations in policy signature indicating differences in clinical practice due to specific patient features. In Figure 4, we provide an overview of the policies for patients in cluster 1 and 2 displaying the deviation in policy with respect to the averaged policy signature $\bar{\rho}$ for all patients (see the Appendix for detailed discussion).




(a) Patient cluster 1.

(b) Patient cluster 2.

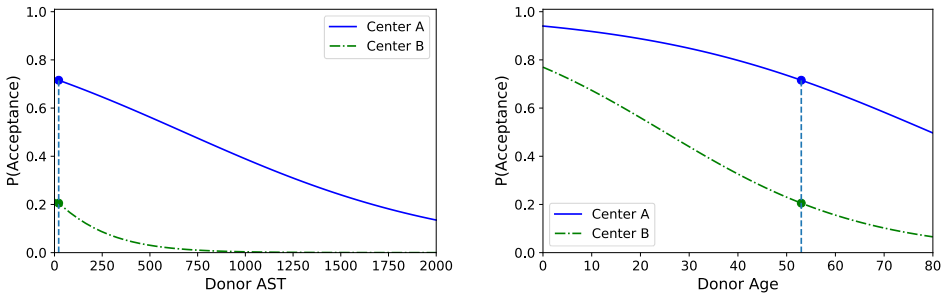
Figure 4: The deviations of policy signature $\Delta\rho = \rho - \bar{\rho}$ of two patient clusters.

Organs procured from regional donors are significantly more likely to be accepted for patients in cluster 2 than cluster 1 (Figure 4(a)). One significant difference between clusters 2 and 3 is whether patients have hepatocellular cancer (HCC). In cluster 2, 62.1% of patients are HCC positive with a MELD score below 20, while only 1.4% of patients in cluster 1 have both low MELD scores and the positive HCC status. Organs from regional donors typically have increased CIT compared to local donors. Prolonged CIT significantly reduces graft and patient survival, and thus organs from local donors are typically preferred (Figure 3). However, HCC positive patients with low MELD scores ($< 20$) are able to tolerate longer ischemia times without significant impact on graft survival [24], explaining the increased weight for regional donor (Figure 4(a)).

To illustrate the advantage and effectiveness of iTransplant in learning individualized policies for different patient groups, the policies learned by iTransplant and by logistic regression method are compared in the Appendix.

**Discovering variations in polices across transplant centers** Here, we illustrate the use of iTransplant as a tool for discovering patient-level policy variations across transplant centers. To achieve this, policies $\hat{\pi}^A$ and $\hat{\pi}^B$ are trained with organ offer data from center A and B, respectively, and are compared on the test set for center A. The probability of offer acceptance predicted by policies $\hat{\pi}^A$ and $\hat{\pi}^B$ was similar when averaged across all patients (difference: 0.005). However, there were significant differences on a per-patient basis, with a mean absolute error of $0.11 \pm 0.14$, resulting in different decisions for many patients. We identify a major divergence (predicted probabilities of offer acceptance are 0.72 and 0.21 for $\hat{\pi}^A$ and $\hat{\pi}^B$, respectively) of these two policies for an accepted organ offer associated with a 38 year old patient with high MELD score (39.0). The organ offered to this patient comes from a 53 year old donor with aspartate aminotransferase (AST) test value of

22.0, and the donor AST value and donor age were found to be two important factors related to such divergence in the considered policies. To demonstrate how this divergence in policies $\hat{\pi}^A$ and $\hat{\pi}^B$ affects the patient outcome, we assess the counterfactual impact of two donor characteristics on the organ offer acceptance probability by altering the age and AST test value of the donor (Figure 5).




(a) Counterfactual impact of donor AST test value. (b) Counterfactual impact of donor age.

Figure 5: Patient-level variations in policies from two different transplant centers.

The AST test value is an indirect measure of liver cell integrity, with a higher value in a donor suggesting more severe damage and therefore potentially worse post transplant outcome. Both policy $\hat{\pi}^A$ and $\hat{\pi}^B$ show preferences for organ offers with lower AST test value. However, for the considered patient, policy $\hat{\pi}^A$ is more tolerant to changes in donor AST test results. The low AST test result of the donor leads to a significantly higher likelihood of offer acceptance compared to policy $\hat{\pi}^B$ (Figure 5(a)). We find similar variations when examining the impact of donor age on the offer acceptance probability. While policy $\hat{\pi}^B$ has a high chance of accepting the organ offer from a young donor, the acceptance rate decreases steadily as the counterfactual donor age increases. For the policy $\hat{\pi}^A$, the acceptance rate is largely insensitive to the donor age until the age of the donor exceeds 30.

With the counterfactual impacts of donor features plotted in Figure 5, we can intuitively observe the variations in transplantation practices in the investigated transplant centers. This demonstrates the potential of iTransplant as a tool for clinicians to examine their own decision-making policies with existing transplantation data and the impact of changes to patient and donor characteristics.

## 6 Conclusion

In this paper, we propose an individualized transparent policy learning framework iTransplant to understand the organ offer related decision-making of clinicians. The individualized policy learned by iTransplant shows significant advantage compared to a global policy from logistic regression, highlighting the need to learn personalized policies. With real-world liver transplantation data, we conducted three investigative experiments demonstrating the use of iTransplant as a tool for discovering patient-level variations in transplantation practice and thus informing iterations of clinical decision support tools. While our experiments demonstrate that most decisions can be explained by clinical features recorded in observational datasets, it is likely that some decisions were made based on factors that were not considered. Future work could explore the impact of including additional match criteria and contextual information concerning unit activity and outcomes.

## Acknowledgements

This work was supported by Cystic Fibrosis Trust, Alzheimer's Research UK, the US Office of Naval Research (ONR), and the National Science Foundation (NSF, grant number 1722516). We thank the anonymous reviewers for their comments and suggestions. We thank the clinicians who participated in our survey. We thank Dr. Brent D. Ershoff for valuable discussions and counsel on the OPTN dataset and the observational features and decision criteria considered in this work. This work is based on the liver transplant data from OPTN, which was supported in part by Health Resources and Services Administration contract HHSH250-2019-00001C.

## References


[1] S. Almashat, B. Ayotte, B. Edelstein, and J. Margrett. Framing effect debiasing in medical decision making. *Patient education and counseling*, 71(1):102–107, 2008.

[2] F. Atsma, G. Elwyn, and G. Westert. Understanding unwarranted variation in clinical practice: a focus on network effects, reflective medicine and learning health systems. *International Journal for Quality in Health Care*, 32(4):271–274, 2020.

[3] M. Bain and C. Sammut. A framework for behavioural cloning. In *Machine Intelligence 15*, pages 103–129. Oxford University Press, 1996.

[4] J. Berrevoets, J. Jordon, I. Bica, M. van der Schaar, et al. Organite: Optimal transplant donor organ offering using an individual treatment effect. *Advances in Neural Information Processing Systems*, 33, 2020.

[5] T. Bittermann and D. S. Goldberg. Quantifying the effect of transplanting older donor livers into younger recipients: the need for donor-recipient age matching. *Transplantation*, 102(12):2033, 2018.

[6] J. Chhatwal, S. Samur, E. D. Bethea, T. Ayer, F. Kanwal, C. Hur, M. S. Roberts, N. Terrault, and R. T. Chung. Transplanting hepatitis c virus-positive livers into hepatitis c virus-negative patients with preemptive antiviral treatment: A modeling study. *Hepatology (Baltimore, Md.)*, 67(6):2085–2095, 2018.

[7] P. Croskerry, G. Singhal, and S. Mamede. Cognitive debiasing 1: origins of bias and theory of debiasing. bmj qual saf 22, ii58–ii64, 2013.

[8] P. Croskerry, G. Singhal, and S. Mamede. Cognitive debiasing 1: origins of bias and theory of debiasing. bmj qual saf 22, ii65–ii72, 2013.

[9] P. Dutkowski, C. E. Oberkofler, K. Slankamenac, M. A. Puhan, E. Schadde, B. Müllhaupt, A. Geier, and P. A. Clavien. Are there better guidelines for allocation in liver transplantation?: A novel score targeting justice and utility in the model for end-stage liver disease era. *Annals of surgery*, 254(5):745–754, 2011.

[10] S. Feng, N. Goodrich, J. Bragg-Gresham, D. Dykstra, J. Punch, M. DebRoy, S. M. Greenstein, and R. Merion. Characteristics associated with liver graft failure: the concept of a donor risk index. *American Journal of Transplantation*, 6(4):783–790, 2006.

[11] J. Godown, M. McKane, K. A. Wujcik, B. A. Mettler, and D. A. Dodd. Regional variation in the use of 1a status exceptions for pediatric heart transplant candidates: is this equitable? *Pediatric transplantation*, 21(1):e12784, 2017.

[12] D. S. Goldberg, B. French, J. D. Lewis, F. I. Scott, R. Mamtani, R. Gilroy, S. D. Halpern, and P. L. Abt. Liver transplant center variability in accepting organ offers and its impact on patient survival. *Journal of Hepatology*, 64(4):843–851, 2016.

[13] M. A. Hall. *Correlation-based feature selection for machine learning*. University of Waikato, 1999.

[14] A. Hart, J. M. Smith, M. A. Skeans, S. K. Gustafson, A. R. Wilk, S. Castro, J. Foutz, J. L. Wainright, J. J. Snyder, B. L. Kasiske, and A. K. Israni. Optn/srtr 2018 annual data report: Kidney. *American Journal of Transplantation*, 20(s1):20–130, 2020.

[15] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, pages 417–424, 2006.

[16] K. Hung, J. Gralla, J. L. Dodge, K. M. Bambha, M. Dirchwolf, H. R. Rosen, and S. W. Biggins. Optimizing repeat liver transplant graft utility through strategic matching of donor and recipient characteristics. *Liver Transplantation*, 21(11):1365–1373, 2015.

[17] S. A. Husain, K. L. King, S. Pastan, R. E. Patzer, D. J. Cohen, J. Radhakrishnan, and S. Mohan. Association between declined offers of deceased donor kidney allograft and outcomes in kidney transplant candidates. *JAMA Network Open*, 2(8):e1910312–e1910312, 2019.

[18] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

[19] C. J. Imber, S. D. St. Peter, I. Lopez, L. Guiver, and P. J. Friend. Current practice regarding the use of fatty livers: a trans-atlantic survey. *Liver Transplantation*, 8(6):545–549, 2002.

[20] A. K. Israni, D. Zaun, N. Hadley, J. Rosendale, C. Schaffhausen, W. McKinney, J. J. Snyder, and B. L. Kasiske. Optn/srtr 2018 annual data report: Deceased organ donation. *American Journal of Transplantation*, 20(s1):509–541, 2020.

[21] P. Kamath and W. Kim. Advanced liver disease study groupthe model for end-stage liver disease (MELD). *Hepatology*, 45:797–805, 2007.

[22] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

[23] A. Kwong, W. R. Kim, J. R. Lake, J. M. Smith, D. P. Schladt, M. A. Skeans, S. M. Noreen, J. Foutz, E. Miller, J. J. Snyder, A. K. Israni, and B. L. Kasiske. Optn/srtr 2018 annual data report: Liver. *American Journal of Transplantation*, 20(s1):193–299, 2020.

[24] V. J. Lozanovski, B. Döhler, K. H. Weiss, A. Mehrabi, and C. Süsal. The differential influence of cold ischemia time on outcome after liver transplantation for different indications—who is at risk? a collaborative transplant study report. *Frontiers in Immunology*, 11:892, 2020.

[25] V. J. Lozanovski, E. Khajeh, H. Fonouni, J. Pfeiffenberger, R. von Haken, T. Brenner, M. Mieth, P. Schirmacher, C. W. Michalski, K. H. Weiss, et al. The impact of major extended donor criteria on graft failure and patient mortality after liver transplantation. *Langenbeck's archives of surgery*, 403(6):719–731, 2018.

[26] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[27] L. McCormack, P. Dutkowski, A. M. El-Badry, and P.-A. Clavien. Liver transplantation using fatty livers: always feasible? *Journal of hepatology*, 54(5):1055–1062, 2011.

[28] J. Neuberger, A. Gimson, M. Davies, M. Akyol, J. O'Grady, A. Burroughs, M. Hudson, U. Blood, et al. Selection of patients for liver transplantation and allocation of donated livers in the uk. *Gut*, 57(2):252–257, 2008.

[29] A. Y. Ng, D. Harada, and S. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287, 1999.

[30] A. Y. Ng and S. J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 663–670, 2000.

[31] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[32] D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pages 2586–2591, 2007.

[33] A. Rana, R. R. Sigireddi, K. J. Halazun, A. Kothare, M.-F. Wu, H. Liu, M. L. Kueht, J. M. Vierling, N. L. Sussman, A. L. Mindikoglu, et al. Predicting liver allograft discard: the discard risk index. *Transplantation*, 102(9):1520–1529, 2018.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[35] D. L. Segev, W. R. Maley, C. E. Simpkins, J. E. Locke, G. C. Nguyen, R. A. Montgomery, and P. J. Thuluvath. Minimizing risk associated with elderly liver donors by matching to preferred recipients. *Hepatology*, 46(6):1907–1918, 2007.

[36] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[37] G. Thabut, J. D. Christie, W. K. Kremers, M. Fournier, and S. D. Halpern. Survival differences following lung transplantation among us transplant centers. *Jama*, 304(1):53–60, 2010.

[38] F. Torabi, G. Warnell, and P. Stone. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4950–4957, 2018.

[39] A. Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, pages 1–15, 2019.

[40] F. R. Vogenberg, C. Isaacson Barash, and M. Pursel. Personalized medicine: part 1: evolution and development into theranostics. *Pharmacy and Therapeutics*, 35(10):560–576, 2010.

[41] M. L. Volk, N. Goodrich, J. C. Lai, C. Sonnenday, and K. Shedden. Decision support for organ offers in liver transplantation. *Liver transplantation*, 21(6):784–791, 2015.

[42] J. E. Wennberg. Unwarranted variations in healthcare delivery: implications for academic medical centres. *BMJ*, 325(7370):961–964, 2002.

[43] A. Wey, M. Valapour, M. A. Skeans, N. Salkowski, M. Colvin, B. L. Kasiske, A. K. Israni, and J. J. Snyder. Heart and lung organ offer acceptance practices of transplant programs are associated with waitlist mortality and organ yield. *American Journal of Transplantation*, 18(8):2061–2067, 2018.

[44] R. Wiesner, E. Edwards, R. Freeman, A. Harper, R. Kim, P. Kamath, W. Kremers, J. Lake, T. Howard, R. M. Merion, et al. Model for end-stage liver disease (meld) and allocation of donor livers. *Gastroenterology*, 124(1):91–96, 2003.

[45] R. Wolfe, F. LaPorte, A. Rodgers, E. Roys, G. Fant, and A. Leichtman. Developing organ offer and acceptance measures: when 'good' organs are turned down. *American journal of transplantation*, 7:1404–1411, 2007.

[46] H. Yersiz, C. Lee, F. M. Kaldas, J. C. Hong, A. Rana, G. T. Schnickel, J. A. Wertheim, A. Zarrinpar, V. G. Agopian, J. Gornbein, et al. Assessment of hepatic steatosis by transplant surgeon and expert pathologist: a prospective, double-blind evaluation of 201 donor livers. *Liver Transplantation*, 19(4):437–449, 2013.

[47] J. Yoon, A. Alaa, M. Cadeiras, and M. Van Der Schaar. Personalized donor-recipient matching for organ transplantation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[48] J. Yoon, J. Jordon, and M. van der Schaar. Invase: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*, 2018.

[49] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

# A Appendix

## A.1 Data used for experiments

**Data source** The experiments in our paper are conducted on a custom limited dataset (covering organ offers from January 1, 2003 to December 4, 2020) based on OPTN data as of December 4, 2020. The OPTN data are publicly available, and no patient identifiers or possibly identifiable information are included in the custom limited OPTN dataset used in this paper. Please see instructions on requesting OPTN data and the differences between custom limited datasets and patient-identified datasets on https://optn.transplant.hrsa.gov/data/request-data/data-request-instructions/.

**Selection of organ offer data** In the OPTN system, when an organ offer is assigned to a potential recipient in one transplant center, there are two types of initial response – *provisional acceptance* and *rejection*. More details of the donor organ are then provided to centers of *provisional acceptance*, and the organ will be offered to one of these centers based on their final decisions (*acceptance* or *rejection*) and the priority of their patients in the organ offering system. Detailed OPTN policies for organ allocation and organ offer acceptance can be found in https://optn.transplant.hrsa.gov/governance/policies/. The reason of rejection was provided if the organ offer was declined by a transplant center, while no reason was recorded for *acceptance* and *provisional acceptance*. In some cases, the initial response of *provisional acceptance* is not updated to *acceptance* or *rejection*, and the final decision is not known. Note that in our custom limited dataset from OPTN, with 0.96% of *acceptance* and 95.58% of *rejection*, only 3.46% of responses to organ offers are *provisional acceptance*. Thus, in our experiment, we only include organ offers with a final decision of *acceptance* or *rejection*. In some special cases, the reasons for organ offer rejection are either administrative in nature (e.g., "surgeon unavailable" and "heavy workload") and do not directly relate to possible patient or organ factors available in our datasets (e.g., "patient's condition improved, transplant not needed"). For the experiments in this paper, we select rejected organ offers of twelve cases, covering 87.7% of all rejected offers. These considered reasons of rejection include: donor age or quality (code 830, 922), donor size or weight (code 831, 923), organ-specific donor issue (code 837), distance to travel or ship (code 824), donor quality (code 921), positive serological tests of donor (code 834), donor organ anatomical damage or defect (code 836), donor blood type (code 832, 924) and abnormal biopsy of donor liver (code 942).

**Patient and donor organ features used in experiments** Based on domain knowledge on liver transplantation, subsets of patient and donor organ features related to decisions on organ offers are selected to construct the patient feature space $\mathcal{X}$ and organ feature space $\mathcal{O}$, respectively. The patient feature vector $\mathbf{X} \in \mathcal{X}$ and organ feature vector $\mathbf{O} \in \mathcal{O}$ are calculated using the original transplantation data. Organ offers which contain missing values in patient or organ features were excluded from the dataset.

The list of patient features selected includes: age, gender, height, weight, blood type, body mass index (BMI), creatinine concentration, international normalized ratio (INR), bilirubin concentration, sodium concentration (Na), MELD score, albumin concentration, status of ascites, ethnicity, status of dialysis, status 1a, status of hepatocellular carcinoma (HCC), indicator of hepatocellular carcinoma exception point, functional status of patient, indicator of life support status, indicator of whether the patient was on a mechanical ventilator, status of portal vein thrombosis, history of prior abdominal surgery, number of previous transplants and primary cause of liver disease.

The list of donor organ features selected includes: age, gender, height, weight, blood type, creatinine concentration, bilirubin concentration, aspartate aminotransferase (AST) test result, alanine aminotransferase (ALT) test result, indicator of non-heart beating donor, cause of death for deceased donor, ethnicity, donor organ origin, death mechanism of donor, circumstances of death (natural death or not), organ procure type (whole or split), percentage of macrosteatosis (MaS), percentage of microsteatosis (MiS), hepatitis B serostatus and hepatitis C serostatus.

For the experiments in this paper, we only consider adult patients that joined the waitlist after 2002. We also removed patients who have been retransplanted, cases of living donor transplantation, and multiorgan transplantation. Thereafter, records with missing values for selected patient and donor organ features are removed.

**Match criteria used in experiments** Most match criteria used in the experiments are directly obtained as donor organ features. Here, we provide additional explanations for some of match criteria related to patient features. The donor organ origin (local/regional/national donor) with respect to patient location is calculated based on distance information in the patient-organ match classification from OPTN data. According to [21, 44], for the MELD and MELD-Na score calculation, we use the following:

$$\text{MELD} = \lfloor 6.43 + 9.57 \times \log(\text{creatinine}) + 3.78 \times \log(\text{bilirubin}) + 11.2 \times \log(\text{INR}) \rceil,$$

$$\text{MELD-Na} = \begin{cases} \text{MELD}, & \text{if MELD} < 12, \\ \text{MELD} + 1.32 \times (137 - \text{Na}) - 0.033 \times (137 - \text{Na}) \times \text{MELD}, & \text{otherwise}, \end{cases}$$

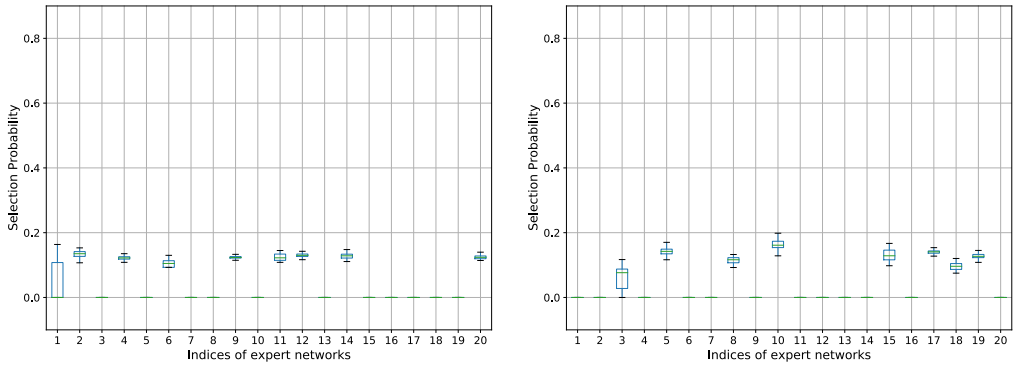
where $\lfloor \cdot \rceil$ rounds its input to the nearest integer. In addition, the creatinine value is clamped into $[1, 4]$ and is adjusted to $4.0$ if the patient had dialysis at least twice in the past week. The INR and bilirubin values are clamped into $[1, 1 \times 10^{10}]$ to avoid negative MELD scores. For MELD-Na score, the Na value is adjusted to the range of $[125, 137]$, and the final MELD-Na score is clamped into range $[6, 40]$.

### A.2 Additional explanations for the network structure of iTransplant

**Reconstruction loss** For the auto-encoder part in iTransplant, the standard reconstruction loss $\mathcal{L}_{Reconstruction}$ is adopted to guide the learning of latent representation of patient features. Here, $\mathcal{L}_{Reconstruction}$ is calculated as the mean squared error between the reconstructed patient feature $\hat{\mathbf{X}}$ from the decoder network and the original patient feature vector $\mathbf{X} \in \mathcal{X}$ as shown in equation (3)

$$\mathcal{L}_{Reconstruction} = \int_{\mathcal{X}} \|\mathbf{X} - \hat{\mathbf{X}}\|_2^2 \mathrm{d}\mathbf{X}. \tag{3}$$

**Joint loss function** The three loss functions of policy loss $\mathcal{L}_{Policy}$, reconstruction loss $\mathcal{L}_{Reconstruction}$ and consistency loss $\mathcal{L}_{Consistency}$ are added together for the training of iTransplant. Note that the gradients from $\mathcal{L}_{Consistency}$ only apply to the policy selector network while gradients from $\mathcal{L}_{Policy}$ are passed to the encoder network to guide the representation learning of patient features in the latent space $\mathcal{Z}$ together with $\mathcal{L}_{Reconstruction}$.




(a) Cluster 1. (b) Cluster 2.

Figure 6: The selection probability of expert networks for patients in two different clusters.

**Mixture of expert (MoE) layer** The MoE layer of the policy selector network in iTransplant contains a gating network and $K$ expert networks. In this paper, the gating network is represented by a $K$-by-$m$ matrix $\mathbf{G}$, where $m$ is the dimension of the latent space $\mathcal{Z}$. The expert networks are ranked based on the selection likelihoods calculated from the matrix product $\mathbf{GZ}$, where $\mathbf{Z} \in \mathcal{Z}$ is the latent representation of patient features $\mathbf{X}$ , and outputs of the top-$k$ experts in the MoE layer are combined to make predictions. For patients from different cohorts, variations in their latent representations in $\mathcal{Z}$ would lead to the selection of different expert networks. The distribution of expert selection

probability with organ offer data from transplant center with ID code 16864 is illustrated in Figure 6. With $K = 20$ experts available, the gating network in MoE layer tends to assign different sets of expert networks for patients from different clusters. Note that the patient clusters here are the same with those in the main manuscript.

### A.3 Experiment setup

**Selection of transplant centers** For experiments in this paper, organ offer data from ten transplant centers with sufficient number (over 1,000) of accepted offers after removal of missing data are selected. Some basic information of these ten centers can be found in Table 4.

Table 4: Statistics of organ offers in the selected centers.

| CENTER ID CODE | ORGAN OFFERS[2] | ACCEPTANCE RATE | ACCEPTED ORGAN OFFERS |
|---|---|---|---|
| 20522 | 12140 | 18.59% | 2257 |
| 19034 | 17608 | 11.36% | 2001 |
| 6820 | 13120 | 12.90% | 1693 |
| 7471 | 16429 | 10.29% | 1691 |
| 23312 | 31013 | 5.18% | 1607 |
| 14942 | 11148 | 13.37% | 1491 |
| 16864 | 14030 | 10.63% | 1491 |
| 23808 | 29964 | 4.86% | 1457 |
| 25110 | 31313 | 4.38% | 1371 |
| 13609 | 13764 | 9.63% | 1326 |

**Model configurations** In the main manuscript, baselines of (per-cluster) logistic regression, (per-cluster) decision tree, locally weighted regression (LOWESS), random forest, global variable selection with LASSO, instance-wise variable selection with INVASE and a neural network-based behavioural cloning (BC) model are used for comparison with iTransplant. For the baselines of logistic regression, decision tree, random forest and LASSO, implementations in scikit-learn[3] are adopted. The logistic regression-based policies in iTransplant include no interception terms, and no penalty is applied to the policy signatures identified by iTransplant. In line with this, we set zero-penalty and zero-interception options for the coefficients in logistic regression. For the baselines of per-cluster logistic regression and per-cluster decision tree, KMeans algorithm is applied to discover patient clusters. Logistic regression models and decision trees are then trained independently for each of these clusters. For BC, a multi-layer perceptron (MLP) is used to model the map from an organ offer $\mathbf{s}$ to the probability of offer acceptance. For INVASE, three MLPs are utilized as the baseline, critic and selector networks, respectively. Similarly, in iTransplant, the encoder and decoder networks and the experts in the MoE layer are all MLPs with linear activation functions for outputs. The activation functions for non-output layers in BC, INVASE and iTransplant are set to be *LeakyReLU*.

**Model inputs** For all baselines, match criteria $\mathcal{T}(\mathbf{s})$ in the criteria space $\mathcal{M}$ are used by the models to achieve action matching with the expert policy. The patient feature vector $\mathbf{X}$ is provided additionally to per-cluster logistic regression and per-cluster decision tree models for the clustering of patients via KMeans. Similarly, the patient feature vector $\mathbf{X}$ is also provided to LOWESS for the measurement of patient similarity. To provide the upper bounds of performance, both patient features space $\mathcal{X}$ and the criteria space $\mathcal{M}$ are used as inputs to BC. Further evaluation with respect to different model inputs for the baselines can be found in Section A.4.

**Selection of hyperparameters** For hyperparameter selection, benchmark and investigative experiments, the learning rate for all neural networks is set to be $1 \times 10^{-3}$ and the maximum number of training iterations is set as 200. The early stopping technique is adopted for BC and iTransplant in the training phase to avoid over-fitting.

For MLPs in BC, INVASE and iTransplant, the number of hidden layers and the number of units in each layer are denoted as $l_n$ and $h_n$, respectively. For INVASE, the hyperparameter $\lambda$ refers to the

[2]Organ offers with missing values in selected patient and organ features are not counted.

[3]https://scikit-learn.org/stable/

coefficient for $l_1$ penalty term for its instance-wise feature selection masks. For iTransplant, the multiplier for load-balancing in the MoE layer is kept as is in the implementation[4] (loss_coef=0.01). The number of experts in the MoE layer, the number of selected experts for policy signature identification and coefficient for the consistency loss are denoted as $K$, $k$ and $\lambda$, respectively.

For (per-cluster) logistic regression and decision tree models and random forest, the number of clusters, number of decision trees and maximum depth of decision trees are denoted as $n_{cluster}$, $n_{tree}$ and $h_{depth}$, respectively. For LASSO, the $l_1$ regularizer coefficient is denoted with $\alpha$. The bandwidth of the LOWESS algorithm is denoted as $\tau$.

The range of hyperparameters considered for each method mentioned above are given as follows.

- Logistic Regression: $n_{cluster} \in \{1, 2, 4, 8\}$.
- Decision tree: $n_{cluster} \in \{1, 2, 4, 8\}, h_{depth} \in \{5, 10, 15\}$.
- LASSO: $\alpha \in \{0.01, 0.1, 1.0, 5.0\}$.
- Random Forest: $n_{tree} \in \{10, 50, 100, 150\}, h_{depth} \in \{5, 10, 15\}$.
- LOWESS: $\tau \in \{0.01, 0.1, 1, 10\}$.
- BC: $h_n \in \{10, 20, 30, 40, 50\}, l_n \in \{2, 4, 6, 8\}$.
- INVASE: $h_n \in \{20, 30\}, l_n \in \{2, 4, 6\}, \lambda \in \{0.01, 0.1, 0.5\}$.
- iTransplant: $h_n \in \{30, 40, 50\}, \lambda \in \{0.01, 0.05\}, K \in \{10, 20\}, k \in \{4, 8\}$.

These hyperparameters are selected via a grid search with five-fold cross-validation on organ offer data from transplant center with ID code 20522 as listed in Table 4. The hyperparameter selection results are given as follows.

- Logistic Regression: $n_{cluster} = 1$.
- Per-cluster Logistic Regression: $n_{cluster} = 2$.
- Decision tree: $n_{cluster} = 1, h_{depth} = 5$.
- Per-cluster Decision tree: $n_{cluster} = 2, h_{depth} = 5$.
- LASSO: $\alpha = 0.01$.
- Random Forest: $n_{tree} = 150, h_{depth} = 15$.
- LOWESS: $\tau = 1$.
- BC: $h_n = 50, l_n = 4$.
- INVASE: $h_n = 30, l_n = 2, \lambda = 0.01$.
- iTransplant: $h_n = 30, \lambda = 0.01, K = 20, k = 8$.

**Data split and random seeds** For the benchmark and investigative experiments, organ offer data from each transplant center are split to training and test sets with the ratio of $8:2$. For the purpose of early stopping of iTransplant and BC, 20% of organ offers are randomly extracted from the training set to evaluate the action matching performance during training process. In the benchmark, random seed for data split is set to be 40108642, and the models are initialized with ten different integers generated with the random seed of 19260817 to calculate error bars.

**Computing resources and environment** All of the experiment results in this paper are obtained on a CPU of Intel Core™ i5-10210U with maximum RAM of 16 GB. The Python environment is set up under the Windows Subsystem for Linux (WSL 2) with Debian GNU/Linux 10 (buster). The Python version is 3.8.7 with the cpu version of torch 1.9.0. Detailed information can be found in our public code repositories (https://github.com/yvchao/iTransplant and https://github.com/vanderschaarlab/iTransplant).

[4] https://github.com/davidmrau/mixture-of-experts

### A.4 Supplementary experiment results

**Contribution of additional donor organ features** As complementary evidence to the benchmark results reported in the main manuscript, the performances of LASSO, INVASE, BC with different input spaces are compared with iTransplant in Table 5. No significant changes in performance are observed from the benchmark results of BC and INVASE when the match criteria space $\mathcal{M}$ constructed used by iTransplant is replaced by donor organ feature space $\mathcal{O}$. According to Table 5, with full set of organ feature variables in input space $\mathcal{X} \times \mathcal{O}$, the prediction performances of LASSO are largely improved. However, iTransplant still significantly outperforms LASSO even with a subset of donor organ features in the criteria space $\mathcal{M}$.

Table 5: Benchmark results with different input space.

| METHOD | INPUT SPACE | AUC-ROC | AUC-PRC | LL |
|---|---|---|---|---|
| LASSO | $\mathcal{M}$ | 0.777±0.064 | 0.314±0.068 | -0.570±0.045 |
| LASSO | $\mathcal{X} \times \mathcal{M}$ | 0.875±0.053 | 0.439±0.066 | -0.482±0.049 |
| LASSO | $\mathcal{X} \times \mathcal{O}$ | 0.877±0.051 | 0.443±0.064 | -0.478±0.048 |
| INVASE | $\mathcal{M}$ | 0.790±0.062 | 0.341±0.071 | -0.541±0.064 |
| INVASE | $\mathcal{X} \times \mathcal{M}$ | 0.875±0.053 | 0.439±0.066 | -0.482±0.049 |
| INVASE | $\mathcal{X} \times \mathcal{O}$ | 0.877±0.051 | 0.443±0.064 | -0.478±0.048 |
| BEHAVIORAL CLONING | $\mathcal{M}$ | 0.806±0.059 | 0.361±0.073 | -0.515±0.072 |
| BEHAVIORAL CLONING | $\mathcal{X} \times \mathcal{M}$ | 0.899±0.043 | 0.502±0.067 | -0.383±0.067 |
| BEHAVIORAL CLONING | $\mathcal{X} \times \mathcal{O}$ | 0.899±0.042 | 0.505±0.067 | -0.390±0.070 |
| ITRANSPLANT (OURS) | $\mathcal{X} \times \mathcal{M}$ | 0.895±0.045 | 0.502±0.062 | -0.396±0.069 |
| ITRANSPLANT (OURS) | $\mathcal{X} \times \mathcal{O}$ | 0.898±0.048 | 0.508±0.064 | -0.385±0.076 |

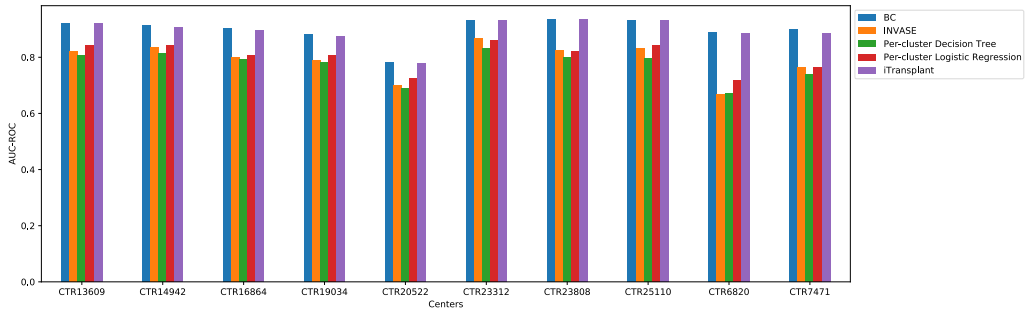
**Comparison with per-cluster logistic regression with additional interaction terms** As shown in Table 6, although including additional pair-wise interaction terms between the match criteria considered in space $\mathcal{M}$ can help to improve the performance of per-cluster logistic regression model, there is still a large gap to the action matching performance achieved by iTransplant. Although the prediction performance of a (per-cluster) logistic regression model is improved slightly with additional interaction terms, a significant gap in performance to iTransplant remains. In comparison, our proposed method, iTransplant, achieves similar performance to the black-box BC model without any interaction terms. Here, we highlight that our method allows clinicians to choose the match criteria considered in space $\mathcal{M}$ based on their expertise, and additional interaction terms can be easily included in the criteria space $\mathcal{M}$.

Table 6: Benchmark with per-cluster logistic regression with interaction terms.

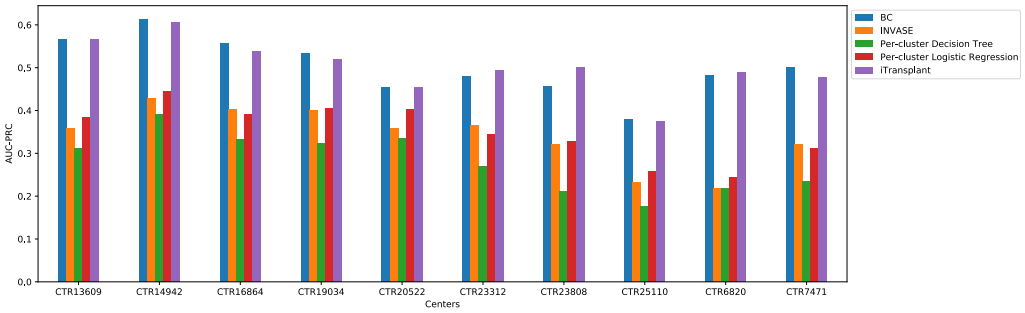
| METHOD | AUC-ROC | AUC-PRC | LL |
|---|---|---|---|
| LOGISTIC REGRESSION (LR) | 0.794±0.054 | 0.341±0.061 | -0.538±0.051 |
| PER-CLUSTER LR | 0.803±0.049 | 0.352±0.063 | -0.527±0.048 |
| PER-CLUSTER LR (WITH INTERACTION TERMS) | 0.824±0.054 | 0.371±0.073 | -0.490±0.063 |
| ITRANSPLANT (OURS) | 0.898±0.048 | 0.508±0.064 | -0.385±0.076 |

**Average performance on individual centers** In addition to the benchmark result in our main manuscript, the average AUC-ROC and AUC-PRC scores of BC, INVASE, per-cluster decision tree, per-cluster logistic regression and iTransplant calculated on a per-center basis are plotted in Figure 7. In all centers, iTransplant effectively reduces the performance gap with (and in some cases surpasses) black-box models (BC and INVASE) while significantly outperforming the white-box models of logistic regression and decision tree, even when compared on a per-cluster basis. This demonstrates the benefit of identifying policy aligned personalized decision policies compared to learning policy-agnostic cluster-wise policies.

**Details of consistency evaluation** For the experiment on consistency, we focus on the twenty criteria in the criteria space $\mathcal{M}$. For perturbations to donor organ features in criteria space, iTransplant




(a) AUC-ROC.




(b) AUC-PRC.

Figure 7: Average performance on organ offer data from ten selected centers.

is innately consistent with its predictions. In the meantime, when patient features related to MELD and MELD-Na scores changes, iTransplant could be inconsistent due to the corresponding deviations of policy signature $\rho$. As mentioned in the main manuscript, we measure the average consistency score over all twenty match criteria, and perturbed samples leading to no changes in match criteria are ignored. The consistency scores of considered methods on 200 organ offers randomly selected from test set of transplant center with ID code 20522 and the average AUC-ROC and AUC-PRC scores on all ten centers are reported in Table 7. According to the consistency scores, the post-hoc interpretations by applying LIME to BC and INVASE provide almost no useful insights under moderate perturbations to input features. In contrast, the white-box part of iTransplant contributes hugely to the consistency score. As reported in Table 7, The consistency of iTransplant can be further improved via a strong

Table 7: Consistency evaluation of interpretations.

| METHOD | CONSISTENCY |
|---|---|
| BC | 46.8% (WITH LIME) |
| INVASE | 41.7% (WITH LIME) |
| ITRANSPLANT, $\lambda = 0.0$ | 91.3% |
| ITRANSPLANT, $\lambda = 0.01$ | 85.1% |
| ITRANSPLANT, $\lambda = 0.1$ | 99.7% |

consistency regularization ($\lambda = 0.1$). However, due to the complex interactions between the policy loss, reconstruction loss and the consistency loss, smaller consistency regularization ($\lambda = 0.01$) may not always lead to an increase in consistency. In practical applications, the trade-offs between consistency and the action-matching performance should be carefully considered.

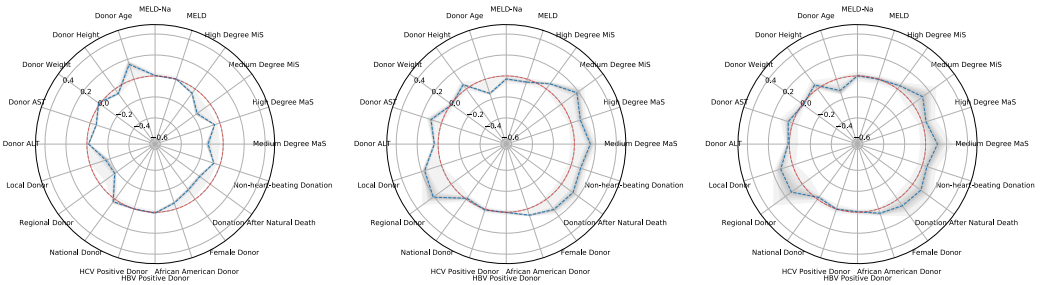
**Performance gain of iTransplant over logistic regression** As discussed in the main manuscript, the patients are clustered into six different clusters based on their latent representation in space $\mathcal{Z}$. With AUC-PRC score as the performance metric, detailed comparison of the prediction performance

of iTransplant and logistic regression in each of these patient clusters is shown in Table 8. We can find that the proposed method iTransplant achieves significant performance gains in all six clusters.

Table 8: Performance (AUC-PRC) gain over logistic regression in each patient cluster.

| CLUSTER | SAMPLE NUMBER | ITRANSPLANT | LOGISTIC REGRESSION |
|---|---|---|---|
| 1 | 588 | **0.152** | 0.133 |
| 2 | 311 | **0.449** | 0.406 |
| 3 | 725 | **0.551** | 0.514 |
| 4 | 614 | **0.032** | 0.025 |
| 5 | 522 | **0.654** | 0.576 |
| 6 | 46 | **0.411** | 0.187 |

**Policy deviations in each patient clusters** As supplementary results of the main manuscript, the policy deviations in the first three patient clusters with respect to the averaged policy $\hat{\pi}_{\bar{\rho}}$ are plotted in Figure 8. It can be found that iTransplant is able to identify personalized organ offer acceptance policies for each patient group, and we argue that this is one of the main reasons for the performance gains observed in Table 8.




(a) Patient cluster 1. (b) Patient cluster 2. (c) Patient cluster 3.

Figure 8: Deviations of policy signature $\Delta\rho = \rho - \bar{\rho}$ in all three clusters.

**Statistics for key patient features in each cluster** Statistics about the distribution of two key patient features (MELD score and HCC status) in different clusters are given in Table 9.

Table 9: Key patient feature distributions in each patient cluster.

| CLUSTER | MELD SCORE | | | | | HCC STATUS | |
|---|---|---|---|---|---|---|---|
| | $[0, 10)$ | $[10, 20)$ | $[20, 30)$ | $[30, 40)$ | $[40, +\infty)$ | POSITIVE | NEGATIVE |
| 1 | 14.8 % | 76.36% | 8.67% | 0.17% | 0.00% | 1.53% | 98.47 % |
| 2 | 47.59% | 51.13% | 1.29% | 0.00% | 0.00% | 62.38% | 37.62% |
| 3 | 0.97% | 65.52% | 27.59% | 4.41% | 1.52% | 0.00% | 100.00% |
| 4 | 21.82% | 67.92% | 9.45% | 0.81% | 0.00% | 0.33% | 99.67% |
| 5 | 3.26% | 51.34% | 33.14% | 9.00% | 3.26% | 0.00% | 100.00% |
| 6 | 0.00% | 2.17% | 84.78% | 6.52% | 6.52% | 0.00% | 100.00% |

**Decision boundaries in all patient clusters** Together with the policy deviations in each cluster, decision boundaries of the in-cluster average policies identified by iTransplant are plotted in Figure 9 (only clusters with no less than 30 accepted organ offers are considered). We can find that iTransplant can properly separate the positive and negative samples with the personalized policies identified in each cluster. We further demonstrate the benefit of personalized policy learning with the comparison of decision boundaries provided by a logistic regression model and iTransplant for patient cluster

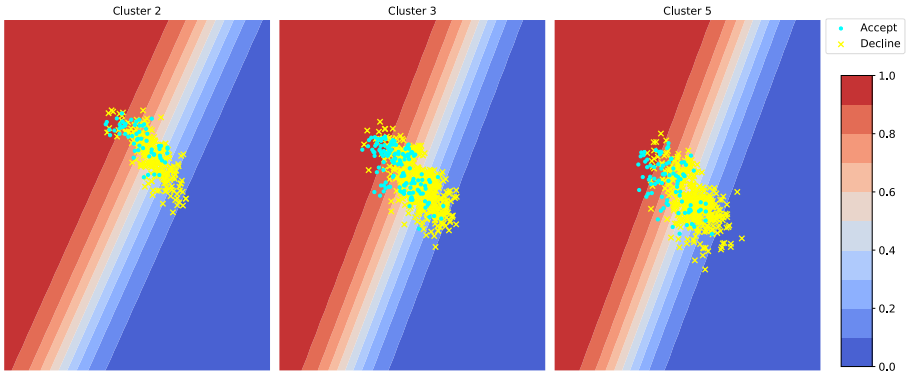



Figure 9: Decision boundaries in considered clusters. The probabilities of organ offer acceptance given by iTransplant are plotted in the background with colors of red and blue for high and low likelihood of offer acceptance, respectively. Organ offers for each patient clusters are projected to the hyperplanes determined by the averaged policy signatures $\rho$ of the policies given by logistic regression and iTransplant in policy space $\Pi$. For each cluster, declined organ offers are marked with yellow crosses while accepted offers are marked with cyan dots.

2 in Figure 10. The optimal decision boundary for positive and negative samples in the projection hyperplane is very close to the one from iTransplant while a logistic regression model is unable to provide customized decision boundaries for patients in different subgroups, which explains the performance gaps observed in Table 8.

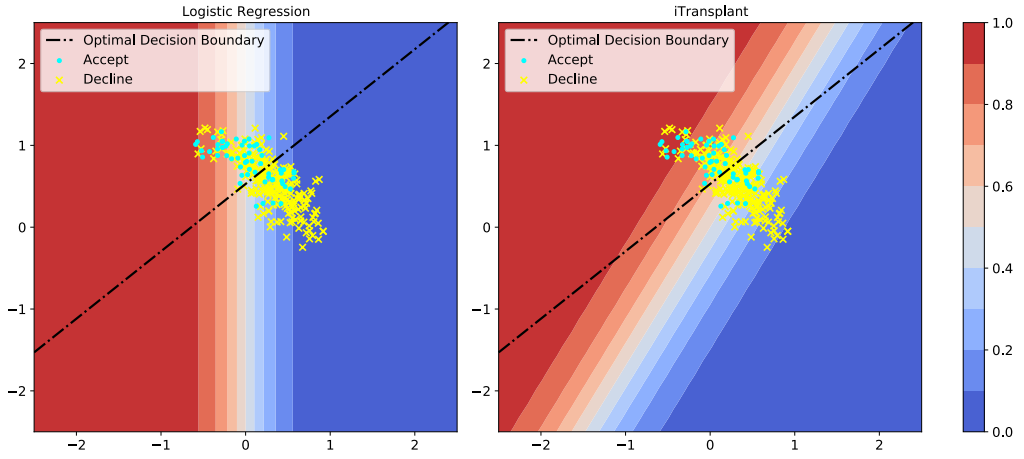



Figure 10: Decision boundaries in patient cluster 2. (L) Logistic regression, (R) iTransplant. The optimal decision boundary in the projection hyperplane is obtained via the support vector machine algorithm and is plotted as black dotted lines.

### A.5 Further discussion on consistency

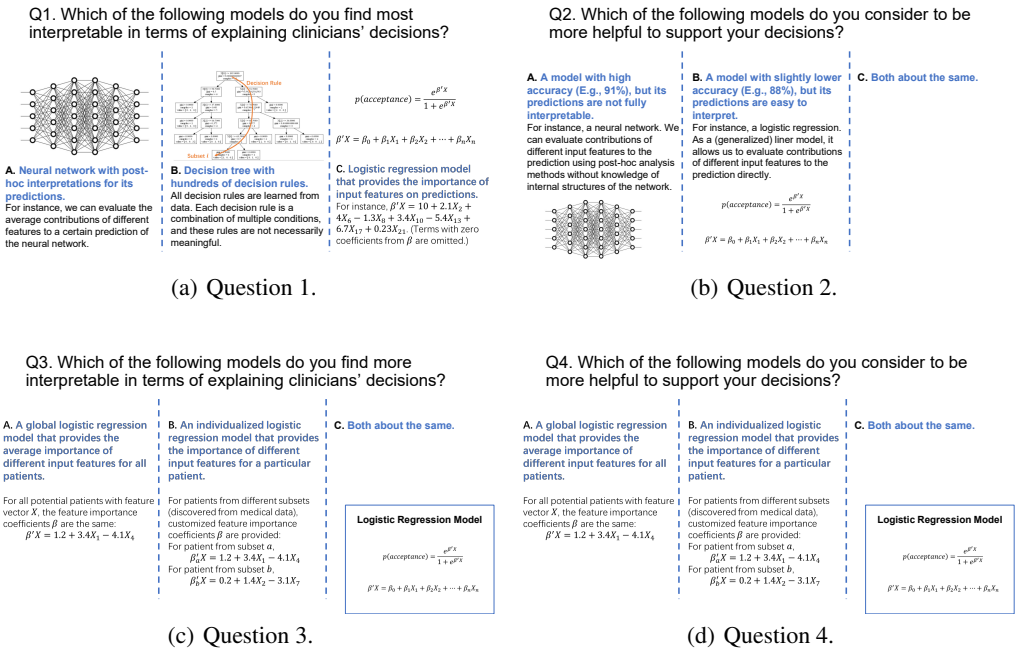
**Importance of consistency** Consistency is proposed as one of the key desiderata of practical inverse decision-making approaches in this paper. For the insights from inverse decision-making methods to be helpful for decision makers, the interpretations need to be consistent under moderate perturbations in input space. Practically, this means that if a feature has a positive weight, a moderate increase to the value of the feature should increase the probability of acceptance (and for a binary feature with positive weight, flipping the feature should reduce the probability of acceptance). Without this

property, counterfactual analysis cannot be reliably performed, limiting the value of any interpretation or understanding gained from inverse decision-making.

**Consistency and non-smooth preferences over patient features** Clinicians' preferences over donor organs could be non-smooth functions over patient features, which may lead to potential failure of methods (like logistic regression) that issues extreme consistent interpretations to clinicians' decisions. In this paper, we requires the consistency between latent space representations of patient features and policies, rather than consistency directly between the patient features and policies. Thus, despite having a non-zero consistency regularization term defined in equation (2) of our main manuscript, altering patient features could still drastically change the latent representation, thus changing the policy significantly, which is the desired outcome in our method. In addition, we note that the logit model in equation (1) allows a single criterion in the criteria space $\mathcal{M}$ to flip the decision and thus our modeling framework is able to sufficiently accommodate the non-smooth preferences scenarios.

### A.6 A short survey on interpretability of machine learning methods

We have conducted a short survey with six clinicians in an attempt to further validate the usefulness of our method. After a brief introduction to decision tree, logistic regression and neural network models, four questions as shown in Figure 11 were shown to each clinician individually, and the surveyed clinicians are asked to give their feedback based on their own understanding of these questions. The survey was conducted blindly, i.e., the clinicians were given no context as pertains to our proposed method.




(a) Question 1.

(b) Question 2.

(c) Question 3.

(d) Question 4.

Figure 11: Four questions in the small survey on interpretability.

The results of our survey are shown in Table 10. Four out of the six surveyed clinicians consider logistic regression to be more or equally helpful compared to a high-precision black-box model. In addition, only one surveyed clinician considers the post-hoc interpretations of neural network-based models to be more interpretable than white-box models like decision tree and logistic regression. Notably, all six clinicians deemed individualized logistic regression models more informative and practical than a general logistic regression model. Despite the relatively small sample size, we believe that these results provide evidence of the importance of interpretability in our proposed approach and

the usefulness of our method compared to alternate forms of interpretability, including basic logistic regression models.

Table 10: Responses from six surveyed clinicians.

| CLINICIAN | QUESTION 1 | QUESTION 2 | QUESTION 3 | QUESTION 4 |
|---|---|---|---|---|
| CLINICIAN 1 | B,C | B | B | B |
| CLINICIAN 2 | B | B | B | B |
| CLINICIAN 3 | C | B | B | B |
| CLINICIAN 4 | B,C | C | B | B |
| CLINICIAN 5 | A | A | B | B |
| CLINICIAN 6 | B | A | B | B |

### A.7 Limitations of our work

While our experiments demonstrate that most decisions can be effectively explained by iTransplant with clinical features recorded in observational datasets, some decisions were still likely to be made based on factors that were not considered in this paper. It is therefore important to explore the impact of including additional match criteria and contextual information concerning unit activity and outcomes in iTransplant for practical applications. In the meantime, the performance gains of iTransplant comes from the black-box policy selector model. Although the policies identified for individual patients are fully human comprehensible, for real-world applications, it is necessary to build the trust of clinicians on the black-box part of iTransplant via properly designed mechanisms for human-machine interactions.

We would like to discuss some additional relevant limitations as follows.

**Misinterpretation of correlations** The insights provided by our method could be used to positively impact future organ transplant practices. However, we need to highlight that the decision drivers discovered by iTransplant only signify potential correlations between certain criteria and real decisions on organ offers. Such correlations between features in the criteria space and decisions identified by our method should not be interpreted as causal, and care must be taken when interpreting the output of any inverse decision-making approach, including our method.

**Impact from organ offering policies** Currently, specific organs are offered to clinicians for a specific recipient, with that offer being predicated on nationally agreed guidelines. A clinician will accept or reject the offer based on their knowledge, experience and discussion with the recipient. Nevertheless, transplant centers have very variable acceptance rates of offered organs, with some consistently accepting a higher percentage than others. Hence, indeed, the organ offering policy determines which organ offers clinicians must make decisions on, which ultimately shapes the observed distribution of organ offers in the data. However, the goal of our method is precisely to understand the observed decisions of clinicians—i.e., given the organ allocation policy in the real transplantation system.

To this end, we seek interpretable parameterizations of the drivers of human decisions that are predictive of these observed decisions. Of course, these parameterizations are only valid for the observed data (i.e., "in-distribution" prediction), and should not be used in scenarios where the organ offering policy is significantly different (i.e., "out-of-distribution" prediction). In this sense, this caveat is no different from any inverse decision-making or supervised learning setting—a change in domain will generally degrade the analysis.

**No hidden confounders** In our method, it is assumed that drivers of observed decisions in the dataset are fully observable or can be inferred from observable features in the dataset. However, as a limitation of many inverse decision-making approaches, it is likely that some decisions were made based on factors that were not included in the observational dataset, which may bias the insights discovered by our method and lead to lower prediction performance.

**Confirmation bias** Any endeavor aimed at "pattern discover" will run the risk of confirmation bias. In most realistic use cases, there will be no absolute "ground truth" available for validation, and

our inverse decision-making approach is not immune to this limitation. Importantly, however, we emphasize that we are not in the business of performing statistical inference or testing (whence the formal quantification of "statistical significance" would require rigid procedures to ensure validity). To the contrary, we are engaged in the orthogonal mission of generating interpretable hypotheses on the underlying drivers of human decisions—in particular, drivers that are highly predictive of observed decisions. The purpose of the investigative experiments is to illustrate that our method is capable of proposing hypothetical decision policies that could be matched with real clinical findings or guidelines, which makes our method useful for clinicians to review and improve their decisions. In the meantime, for practical applications of our method, validation and additional analysis by domain experts is necessary to validate the discovered insights.

**Linearity in the criteria space** The assumption of linearity in the criteria space (or the reward structure) is common in inverse decision-making literature that aims to identify interpretable decision policies from the observable decision history of an agent. We take the same assumption of linearity to obtain human comprehensible decision policies. However, the validity of this assumption strongly depends on the selection of feature maps (criteria) in the criteria space. Misspecification of the criteria space may lead to misleading interpretations and poor performance of inverse decision-making methods. Thus, careful validation by domain experts is necessary for practical applications.

**Greedy decision policy** The OPTN database describes donor organ-recipient pairs, where the organ was offered according to national guidelines to an individual recipient, and accepted or declined by a clinician. The decision to accept or decline was not informed by knowledge of the whole waitlist of potential recipients nor the future availability of donor organs. In line with this, we take the assumption that the decision sequences of clinicians are greedy and not purposeful (accounting for outcomes of future decisions). The assumption of greedy decision policy is necessary for modeling the organ offer acceptance as a one-step decision-making problem and was deemed reasonable in our discussions with clinicians. However, this assumption blocks the application of our approach on sequential decision-making settings where the current decision is affected by potential outcomes at future states. We leave the extension of iTransplant to such settings for our future work.

**Missing data** Missing data is an important issue for practical applications of machine learning. That said, our proposed method tackles a problem orthogonal to that of missing data. To be clear, neither our method nor any of the included benchmarks have any "built-in" capabilities designed for missing data. However, in practice our method (and benchmarks) is compatible with any existing data imputation algorithms.

### A.8 Potential applications

As mentioned in the main manuscript and the Appendix, our model is proposed as a tool to help clinicians to review and identify potential drivers of their clinical decisions. As described in the introduction, clinical decision-making in organ transplantation is poorly understood with a high proportion of organ offers declined and substantial variation in clinical practice. The intended use of a tool such as iTransplant would be to stand beside the clinician decision-makers so that they can reflect upon and better understand the drivers for their decision. These issues are not unique to liver transplantation, and variation in practice has been studied across medicine [2, 12], including cancer [42, 37] and intensive care [11].

As illustrated in our experiments, our method can be used to study and gain greater understanding of these clinical phenomena, and the insights obtained via our method could ultimately be used to improve future decisions. It is worth noting that the criteria space in our method is not limited to the one discussed in our paper. In fact, clinicians could use iTransplant to generate highly predictive hypotheses for clinical decision-making with respect to any set of criteria.

Our long-term goal is to help clinicians review past clinical decisions with the insights from iTransplant and to improve future decisions of clinicians and suggestions from decision support tools. However, the analyses and experiments in our paper are mainly used to illustrate the potential usage of our method as a tool of understanding human decisions.

**Interpretation of the normalized policy signature** As illustrated in Fig. 3 in our main text, the policy signatures identified by iTransplant indicate the relative importance (weights) of different

criteria in explaining the observed decisions. These weights can be used to discover the criteria or features that are correlated with clinicians' decisions in the observational dataset and can be interpreted as a set of potential drivers of human decisions. However, such correlations identified by our method should not be interpreted as causal, and validation by domain experts is always necessary for practical applications of inverse decision-making approaches, including our method.

**Interpretation of patient clusters** The clustering of patients in the latent space learned by iTransplant is another important result of our method. By adopting the consistency loss in equation (2), similar latent representations of patients are encouraged to yield similar decision policies over the criteria space. Thus, clustering in the latent space can help us to group similar patients and explore correlations between patient features and certain decision drivers, which can provide insights into the observed decisions. However, we would like to emphasize that the insights discovered by our method should not be interpreted as causal, and careful validation by domain experts is always necessary.

### A.9 Potential negative social impacts

The proposed method iTransplant attempts to understand clinical decisions and does not directly recommend actions to clinicians. However, inaccuracies in the interpretations of iTransplant (or any inverse decision-making method) without proper audit could lead to negative consequences if they lead to changes in decision-making that adversely impacted outcomes, which should be taken into account in any practical use of such systems. For instance, misinterpreting the insights from our method as causal relationships between criteria and decision policies could lead to negative impacts on transplant decisions. Validation and additional analysis by domain experts is needed for practical application of our method and inverse decision-making methods more broadly.

Secondly, when patient or donor organ features related to (potentially) sensitive characteristics (e.g. gender or ethnicity) are included in the criteria space, analyses of biases learned by iTransplant are necessary in practice to avoid potential consequences on equality, fairness, etc. In addition, as a limitation of many inverse decision-making approaches, it is likely that some decisions were made based on factors that are not included in the observational dataset, which may bias the insights discovered by our method, while misspecification of the criteria space may also lead to incorrect interpretations.