Learning Queueing Policies for Organ Transplantation Allocation using *Interpretable* Counterfactual Survival Analysis

Jeroen Berrevoets [1] **Ahmed M. Alaa** [2] **Zhaozhi Qian** [1]
James Jordon [3] **Alexander Gimson** [4] **Mihaela van der Schaar** [1 2 5]

Abstract

Organ transplantation is often the last resort for treating end-stage illnesses, but managing transplant wait-lists is challenging because of organ scarcity and the complexity of assessing donor-recipient compatibility. In this paper, we develop a data-driven model for (real-time) organ allocation using observational data for transplant outcomes. Our model integrates a *queuing-theoretic* framework with unsupervised learning to cluster the organs into "organ types", and then construct *priority queues* (associated with each organ type) wherein incoming patients are assigned. To reason about organ allocations, the model uses *synthetic controls* to infer a patient's survival outcomes under counterfactual allocations to the different organ types—the model is trained end-to-end to optimize the trade-off between patient *waiting time* and expected *survival time*. The usage of synthetic controls enable patient-level interpretations of allocation decisions that can be presented and understood by clinicians. We test our model on multiple data sets, and show that it outperforms other organ-allocation policies in terms of added life-years, and death count. Furthermore, we introduce a novel organ-allocation simulator to accurately test new policies.

1. Introduction

Over the years, organ transplantation surgeries have become increasingly prevalent in the developed world—currently, a new patient is added to the transplant wait-list every nine minutes in the US (Organ Procurement and Transplantation Network (OPTN), 2021b). A major reason for this is the recent improvement in case selection, anaesthesia, surgical techniques, immunosupression strategies, and expanded donor supply. Allocation strategies used to suggest donor organs to patients, however, have not improved as much. For example, liver allocation in the US and Europe rely on the MELD score which is calculated using only three lab parameters[1] (Organ Procurement and Transplantation Network (OPTN), 2021a; Eurotransplant Reference Laboratory (ETRL), 2020); in the UK, allocation relies on simple linear models that vastly underestimate the complexity of organ-to-patient interaction (Neuberger et al., 2008; NHSBT Liver Advisory Group (LAG), 2019). In addition, these allocation protocols only score patient-organ compatibility, but do not provide a systematic approach to managing the influx of patients and organs arriving into the system over time.

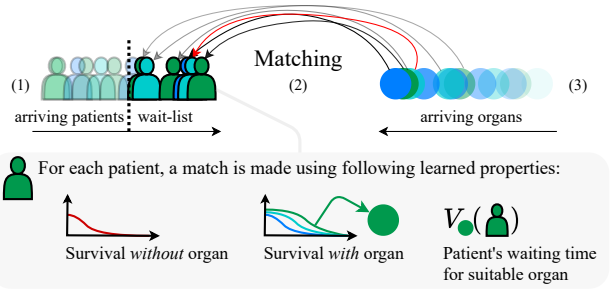

Figure 1. **Overview of transplant system and OrganSync.** Various patients arrive at seemingly irregular intervals (1); and various organs arrive at equally irregular intervals (3). Organ-allocation algorithms match patients on the wait-list (right of the dashed line), with arriving organs (2). *OrganSync* not only models (2), but also (1) and (3). For this, OrganSync estimates each patient's survival without an organ, predicts which organ-type would be most beneficial, and compares the patient's survival time with the waiting time for their suitable organ-type. Indicated by the red arrow; a patient is kept on the wait-list (despite arriving second), as our model anticipates a suitable organ arriving in the future.

Organ allocation is a hard problem. In this paper, we pro-

[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK [2]Department of Electrical Engineering, University of California, Los Angeles (UCLA), Los Angeles CA, USA [3]Department of Engineering Science, University of Oxford, Oxford, UK [4]Cambridge University Hospitals, Cambridge, UK [5]The Alan Turing Institute, London, UK. Correspondence to: JB <jeroen.berrevoets@maths.cam.ac.uk>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

[1]International Normalised Ratio, Creatinine, and Bilirubin; USA also includes Sodium.

vide a first formulation for the problem of allocating organs to recipient patients over time. We identify four challenges associated with this problem: (i) an organ is unique and can only be assigned to a single patient, hence there are many unobserved counterfactual combinations of patients and organs that we do not observe in the data. The difficulty does not end there, as the available data suffers from selection bias since organs are not distributed randomly (they are distributed according to inplace assignment policies); (ii) organ and patient interaction is highly complex, yet outcome prediction for decision support requires interpretability, making a naive non-linear model unfit for the task; (iii) allocating an organ to a single individual affects all others on the wait-list as their waiting is inevitably prolonged. As such, it is important to take into account not only outcome prediction, but also every patient's maximum estimated waiting time until premature death; and (iv), there are many valid objectives in organ-allocation, all boiling down to who is prioritised over whom. An allocation policy should be able to easily adapt to other objectives than, for example, population life years (Gimson, 2020).

Contributions. To address these challenges, we introduce *OrganSync*; a novel decision support system for organ allocation. With OrganSync, we innovate in two ways: by introducing an *interpretable high-dimensional potential outcomes estimator*, addressing challenge (i) and (ii); and we integrate its predictions into a *queueing-theoretic* framework, addressing challenge (iii) and (iv). Furthermore, we introduce a simulation to evaluate organ-allocation policies—accurately testing organ-allocation is non-trivial as tested policies will almost immediately deviate from data that is collected under an alternative policy.

Using machine learning, allocation has improved in terms of added life-years (Yoon et al., 2017; Berrevoets et al., 2020). However, such models lack interpretability, limiting their adoption. Furthermore, these models (including those currently in use) rely heavily on ranking in function of the available organ, offering little perspective for patients on the wait-list as their future rank is essentially random due to the randomness of available organs; i.e. once on the wait-list, a patient has no idea how long they will have to wait, often in agony. We argue that machine learning should *aid* allocations polices, and work in tandem with queueing theory to provide a fixed yet accurate ranking.

2. Problem formulation

Conceptually, a transplant system involves three components that interact and evolve in *real-time*: the patients arriving and leaving the wait-list, the organs arriving for transplant, and a policy that decides which patient to receive the organ (an overview of this is provided in Figure 1).

Notation. Let $\mathcal{X} \subset \mathbb{R}^d$ denote the space of all possible patients, and let $\mathcal{O} \subset \mathbb{R}^e$ denote the space of all possible organs. Let $\mathbf{X} \in \mathcal{X}$ denote the feature vector of a patient and $\mathbf{O} \in \mathcal{O}$ the feature vector of an organ. We assume we have an observational dataset containing N patients, $\mathbf{X}_1, ..., \mathbf{X}_N$. Each patient either received an organ $\mathbf{O}_i$ (unique to each patient) or did not receive an organ, for which we will slightly abuse notation and write $\mathbf{O}_i = \emptyset$. Each patient has an associated survival time $Y_i + W_i$ that breaks down into their waiting time $W_i \in \mathbb{R}_+$, indicating the time spent on the wait-list (until they either receive an organ or die while waiting), and their survival post-transplant $Y_i \in \mathbb{R}_+$ (which is 0 if the patient did not receive an organ). In addition, some patients are censored, where we write $\delta_i = 1$ if the patient died and $\delta_i = 0$ if they were censored. Note that censoring can happen both during W_i and during Y_i. Together, these create a dataset $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{O}_i, Y_i, W_i, \delta_i) : i = 1, ..., N\}$.

We assume that the survival times, Y_i, are generated according to the Neyman-Rubin potential outcomes framework (Rubin, 2005) so that associated with each patient, we assume there is a set of potential outcomes $\{Y(\mathbf{o}) : \mathbf{o} \in \mathcal{O}\}$ and that the observed survival time $Y_i = Y(\mathbf{O}_i)$ is consistent with the potential outcome for the observed organ.

Our goal is to perform inference in a live setting with streams of both patients and organs. At discrete times, $t = 1, 2, ...$, we denote by $\mathcal{Q}(t) = \{\mathbf{X}_j : j = 1, ..., n(t)\} \subset \mathcal{X}$ the patients on the transplant wait-list[2]. At each time step, an organ, $\mathbf{O}(t)$, arrives and must be assigned to a patient on the wait-list.

Goal. We wish to define an allocation-policy,

$$\pi(t) = \pi(\mathbf{O}(t), \mathcal{Q}(t)) = j \in \{1, ..., n(t)\}, \tag{1}$$

that maximises the population life-years given by

$$\max_{\pi} \lim_{t\to\infty} \frac{1}{|\mathcal{X}(t)|} \sum_{\mathbf{X}\in\mathcal{X}(t)} \mathbb{E}_{\pi,\mathcal{O}(t)}\Big[W + Y|\mathbf{X}\Big] \tag{2}$$

where $\mathcal{X}(t) = \bigcup_{s\le t} \mathcal{Q}(s)$ (i.e. all patients that have ever been in the system) and $\mathcal{O}(t) = \{\mathbf{O}(1), ..., \mathbf{O}(t)\}$ (i.e. the sequence of organs observed in the system).

In order to optimise Eq.(2), a policy, π, must not only consider the immediate benefit that a current organ has for each individual in the wait-list, but also the potential future organs that will become available, and how long each person on the wait-list can survive without an organ. Importantly, we consider the setting in which organs are perishable (e.g. heart, liver) and as such it never makes sense to wait with an organ for a future patient that is not yet in the queue and might be a "better" fit - we therefore do not model the arrival of future patients.

[2]Note that the index j does not correspond to the dataset index i in any way - these are new patients

Therefore at each time-step, t, we require the following: *(i) A good estimate of survival given an organ,* $\mathbb{E}[Y(\mathbf{O}(t))|\mathbf{X}]$*, for each* $\mathbf{X} \in \mathcal{Q}(t)$. In order to obtain this estimate using the censored data we have available in $\mathcal{D}$ (indicated by δ_i), we model $Y(\mathbf{O})$ through *survival analysis*, with a patient's conditional survival function,

$$S(y|\mathbf{X}, \mathbf{O}) = p(Y(\mathbf{O}) > y|\mathbf{X}), \ \forall y \in \mathbb{R}_+. \tag{3}$$

(ii) A good estimate of survival without an organ, $\mathbb{E}[W|\mathbf{X}, \mathbf{O} = \emptyset, \delta = 1]$*, for each* $\mathbf{X} \in \mathcal{Q}(t)$. As with our estimate for the survival with organ, we use survival analysis to leverage the censored dataset to compute the expected survival without an organ.

(iii) An approximation of the arrival distribution of future organs. To obtain such an approximation, we reduce the problem to a system with multiple queues and leverage Little's law (Little, 1961) in order to obtain wait-times for specific organ "types".

To our knowledge, OrganSync is the only method that estimates and uses these elements to optimise Eq.(2) through leveraging knowledge gained from each.

3. OrganSync

OrganSync weighs three different components: (i) patient survival with the current organ; (ii) patient survival without any organ; and (iii) an approximation of the arrival distribution of future organs. For a policy to be truly useful, these components have to be learned in an interpretable way (Doshi-Velez & Kim, 2017).

Overview. OrganSync operates in five major steps: [*Step 1*], we learn a non-linear representation of the patient-organ space; [*Step 2*], in this representation space we compose a patient-organ pair's synthetic control; [*Step 3*], using the synthetic control, we compute a patient's survival; [*Step 4*], with the survival estimates, we search for a patient's optimal organ-type; and [*Step 5*], with the priority and optimal organ-type, we assign the patient to one of K priority queues. Each step is explained in detail below and annotated in Figure 2.

3.1. Interpretable survival analysis (Steps 1, 2, and 3).

For each patient, we wish to make two survival estimates: *without* an organ and *with* an organ. This is a challenging problem due to selection bias: simply regressing Y on $\mathcal{X} \times \mathcal{O}$ and W on $\mathcal{X}$ will yield biased estimates due to past applied policies. While Berrevoets et al. (2020) solve this problem using adversarially balancing a representation, we identify two shortcomings of their method which we solve in this paper: (i) our method is interpretable as it is explicitly based on past observations through composition of a synthetic control (Abadie et al., 2015); and (ii) by utilising survival curves instead of a point estimate we leverage censored data.

In what follows, we outline how we build up our estimate for Y, and our estimate for W is done in the same way, without a dependence on $\mathbf{O}$.

[*Step 1*] - A non-linear representation. Let $f : \mathcal{X} \times \mathcal{O} \to \mathcal{U}$ be a non-linear representation of the patient-organ pair with $\mathbf{u} \in \mathcal{U} \subset \mathbb{R}^h$, and let,

$$\hat{Y}(\mathbf{O}) = \beta^\top \mathbf{u} + \epsilon, \tag{4}$$

where $\beta \in \mathbb{R}^h$ are latent coefficients, and ϵ is a noise term.

We learn f and β using supervised learning through Eq.(4), but this comes with a limitation: f and β are not interpretable. Rather than performing inference through Eq.(4), we will use f to compose a synthetic patient-organ pair from other pairs in $\mathcal{D}$.

[*Step 2*] - Synthetic patient-organ pairs. Let $(\mathbf{X}, \mathbf{O})$ be a patient-organ pair for which we want to estimate $Y(\mathbf{O})$. While using Eq.(4) might yield an accurate point-estimate, it remains non-interpretable, nor will it help us in constructing a survival estimate like in Eq.(3)— as is required (outlined in Section 2). Instead, we construct a *similar* synthetic pair in the representation space:

$$\tilde{\mathbf{u}} = \mathbf{U}\mathbf{a} \tag{5}$$

where $\mathbf{U} \in \mathbb{R}^{h \times N}$ is a matrix comprised of the non-linear representations of the patient-organ pairs in $\mathcal{D}$, and $\mathbf{a} = \mathbf{a}(\mathbf{X}, \mathbf{O}) \in [0, 1]^N$ is a vector used to build a convex combination of various patient-organ pairs in the data. We constrain $\mathbf{a}$ in three ways: (i) the sum of elements in $\mathbf{a}$ equal one ($\sum_m a_m = 1$); (ii) $\mathbf{a}$ should be sparse such that $\tilde{\mathbf{u}}$ is interpretable as a composition of only a few past observations; and (iii) each element in $\mathbf{a}$ is positive. Note that $\mathbf{a}$ is a function of the patient-organ pair $(\mathbf{X}, \mathbf{O})$ according to

$$\mathbf{a}(\mathbf{X}, \mathbf{O}) = \min_{\mathbf{a} \in [0,1]^N} \|\mathbf{U}\mathbf{a} - f(\mathbf{X}, \mathbf{O})\|_2^2 + \lambda \|\mathbf{a}\|_1, \tag{6}$$

where the second term corresponds to an L_1 regulariser governed by $\lambda \in [0, 1]$, encouraging sparsity (Tibshirani, 1996). As per Eq.(4), $\hat{Y}(\mathbf{O})$ relates linearly to $\mathbf{u}$, allowing $\hat{Y}(\mathbf{O})$ to be estimated as,

$$\hat{Y}(\mathbf{O}) = \mathbf{Y}^\top \mathbf{a} \tag{7}$$

with $\mathbf{Y}$ the vector of outcomes in $\mathcal{D}$ (Abadie & Gardeazabal, 2003; Abadie et al., 2015). Elements in $\mathbf{a}$ can be interpreted as percentages, due to the constraints discussed above, allowing us to know exactly which past seen cases an estimate is based upon, and by how much.

We have illustrated our estimator in the leftmost part of Figure 2, where: the patient-organ space, $\mathcal{X} \times \mathcal{O}$ is mapped to a non-linear representation space, $\mathcal{U}$, in such a way that it relates linearly to Y; and the patient is then combined synthetically to possible organs in $\mathcal{O}$ from which we predict

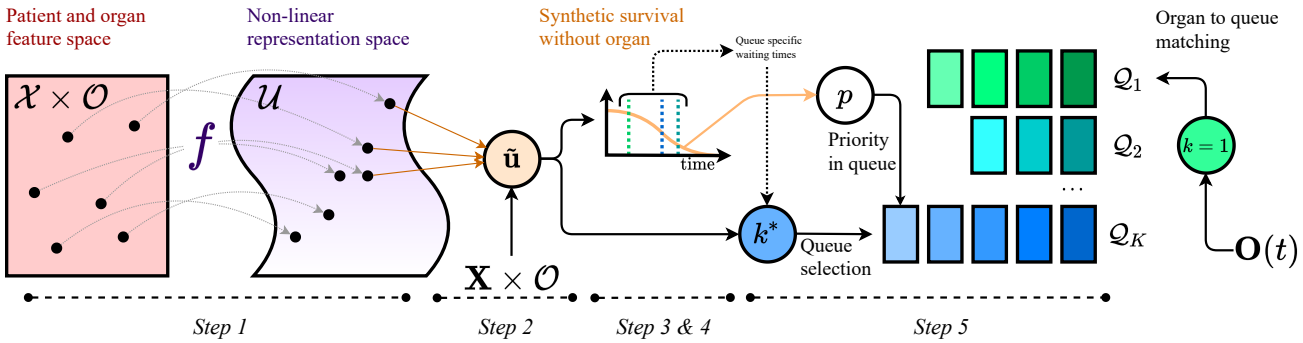

Figure 2. **Detailed overview of OrganSync.** Whenever a patient enters the transplant system, we estimate their survival without organ using a synthetic patient-organ pair, and similarly estimate which organ-type would yield the highest survival. To cope with complex organ-to-patient interaction, we build the synthetic pairs in a non-linear representation space. Using the patient's synthetic survival without an organ, we select an appropriate queue according to their optimal organ-type, and the estimated waiting time; as well as, assign priority within the chosen queue. When a new organ arrives it is predicted a type, and is granted to the first-in-line of the organ type's queue.

an optimal organ-type, as well as patient survival, discussed next.

[*Step 3*] - Survival estimates. To estimate the patient's counterfactual survival under a different organ allocation, we use a Kaplan-Meijer survival estimate (Chen, 2020; Beran, 1981), where our dataset is based on the patient-organ pairs that compose ones synthetic control. Specifically, let $\mathcal{I}(\mathbf{a}) \subset \{1, ..., N\}$ be the non-zero elements in $\mathbf{a}$, which in practice will correspond to elements higher than some small threshold. A conditional survival estimate is then given by the Kaplan-Meijer survival estimate:

$$\hat{S}(y|\mathbf{X}, \mathbf{O}) = \prod_{t<y} \left(1 - \frac{d_{\mathbf{a}}(t)}{n_{\mathbf{a}}(t)}\right), \tag{8}$$

with, $d_{\mathbf{a}}(t) = \sum_{j \in \mathcal{I}(\mathbf{a})} \delta_j \mathbb{I}\{Y_j \leq t\}$ the amount of deaths up to t, and $n_{\mathbf{a}}(t) = \sum_{j \in \mathcal{I}(\mathbf{a})} \delta_j \mathbb{I}\{Y_j > t\} + \kappa$ the amount of patients seen up to t, where $\kappa > 0$ is a small constant to avoid dividing by zero in Eq.(8). Basing our survival estimates on the synthetic control results in interpretable estimates. Furthermore, the *similarity* of patient-organ pairs is calculated in the non-linear representation space while the survival estimates remain non-parametric. These estimates determine priority in $\mathcal{Q}$, discussed in Section 3.2.

Assumptions. As is standard in causal literature on potential outcomes, we make the following assumptions (Hirano & Imbens, 2004). Our supplemental material contains a discussion on these assumptions.

Assumption 1 (Overlap.)**.** *For all* $\mathbf{X} \in \mathcal{X}$ *and all* $\mathbf{O} \in \mathcal{O}$ *that have positive probability of occurring,* $0 < p(\mathbf{O}|\mathbf{X}) < 1$.

Assumption 2 (Unconfoundedness.)**.** *Conditional on* $\mathbf{X}$*, the full set of potential outcomes are independent of the treatment,* $\mathbf{O} \perp\!\!\!\perp \{Y(\mathbf{o}) : \mathbf{o} \in \mathcal{O}\} \mid \mathbf{X}_i$.

Furthermore, through Eq.(4) we make an implicit assumption that there exists a representation $\mathbf{u}$ from which the outcome is a linear combination. Note that this is a very general assumption as, in practice, it merely warrants the use of a neural network with a fully connected output layer.

3.2. Organising the patients in queues (Steps 4 and 5).

We now turn to allocating organs to patients. Contrasting existing allocation-policies— recent (Berrevoets et al., 2020; Yoon et al., 2017) and less recent (Malinchoc et al., 2000; Kim et al., 2008; Neuberger et al., 2008) —we resort to queueing theory, rather than ranking.

[*Step 4*] - Queue assembly. In order to optimise our goal of total life years, we need to model the arrival of organs. The consideration of what organs will arrive in the future is crucial to the problem of who to assign the current organ to. Unfortunately, modelling the arrival distribution of the high-dimensional organ space is difficult. To address this difficulty we introduce K *queues*, $\mathcal{Q}_1, ..., \mathcal{Q}_K$, that correspond to K *clusters* of organs. In doing so, we reduce the problem of estimating the full organ arrival process, to estimating the arrival process of k distinct "types" of organs. In doing so we introduce a trade-off, a larger K leaders to better estimates of organ-recipient outcomes *for a specific organ*, but the estimation of the arrival process becomes worse. In particular, as we take K to infinity, we recover the uniqueness of every organ; each queue corresponds to a specific organ in the space $\mathcal{O}$.

We assign each patient to one of the K clusters, by estimating their survival given each of the K cluster-centers and comparing their survival without an organ to the estimated waiting time associated with each cluster $V_k(\mathbf{X})$, $k \in \{1, \ldots, K\}$.

Table 1. **Various allocation policies in transplantation.** OrganSync satisfies our key criteria: (1) use survival analysis; (2) estimate potential outcomes; (3) provide interpretable suggestions; (4) have waiting time estimates. Data, estimands, processing, and outcomes are denoted $\mathcal{D}$, e, p, i, respectively; text above the arrow specifies how the method arrives at each node. Let $\tau = Y(\mathbf{O}) - Y(\emptyset)$.

Method	Reference	Overview	(1)	(2)	(3)	(4)
FIFO	n.a.	$\mathcal{D} \xrightarrow{\max\{\text{waiting time}\}} i$	✗	✗	✓	✓
MELD	(Malinchoc et al., 2000)	$\mathcal{D} \xrightarrow{\propto W_i} e \xrightarrow{\max\{e\}} i$	✗	✗	✓	✗
MELD-na	(Kim et al., 2008)	$\mathcal{D} \xrightarrow{\propto W_i} e \xrightarrow{\max\{e\}} i$	✗	✗	✓	✗
TransplantBenefit	(Neuberger et al., 2008)	$\mathcal{D} \xrightarrow{\tau} e \xrightarrow{\max\{e\}} i$	✓	✓	✓	✗
ConfidentMatch	(Yoon et al., 2017)	$\mathcal{D} \xrightarrow{Y(\mathbf{O})} e \xrightarrow{\max\{e\}} i$	✗	✗	✗	✗
OrganITE	(Berrevoets et al., 2020)	$\mathcal{D} \xrightarrow{\tau \;\&\; p(O)} e \xrightarrow{\text{weight predictions}} p \xrightarrow{\max\{p\}} i$	✓	✓	✗	✗
OrganSync	(Ours)	$\mathcal{D} \xrightarrow{\hat{S} \;\&\; \rho_k} e \xrightarrow{\text{predict } k^* \text{ and } V_k(\mathbf{X})} p \xrightarrow{\text{select } Q_k} i$	✓	✓	✓	✓

Using Little's law (Little, 1961) we estimate $V_k(\mathbf{X})$ as:

$$V_k(\mathbf{X}) = \frac{L_k(p(\mathbf{X}))}{\rho_k}, \tag{9}$$

where ρ_k is the arrival rate of patients into Q_k; and $L_k(p(\mathbf{X}))$ is the number of patients with higher priority than $\mathbf{X}$ in Q_k. We define the priority, p, in terms of a patient's survival without organ,

$$p(\mathbf{X}) = \left(\int_w \hat{S}(w|\mathbf{X}, \emptyset) dw \right)^{-1}. \tag{10}$$

We estimate ρ_k using the observational dataset $\mathcal{D}$. In particular, the estimate does not account for the situation in which an organ arrives but there is no one present in the queue to receive it. In such a situation, the organ is moved to the next best queue. The more frequently this occurs at inference time, the worse our estimates for ρ_k will actually reflect the rate of arrival of organs into a given queue. In particular, as $K \to \infty$, this event occurs more and more frequently, as more and more queues are left empty at any given time.

[*Step 5*] - Using the queues. When a patient enters the transplant system, we identify their ideal queue, $k^* = \arg\max_k \hat{Y}(\mathbf{O}_k)$ where $\mathbf{O}_k$ is the cluster-center of cluster k. Next, we compare the patient's waiting time in the queue of interest, V_{k^*}, with the patient's expected survival using Eq.(3). Should the queue's waiting time be longer than the patient is expected to live, we move the patient to the next best queue. When an organ enters the transplant system, it is allocated to the first-in-line of the closest queue (in terms of cluster-center) to the available organ. Organ-allocation in OrganSync is outlined in the right-most part of Figure 2, and pseudo-code for our full training procedure is in our supplemental materials.

4. Related work and benchmarks

Before we discuss our experiments, we provide a brief overview of our considered benchmarks in potential outcome estimation, as well as organt-to-patient allocation.

4.1. Potential outcome estimation

A key part in our work is the use of a potential outcomes estimator (Rubin, 2005). Specifically, an estimator for more-dimensional, continuous treatment effects. This is of key importance as, like in many clinical settings (Dahabreh et al., 2016), organ-allocation suffers from selection bias (by definition). Therefore, we require methods that can cope with *biased data*. While potential outcomes estimation for binary treatment (Johansson et al., 2016; 2020; Bertsimas et al., 2017; Athey & Imbens, 2016; Hassanpour & Greiner, 2019; 2020; Yao et al., 2018; Zhang et al., 2020), or categorical treatments (Yoon et al., 2018; Alaa & van der Schaar, 2017; Alaa et al., 2017; Bica et al., 2020a), or continuous one-dimensional treatments (Bica et al., 2020b) has been studied profoundly; little attention has been devoted to more-dimensional treatments (Berrevoets et al., 2020). Contrasting existing methods, our method does not rely on learning a balanced representation of the data (Berrevoets et al., 2020; Bica et al., 2020a; Ganin et al., 2016; Schoenauer-Sebag et al., 2019; Li et al., 2018; Johansson et al., 2016), but rather the composition of a comparable synthetic control (Abadie et al., 2015) as outlined in Section 3.1.

Benchmarks. In our experiments, we compare against four other proposed high-dimensional treatment effect estimators: (i) TransplantBenefit (Neuberger et al., 2008), used for liver allocation-policy in the UK today; (ii) ConfidentMatch (Yoon et al., 2017), a recent organ outcome prediction

method; (iii) a multi-task network predicting organ outcome based on organ-types using a KMeans cluster, also compared against in Berrevoets et al. (2020); and (iv) OrganITE (Berrevoets et al., 2020), to our knowledge the only other high-dimensional potential outcome estimator. While OrganITE can estimate high-dimensional treatment outcomes, their approach differs vastly from OrganSync. OrganITE builds a balanced representation of treated and untreated patients, such that it is harder to predict in which original treatment group a patient belonged; while OrganSync explicitly builds an alternative patient-organ pair from past data. While OrganITE loses all reference to which past patient-organ pairs a prediction is based on, OrganSync keeps a hard link to previous data, lending to interpretability.

Table 2. **Results on organ allocation.** For each dataset we report the allocation performance of the benchmarks outlined in Table 1, in terms of added life-years (ALY), as well as total deaths over the course of one year. We set MELD as the baseline, to compare against. All results are reported in percentages ("%" is dropped for brevity) and ran over ten data-folds, standard deviation in brackets.

	UNOS		**UKReg**	
Method	Deaths	ALY	Deaths	ALY
MELD	compared against			
FIFO	-0.9 (.01)	-2.0 (.11)	-1.1 (.16)	-5 (.01)
M-na	-0.3 (.13)	+1.2 (.10)	-2.1 (.18)	+6 (.01)
TB	+7.0 (.19)	+2.4 (.21)	+0.9 (.11)	+8 (.03)
CM	-0.01 (.09)	+12.8 (.31)	+0.1 (.11)	+7 (.02)
O-ITE	**-3.6 (.18)**	+11.1 (.28)	-3.3 (.12)	+11 (.15)
OS	-3.5 (.15)	**+13.1 (.19)**	**-4.1 (.21)**	**+ 13 (.03)**

4.2. Organ allocation benchmarks

We have summarised various allocation strategies in Table 1, they are: First-in-first-out (FIFO), a naive queueing algorithm where the patient longest in the queue receives the next available organ; MELD (Malinchoc et al., 2000)— in use in the EU —computes a score for every patient based on three lab parameters: INR, Creatinine, and Bilirubin; MELD-na (M-na) (Kim et al., 2008)— used in the USA —takes into account Sodium alongside the MELD score; TransplantBenefit (TB) (Neuberger et al., 2008)— currently in use in the UK —calculates patient survival with and without organ using two distinct Cox Proportional Hazard models; and OrganITE (O-ITE) (Berrevoets et al., 2020), which uses a high-dimensional treatment effect estimator, paired with an organ density to account for rarity across patients in $Q(t)$ by weighting the ITE prediction.

5. Experiments

We evaluate OrganSync's performance on organ-allocation against various benchmarks (detailed in Section 4), and provide extensive analysis in our novel high-dimensional individual treatment effects estimator with respect to performance, as well as interpretability. For this we use two major datasets: the United Network for Organ Sharing (UNOS) dataset (Cecka, 2000), and the UK transplant registry data (available upon request). With each model we have performed an extensive hyperparameter search using a Bayesian optimisation scheme. The chosen hyperparameters are reported in the supplemental material[3].

5.1. Organ-to-patient allocation

Using our novel simulation, we evaluate how well OrganSync allocates organs to patients on the wait-list. Essentially, we evaluate any policy in Table 1 providing an interface such as in Eq.(1). With our simulation we provide highly accurate outcome estimates for patient-organ pairs that are not well represented in the given dataset; report deaths and added life-years; and provide an abstract interface for interaction with any given policy.

Deaths and average added life-years. As shown in Table 2, we report increases up to 13.1% in average life-years as compared to MELD; and decreases of -4.1% of premature deaths on the wait-list. While these percentages may seem small, they represent approximately 200 yearly deaths in the USA liver transplant-system. We also note a comparable performance to OrganITE, *while remaining entirely interpretable*. Having an interpretable machine learning model is important should it ever be adopted in healthcare (Ahmad et al., 2018; Doshi-Velez & Kim, 2017).

5.2. Prediction and interpretability

We evaluate our first component– the potential outcomes estimator –on outcome prediction (both factual and counterfactual), as well as interpretability. Results on prediction are reported in Table 3, and on interpretability in Figure 3.

Factual versus counterfactual evaluation. By predicting outcomes as they are reported in the dataset, we evaluate on factual outcomes. While OrganSync performs well in this regard, performance on counterfactuals is more important for a potential outcomes estimator. That is, by training on biased data, we wish to evaluate using unbiased test data. While indeed desirable, this is impossible using the factual data as is. As such, from our two datasets, we create two semi-synthetic datasets where the outcomes are replaced by a known function: $Y(\mathbf{O}) = \nu \exp\{\theta_1^\top \mathbf{X} + \theta_2^\top \mathbf{O} + \frac{1}{2}\} + \epsilon_Y$. We then leave the training set biased, and shuffle patient and organs to form random matches in the test set. A similar

[3] All our code to reproduce our results is available at `https://github.com/vanderschaarlab/mlforhealthlab`, including: model definition, simulations, data preprocessing, and semi-synthetic data generation.

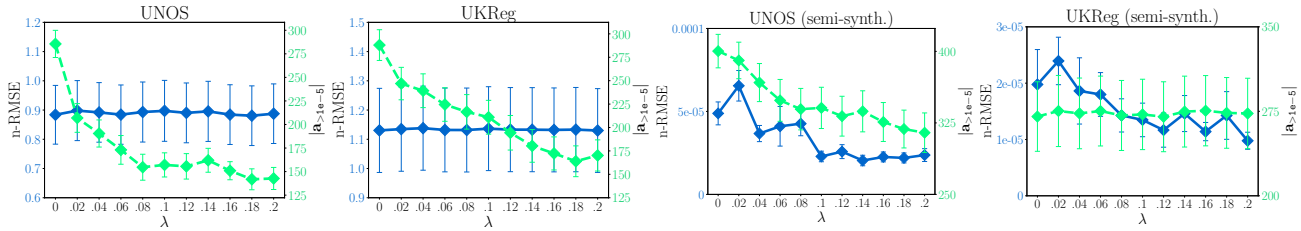

Figure 3. **Performance of OrganSync's potential outcomes estimator with different λ.** We report both n-RMSE, and the average amount of elements in **a** that are above `1e-5`. **On real data:** While λ heavily impacts the amount of contributors to the synthetic control (less than half when $\lambda = 0.2$, compared to 0.0), performance remains largely untouched. This is due to heavy impact from a few "large" contributors versus "small" contributors; corresponding with high and low values in **a**, respectively. **On semi-synth. data:** The reverse is true for semi-synthetic data; performance increases with higher λ, though it seems counterfactual estimation requires more contributors, even with high λ. This makes sense as these are typically out-of-sample estimations which demand more extreme synthetic controls.

Table 3. **Performance on prediction outcomes.** Results are averaged over different 50 data-folds (standard deviation in brackets) on two major datasets. The leftmost part reports n-RMSE, the rightmost part reports the mean absolute error in days. Note, that we did not report the mean absolute error in days for synthetic data as this has no meaningful indication of expected error in reality. Lower is better.

	Semi-synth. (counterf.)		**Real (factual)**			
	n-RMSE				Mean abs. error in days	
	UNOS	UKReg	UNOS	UKReg	UNOS	UKReg
TransplantBenefit	2.66 (.306)	1.32 (.012)	1.75 (.264)	1.32 (.015)	2087 (20.7)	2833 (18.3)
ConfidentMatch	1.17 (.069)	0.84 (.523)	0.83 (.025)	0.90 (.082)	1328 (39.4)	1625 (236)
Multi-task	.027 (.004)	0.24 (.052)	0.81 (.009)	0.76 (.201)	796 (12.1)	1076 (32.7)
OrganITE	.030 (.001)	0.15 (.016)	0.79 (.010)	0.46 (.143)	761 (16.4)	634 (27.4)
OrganSync-β	.049 (.002)	0.29 (.077)	**0.76** (.005)	**0.39** (.004)	**715** (4.13)	**497** (5.72)
OrganSync-SC	**.005** (.004)	**0.09** (.005)	0.81 (.031)	0.42 (.006)	781 (19.1)	541 (14.7)

strategy was also used in Berrevoets et al. (2020).

Prediction. We report the normalised root mean squared error (n-RMSE) for factual and counterfactual estimation, and the mean absolute error in days for factual data in Table 3. Besides the benchmarks described in Section 4.1, we also compare OrganSync *with* synthetic estimates (OrganSync-SC), and *without* synthetic estimates (OrganSync-β). The former regresses through Eq.(7), while the latter is simply regressing Y through Eq.(4), without any consideration of Eq.(5) or Eq.(6), essentially reducing to a standard MLP.

Interpretability. One of the major problems we wish to solve, is that of providing *interpretable* potential outcome estimates for high-dimensional treatments. All our benchmarks are either interpretable (TransplantBenefit), or non-linear (ConfidentMatch, Multi-task networks, and OrganITE); but not both. Non-linear estimates are difficult to interpret as they combine all the past seen data to build a non-linear representation space, which in the case of UNOS amounts to roughly 80k patients, and UKReg to 15k patients (in the training sets). OrganSync is no exception, the non-linear component is similarly not interpretable. Furthermore, learning a second model that functions as a "translator" may result in unfaithful explanations, as stated by Rudin (2019).

By linearly combining patient-organ pairs, OrganSync remains explicit about exactly how much each pair contributes to the estimate. Furthermore, their contribution remain interpretable as it is both linear, and constrained to sum to one. While one might assume that synthetic controlled estimates adversely affect performance, Figure 3 shows this is not the case. In Figure 3 we report prediction performance and the size of the synthetic control given λ. Having more contributors, makes the synthetic control less interpretable, warranting the use of higher λ.

We gain further insight into how OrganSync combines patient-organ pairs into synthetic controls through Figure 4. In Figure 4 we have learned ten KMeans clusters over the patient-organ space, $\mathcal{X} \times O$ (top row), as well as the representation space, $\mathcal{U}$ (bottom row). Using these clusters, we sampled 50 patient-organ pairs per cluster from the test set, and computed their synthetic control. We then counted the amount of patients in each cluster that contributed to their synthetic control and normalised the results. When comparing the top row, with the bottom row, we recognise that the non-linear representation rearranges the patient-organ pairs in such a way that they are easily combined in a linear way. Furthermore, having the non-linear representation allows OrganSync to combine non-obvious pairs from the dataset

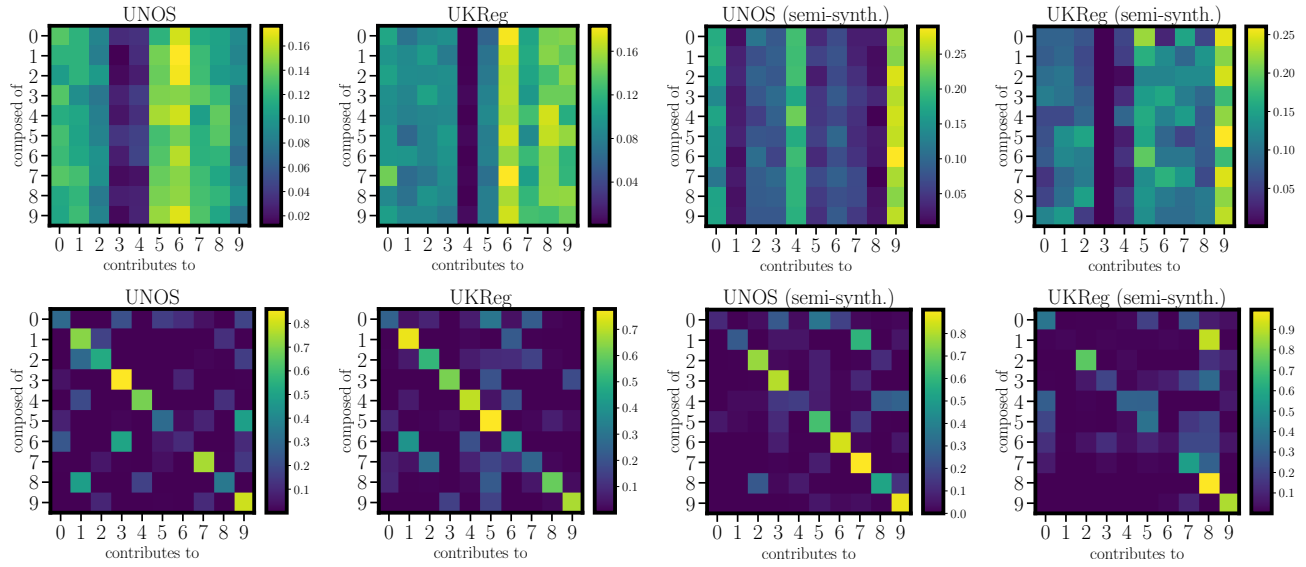

Figure 4. **Contributions of different patient-organ types to synthetic pairs.** We have clustered the patient-organ pairs in the training dataset using KMeans, with `K=10`. The row **above** depicts clusters in the *patient-organ space* ($\mathcal{X} \times \mathcal{O}$), the row **below** depicts clusters in the *representation space* ($\mathcal{U}$). From each cluster (0-9), we sampled 50 patient-organ pairs from the test set, and composed their synthetic pair. For each cluster (0-9), we report the contribution of all clusters to the predicted outcome of their synthetic pairs. In these plots, "contribution" means the information contained in a group of patient-organ pairs (a cluster) has greater or lesser impact to explaining the predicted outcome of a synthetic patient-organ pair. The **vertical axis** shows for each cluster, what the cluster contributes to; the **horizontal axis** shows for each cluster, what other clusters they are composed of. We have normalised these outcomes for each row, such that one square can be interpreted as a percentage. For example, in UNOS (left-above), cluster "6" is the highest frequent contributor, while clusters "3 and 4" contribute least. Note that the cluster types have no significance across datasets. *We learn that OrganSync represents the data such that similar patient-organ pairs are nearby in $\mathcal{U}$, while scattered across $\mathcal{X} \times \mathcal{O}$.*

(illustrated by the lower values in the top row of Figure 4).

Example. To fully cement the interpretability of OrganSync, we provide one concrete example in Table 4. In our example, we report the MELD-na features: Creatinine (Creat.), Bilirubin (Bilir.), INR, and Sodium; for a more complete example we refer to our supplemental materials. Note that searching a synthetic control on the features directly, would likely result in very different contributors as reported in Table 4. Using these types of tables, OrganSync delivers exactly what is claimed by Rudin (2019) to be necessary for machine learning models to be useful in clinical settings— interpretation, directly provided by the model, not by a second model functioning as a "translator".

Table 4. **Example.** We report the top 3 contributors for one example's synthetic control. The absolute error in life-years for OrganSync was 243 days, and 532 days for TransplantBenefit. Note that the top 3 already span 8 years of past cases.

Contrib.	Creat.	Bilir.	INR	Sodium	*Year*
Example	46	169	1.1	140	*2019*
Past patients in synthetic control					
`.399`	254	238	1.6	134	*2019*
`.261`	71	35	1.4	135	*2017*
`.241`	72	20	1.0	142	*2011*
...					

6. Discussion

OrganSync has the potential to transform healthcare by aiding clinicians in treatment decisions while offering new insight through various interpretations. For example, consider Table 4, where the top contributors to a synthetic control do not seem to be very comparable in terms of the MELD-na features which are for the past 13 years considered very important in liver allocation. A sentiment that is further explained by Figure 4 where the synthetic controls seem to span the entire patient-organ feature space. However, It is important to note that machine learning models can fail. In particular, OrganSync relies on two assumptions outlined in Section 3.1: *overlap*, and *unconfoundedness*. When these assumptions are violated, OrganSync might fail to correctly predict counterfactual scenarios. Having domain knowledge about the provided data (Johnson et al., 2014; Lerut et al., 2020) is thus crucially important to confirm these assumptions. As such, we envisage OrganSync as a *decision support tool*, working in tandem with clinicians. In fact, this is exactly why interpretable decisions and predictions are so important in OrganSync.

Acknowledgements

The research presented in this paper was supported by the W.D. Armstrong Trust, The Alan Turing Institute, and by the US Office of Naval Research (ONR). The authors would like to thank the anonymous reviewers, as well as Brent Ershoff (UCLA) for many helpful discussions, and Alicia Curth (University of Cambridge) for reviewing an early draft of our paper.

References

Abadie, A. and Gardeazabal, J. The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132, 2003.

Abadie, A., Diamond, A., and Hainmueller, J. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, 2015.

Ahmad, M. A., Eckert, C., and Teredesai, A. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '18, pp. 559–560, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357944. doi: 10.1145/3233547.3233667. URL https://doi.org/10.1145/3233547.3233667.

Alaa, A. M. and van der Schaar, M. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3424–3432. Curran Associates, Inc., 2017.

Alaa, A. M., Weisz, M., and van der Schaar, M. Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966*, 2017.

Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Beran, R. Nonparametric regression with randomly censored survival data. Technical report, University of California, Berkeley, 1981.

Berrevoets, J., Jordon, J., Bica, I., Gimson, A., and van der Schaar, M. OrganITE: Optimal transplant donor organ offering using an individual treatment effect. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20037–20050. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/e7c573c14a09b84f6b7782ce3965f335-Paper.pdf.

Bertsimas, D., Kallus, N., Weinstein, A. M., and Zhuo, Y. D. Personalized diabetes management using electronic medical records. *Diabetes care*, 40(2):210–217, 2017.

Bica, I., Alaa, A. M., Jordon, J., and van der Schaar, M. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=BJg866NFvB.

Bica, I., Jordon, J., and van der Schaar, M. Estimating the effects of continuous-valued interventions using generative adversarial networks. In *Advances in Neural Information Processing Systems*, 2020b.

Cecka, J. M. The unos scientific renal transplant registry–2000. *Clinical transplants*, pp. 1–18, 2000.

Chen, G. H. Deep kernel survival analysis and subject-specific survival time prediction intervals. In *Machine Learning for Healthcare Conference*, pp. 537–565. PMLR, 2020.

Dahabreh, I. J., Hayward, R., and Kent, D. M. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *International journal of epidemiology*, 45(6):2184–2193, 2016.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Eurotransplant Reference Laboratory (ETRL). Livers transplanted in 2019, 2020. URL https://www.eurotransplant.org/organs/liver/. Online; accessed 4-February-2021.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Gimson, A. Development of a uk liver transplantation selection and allocation scheme. *Current Opinion in Organ Transplantation*, 25(2):126–131, 2020.

Hassanpour, N. and Greiner, R. Counterfactual regression with importance sampling weights. In *IJCAI*, pp. 5880–5887, 2019.

Hassanpour, N. and Greiner, R. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxBJT4YvB.

Hirano, K. and Imbens, G. W. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2(3):259–278, 2001.

Hirano, K. and Imbens, G. W. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84, 2004.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029, 2016.

Johansson, F. D., Shalit, U., Kallus, N., and Sontag, D. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

Johnson, R. J., Bradbury, L. L., Martin, K., Neuberger, J., et al. Organ donation and transplantation in the uk—the last decade: a report from the uk national transplant registry. *Transplantation*, 97:S1–S27, 2014.

Kilambi, V., Bui, K., and Mehrotra, S. Livsim: an open-source simulation software platform for community research and development for liver allocation policies. *Transplantation*, 102(2), 2018.

Kim, W. R., Biggins, S. W., Kremers, W. K., Wiesner, R. H., Kamath, P. S., Benson, J. T., Edwards, E., and Therneau, T. M. Hyponatremia and mortality among patients on the liver-transplant waiting list. *New England Journal of Medicine*, 359(10):1018–1026, 2008. doi: 10.1056/NEJMoa0801209. URL https://doi.org/10.1056/NEJMoa0801209. PMID: 18768945.

Lerut, J., Karam, V., Cailliez, V., Bismuth, H., Polak, W. G., Gunson, B., Adam, R., and European Liver, I. T. A. E. What did the european liver transplant registry bring to liver transplantation? *Transplant International*, 33(11): 1369–1383, 2020.

Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., and Tao, D. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 624–639, 2018.

Little, J. D. A proof for the queuing formula: L= λ w. *Operations research*, 9(3):383–387, 1961.

Malinchoc, M., Kamath, P. S., Gordon, F. D., Peine, C. J., Rank, J., and Ter Borg, P. C. A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. *Hepatology*, 31(4):864–871, 2000.

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

NHSBT Liver Advisory Group (LAG). POLICY POL196/7 - Deceased Donor Liver Distribution and Allocation, 2019. URL https://nhsbtdbe.blob.core.windows.net/umbraco-assets-corp/17449/liver-allocation-policy-pol196.pdf. Online; accessed 4-February-2021.

Organ Procurement and Transplantation Network (OPTN). About meld and peld, 2021a. URL https://optn.transplant.hrsa.gov/resources/allocation-calculators/about-meld-and-peld/. Online; accessed 4-February-2021.

Organ Procurement and Transplantation Network (OPTN). Organ procurement and transplantation network: Organ donation and transplantation can save lives, 2021b. URL https://optn.transplant.hrsa.gov. Online; accessed 4-February-2021.

Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Schoenauer-Sebag, A., Heinrich, L., Schoenauer, M., Sebag, M., Wu, L., and Altschuler, S. Multi-domain adversarial learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Sklv5iRqYX.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3076–3085. JMLR. org, 2017.

Thompson, D., Waisanen, L., Wolfe, R., Merion, R. M., McCullough, K., and Rodgers, A. Simulating the allocation of organs for transplantation. *Health care management science*, 7(4):331–338, 2004.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. Representation learning for treatment effect estimation from observational data. In *Advances in*

Neural Information Processing Systems, volume 31, pp. 2633–2643. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/a50abba8132a77191791390c3eb19fe7-Paper.pdf.

Yoon, J., Alaa, A. M., Cadeiras, M., and van der Schaar, M. Personalized donor-recipient matching for organ transplantation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 1647–1654. AAAI Press, 2017.

Yoon, J., Jordon, J., and van der Schaar, M. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByKWUeWA-.

Zhang, Y., Bellot, A., and van der Schaar, M. Learning overlapping representations for the estimation of individualized treatment effects. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 1005–1014. PMLR, 2020.