Proceedings of Machine Learning Research 248:683–695, 2024 Conference on Health, Inference, and Learning (CHIL) 2024

Learning Social Fairness Preferences from Non-Expert Stakeholder Opinions in Kidney Placement

Mukund Telukunta MT3QB@MST.EDU
Missouri University of Science and Technology

Sukruth Rao RAOSUKRU@MSU.EDU
Michigan State University

Gabriella Stickney STICKN21@MSU.EDU
Michigan State University

Venkata Sriram Siddardh Nadendla NADENDLA@MST.EDU
Missouri University of Science and Technology

Casey Canfield CANFIELDCI@MST.EDU
Missouri University of Science and Technology

Abstract

Modern kidney placement incorporates several intelligent recommendation systems which exhibit social discrimination due to biases inherited from training data. Although initial attempts were made in the literature to study algorithmic fairness in kidney placement, these methods replace true outcomes with surgeons' decisions due to the long delays involved in recording such outcomes reliably. However, the replacement of true outcomes with surgeons' decisions disregards expert stakeholders' biases as well as social opinions of other stakeholders who do not possess medical expertise. This paper alleviates the latter concern and designs a novel fairness feedback survey to evaluate an acceptance rate predictor (ARP) that predicts a kidney's acceptance rate in a given kidney-match pair. The survey is launched on Prolific, a crowdsourcing platform, and public opinions are collected from 85 anonymous crowd participants. A novel social fairness preference learning algorithm is proposed based on minimizing social feedback regret computed using a novel logit-based fairness feedback model. The proposed model and learning algorithm are both validated using simulation experiments as well as Prolific data. Public preferences towards group fairness notions in the context of kidney placement have been estimated and discussed in detail. The specific ARP tested in the Prolific survey has been deemed fair by the participants.

Data and Code Availability This paper uses the kidney matching dataset (STAR file) requested from the Organ Procurement and Transplant Network (OPTN) to generate the data tuples presented to the survey participants. Given the donor/recipient IDs used in both simulation experiments as well as survey dataset are deemed as HIPAA PHI identifiers, both the code and dataset are also not released to the public. However, both code and data can be made available upon request only after obtaining consent from OPTN to avail the STAR file.

Institutional Review Board (IRB) This research paper has undergone ethical review and approval by the University of Missouri Institutional Review Board with the approval number 2092366. The informed consent process, including the information provided to participants and the procedures for obtaining their voluntary and informed consent, has been reviewed and approved by the IRB. Participants were assured of the confidentiality and privacy of their data, and all efforts have been made to minimize any potential risks associated with their involvement in the study.

1. Introduction

The increasing rate of kidney non-utilization in deceased donors (Lentine et al., 2023) has inspired the adoption of machine learning (ML) solutions to identify kidneys with high risk of non-utilization (Barah and Mehrotra, 2021), provide predictive analytics regarding patient's mortality likelihood without a transplant (McCulloh et al., 2023), and predictions to transplant surgeons by predicting the probability of kidney offer acceptance (Ashiku et al., 2022). However, these models are susceptible to social discrimination, as they are trained using past decisions curated during traditional kidney placement practices. For instance, the inclusion of *race* coefficient in the computation of Kidney Donor Profile Index (KDPI) systematically assigns higher scores to kidneys from Black donors irrespective of whether or not they carry the APOL1 gene (one that results in a guaranteed failure of renal transplantation), thereby contributing to an increase in the overall non-utilization rate (Chong et al., 2021). At the same time, the *age* attribute in calculating patient's Estimated Post Transplant Survival (EPTS) score allocates high-quality kidneys to younger recipients at the expense of older patients with a potentially greater medical need (Eidelson, 2012). From a non-ML perspective, several researchers explored fairness in kidney placement either by improving organ allocations for older patients (Mattei et al., 2018), ensuring that similar patients have the same chances of receiving a transplant (Farnadi et al., 2021), or proposing novel metrics to ensure equity in kidney transplant allocations for patients of different racial/ethnic background. However, to the best of our knowledge, there hasn't been much research conducted on fair ML-based system in kidney placement domain. Despite the urgent need to analyze such biases, there are two main challenges in evaluating the fairness of diverse ML-based predictors deployed within the complex decision pipeline in kidney placement.

Firstly, a significant limitation with state-of-the-art fairness notions (especially group-based notions (Mehrabi et al., 2021)) is their reliance on final outcomes, which are usually observed in hindsight. For example, the death of an organ recipient can only be observed in hindsight, only during a two-year post transplantation monitoring period. The process of recording true outcomes is very challenging due to the need to track organ recipients post surgery over at least 2-5 years. One of the initial attempts to address this challenge is the design of a novel individual fairness notion called Discounted Cumulative Fairness (DCF) (Zhang et al., 2023) that quantifies individual unfairness based on user-rank when the final outcomes are unavailable for certain instances. However, such an approach does not provide analysis from a group fairness perspective, which primarily focuses on disparate impact across various protected groups in the society. As an alternative, human perception of fairness is proposed where perceived labels are collected from expert critics for a quick analysis (Srivastava et al., 2019; Grgic-Hlaca et al., 2018).

The second challenge is that selective feedback elicitation from *clinical experts* (e.g. transplant surgeons, organ procurement teams) can lead to myopic fairness analysis. For example, fairness opinions of donors and recipients (a.k.a. *personal experts* who lack technical knowledge but possess the basic understanding through interaction with clinical experts as well as their own peers) are important, since their lives are directly impacted by the decisions made by expert stakeholders in kidney placement pipeline. Most often, their opinions about the kidney placement process could deviate quite significantly from clinical experts' opinions. Furthermore, the analysis behind clinical experts' decisions regarding kidney offers/procurement is not revealed by OPTN in their STAR file due to HIPAA restrictions, which could otherwise turn very useful in fairness analysis. Moreover, there is a large group of personal experts and public critics who are available to elicit opinions regarding fairness preferences, as opposed to a handful of clinical experts who are often available only for a very limited time. Therefore, the main objective of this paper is to elicit fairness opinions from public critics regarding the performance of ML-based predictors used in kidney placement across diverse social groups, and learn their preferences across diverse group fairness notions.

The main contributions of this paper are three-fold. Firstly, this paper investigates ***personal expert's (i.e., public) fairness preferences to evaluate ML-based predictors used in kidney placement*** pipeline. This is the first-of-its-kind effort to elicit fairness opinions from stakeholders other than clinical experts in kidney placement. A human-subject ***survey experiment*** was conducted on Prolific crowdsourcing platform to collect feedback regarding the fairness of a ML-based system from non-expert (public) participants. In contrast to prior efforts, participants are not constrained to any particular fairness perspective, and are free to choose their preferred group fairness notions at will, and assess the fairness of the ML-system for a given sensitive attribute(s). Secondly, a ***novel logit-based feedback model*** is proposed based on encoded Likert choices and *noisy fairness preferences* across group fairness notions. Thirdly, a ***projected gradient-descent algorithm***

DONOR	
Age	29
Race	White
Gender	Male
Kidney Quality	16

Table 1: An Example of 4 Donor Characteristics (out of 17 attributes used in ARP model) Revealed to Participants in the Survey

with an efficient gradient computation is designed to minimize social feedback regret. The proposed approach is validated on a wide range of simulation experiments. Finally, the proposed method was adopted to analyze and ***find public's social preferences recorded in Prolific survey*** dataset.

The remainder of this paper is organized as follows. Section 2 presents a brief literature survey on human fairness perception. The Prolific experiment is discussed in Section 3, which is then followed by the proposed methodology in Section 4. Evaluation methodology is presented in detail in Section 5, followed by results and their discussion in Section 6.

2. Human Fairness Perception: A Brief Literature Survey

In the past, several researchers have attempted to model human perception of fairness. For instance, in an experiment performed by Srivastava et al. (2019), participants were asked to choose among two different models to identify which notion of fairness (demographic parity or equalized odds) best captures people's perception in the context of both risk assessment and medical applications. Likewise, another team surveyed 502 workers on Amazon's Mturk platform and observed a preference towards *equal opportunity* in Harrison et al. (2020). Work by Grgic-Hlaca et al. (2018) discovered that people's fairness concerns are typically multi-dimensional (relevance, reliability, and volitionality), especially when binary feedback was elicited. A very recent work of Lavanchy et al. (2023) conducts four survey experiments to study applicants' perception towards algorithm-driven hiring procedures. Their findings indicate that recruitment processes are deemed less fair compared to human only or AI-assisted human processes, regardless of applicants receiving a positive outcome.

3. Experiment Design

The objective of the survey experiment is to collect non-expert (i.e. public) feedback regarding the fairness of a state-of-the-art kidney acceptance rate predictor (ARP) (Ashiku et al., 2022). This predictor is an analytics tool that predicts kidney acceptance probability based on donor-recipient characteristics (includes both medical features and social demographics) in order to support transplant surgeon decisions regarding deceased donor kidney offers and alleviate kidney non-utilization. The predictor was trained using kidney matching datasets spanning from 2014 to 2018, achieving a testing accuracy of 96%.

3.1. Datasets and Preprocessing

Public participants are provided with predictions from the ARP for various kidney matching instances spanning 2020 and 2021. These predictions are based on datasets called Standard Transplant Analysis and Research (STAR) files, obtained from the Organ Procurement and Transplant Network (OPTN). The STAR files contain anonymized patient-level data on transplant recipients, donors, and matches dating back to 1987. Each dataset typically includes numerous instances where a deceased donor kidney is matched with thousands of potential recipients (refer Appendix A for more details on kidney placement). Since presenting such large datasets can overwhelm the participants, the number of potential recipients for each deceased donor was limited to $K = 10$. This subset includes at least one recipient who received the kidney, ensuring a balanced representation of successful and unsuccessful transplant outcomes. The remaining recipients were randomly selected. Additionally, recipients under 17 years old were excluded due to unique challenges in pediatric transplantation (Magee et al., 2004). The preprocessed dataset comprised 13,628 deceased donors from 2021 and 5,023 from 2022. A sample of 10 data-tuples (each containning one donor, 10 recipients) are randomly selected from the preprocessed STAR dataset. In other words, a total of 100 random potential recipients were selected from the STAR file for our survey experiment. These 100 potential recipients are sampled to maintain the same diversity across different protected groups as observed in the complete STAR file for years 2021 and 2022. The ARP was then applied to this sample to obtain acceptance rates for every potential recipient within each deceased donor kidney.

POTENTIAL RECIPIENTS

Recipient #	1	2	3	4	5	6	7	8	9	10
Age	19	48	73	73	63	67	67	71	73	38
Race	White	Black	Black	White	White	White	White	White	White	Multi-Racial
Gender	Male	Female	Male	Male	Male	Female	Male	Male	Male	Female
Est. Post Transplant Survival	3	16	92	86	60	59	67	61	65	17
Distance from Transplant Center	0.0 miles	210.0 miles	168.0 miles	220.0 miles	82.0 miles	204.0 miles	65.0 miles	0.0 miles	0.0 miles	199.0 miles
Acceptance Rate	59%	81%	49%	97%	54%	78%	78%	78%	78%	59%
Surgeon's Decision	No Transplant	No Transplant	No Transplant	No Transplant	No Transplant	No Transplant	No Transplant	No Transplant	No Transplant	Transplant

Figure 1: An Example of Recipient Characteristics

Q2. Given the **surgeon's decision**, how fair is the Acceptance Rate from the ML Predictor for **older (age > 50) versus younger (age < 50)** recipients?

Completely Unfair ○ Moderately Unfair ○ Slightly Unfair ○ Neither Fair nor Unfair ○ Slightly Fair ○ Moderately Fair ○ Completely Fair ○

Q3. Given the **surgeon's decision**, how fair is the Acceptance Rate from the ML Predictor for **Female versus Male** recipients?

Completely Unfair ○ Moderately Unfair ○ Slightly Unfair ○ Neither Fair nor Unfair ○ Slightly Fair ○ Moderately Fair ○ Completely Fair ○

Q4. Given the **surgeon's decision**, how fair is the Acceptance Rate from the ML Predictor for **Black versus Other Races/Ethnicities** recipients?

Completely Unfair ○ Moderately Unfair ○ Slightly Unfair ○ Neither Fair nor Unfair ○ Slightly Fair ○ Moderately Fair ○ Completely Fair ○

Figure 2: Three Questions Presented to the Participants for Each Data Tuple

A single donor paired with 10 potential recipients is considered as a *data-tuple*.

3.2. Survey Questions

This survey presents data as two distinct tables for each data-tuple. The first table contains information regarding the deceased donor including donor's age, race, gender, and KDPI score. As an illustration, Table 1 presents all the donor characteristics in a fake data-tuple example presented to the survey participant. Note that these attributes are hand-picked amongst the 17 donor attributes used in the original ARP model, to facilitate public critic's evaluation who is not expected to have any medical expertise. The second table presents information on ten fake recipient profiles potentially matched with this donor, which includes each recipient's age, race, gender, EPTS score, distance from the transplant center, prediction from ARP, and the surgeon's decision (transplant or no transplant), as shown in the Figure 1. Subsequently, the participants were instructed to respond to four distinct questions within each data-tuple. The participants were asked to rate the fairness of the ARP using a Likert scale since it is the most widely used approach in research surveys. Specifically, the participants responded using a Likert scale ranging from 1 to 7 (denoted as s), where 1 indicates complete unfairness, and 7 denotes complete fairness. Given that any group fairness evaluation lies in the interval $[-1, 1]$, the chosen Likert scaling allows to capture the extent of (un)fairness perceived by a participant. Following this, the participants were further prompted to assess the fairness of the ARP in context of (i) older recipients (age > 50) versus younger recipients (age < 50), (ii) female versus male recipients, and (iii) Black recipients versus recipients from other racial backgrounds (as shown in Figure 2). Please refer Appendix B for more details regarding the survey design.

3.3. Participant Demographics

The survey experiment was deployed on Prolific (IRB Reference Number 2092366) during December 2023. A total of 85 participants were recruited for the study. Among them, $N = 75$ individuals were chosen, with the exclusion of 8 participants experiencing technical difficulties, and an additional 2 participants failing to answer the attention check questions. On an average, the participants took 20 minutes to complete the entire survey and were compensated at the rate of \$11.58/hour. Table 2 summarizes the demographics of the recruited participants.

Demographic Attribute	Prolific	Census
18-25	8%	13%
25-40	57%	26%
40-60	29%	32%
>60	6%	22%
White	60%	59%
Black	19%	12%
Asian	12%	5.6%
Hispanic	3.4%	18%
Other	5.6%	9%
Male	49%	49.5%
Female	49%	50.5%
Non-binary	2%	-
High School or equivalent	18%	26.5%
Bachelor's (4 year)	40%	20%
Associate (2 year)	15%	8.7%
Some college	12%	20%
Master's	11%	13%

Table 2: Participants demographics compared to the 2021 U.S. Census Data.

The recruited participants consisted of fewer Hispanics (3.4%), more Blacks (19%), more educated (51%) and more younger (65%) individuals compared to the 2021 U.S. Census (Bureau). Such skew in demographics is commonly observed while recruiting participants from crowd-sourcing platforms. In addition, recruiting older participants for research studies can be difficult because of barriers associated with aging, such as participation online and on social media (Turner et al., 2020). Moreover, Prolific has been noted to have demographics that skew younger participants (Charalambides, 2021).

4. Methodology

4.1. Fairness Feedback Model

Consider N non-expert participants who evaluate the acceptance rate predictor (ARP) from the perspective of group fairness across sensitive demographics. The n^{th} participant investigates the m^{th} representative *data-tuple* $\boldsymbol{d}_m = \{\boldsymbol{x}_{1:K}^{(m)}, \boldsymbol{y}_{1:K}^{(m)}, \hat{\boldsymbol{y}}_{1:K}^{(m)}\}$ from ARP, which comprises of the donor-recipient attributes $\boldsymbol{x}_{1:K}^{(m)}$, surgeon's decisions $\boldsymbol{y}_{1:K}^{(m)}$ and the ARP's predictions $\hat{\boldsymbol{y}}_{1:K}^{(m)}$ across K donor-recipient pairs. Upon investigation, the n^{th} participant presents a fairness feedback score $s_{n,m} \in \{1, 2, \cdots, 7\}$ to the evaluation platform (ref. Figure 3), where $s_{n,m} = 1$ indicates an unfair ARP and $s_{n,m} = 7$ indicates a fair ARP.

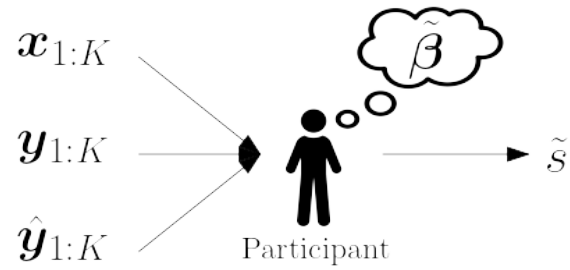


Figure 3: Non-Expert Participant's Feedback Model

In this section, the n^{th} participant's fairness feedback score $s_{n,m}$ is modeled as follows. Assume that the n^{th} participant exhibits an unknown *preference weight* $\boldsymbol{\beta}_n = \{\beta_{n,1}, \cdots, \beta_{n,L}\}$ over L group fairness notions. In other words, $\beta_{n,l} \in [0, 1]$ and $\sum_{l=1}^{L} \beta_{n,l} = 1$, for all n, l. Let $\phi_\ell(\boldsymbol{d}_m)$ denote the evaluation of ARP from the perspective of ℓ^{th} fairness notion. For the sake of brevity, the computation of group fairness notions is discussed in detail in Appendix C. Let the n^{th} participant aggregate the L fairness evaluations of ARP as

$$\psi_{n,m}(\boldsymbol{\beta}_n) = \sum_{l=1}^{L} \beta_{n,l} \cdot \phi_l(\boldsymbol{d}_m). \tag{1}$$

Since any fairness evaluation $\phi_l(\boldsymbol{d}_m)$ lies between -1 and 1, the aggregated fairness evaluation $\psi_{n,m}(\boldsymbol{\beta}_n) \in [-1, 1]$. Consequently, if $\psi_{n,m}(\boldsymbol{\beta}_n) = 0$, the n^{th} participant deems the ARP as a fair system. On the contrary, if $\psi_{n,m}(\boldsymbol{\beta}_n) = 1$ or -1, the n^{th} participant will deem the ARP system as an unfair one. However, the n^{th} participant encodes their aggregated fairness evaluation $\psi_{n,m}(\boldsymbol{\beta}_n)$ using Likert scale and reports a fairness feedback score $s_{n,m} \in \{1, \cdots, 7\}$. For the sake of simplicity, assume that the Likert encoding is accomplished by dividing the interval $[-1, 1]$ into 14 equal partitions, each with width $\delta = 1/7$. The boundaries of these partitions are therefore given as $b_i = -1 + i \cdot \delta$ for all $i = 0, 1, 2, \cdots, 14$. Let $\mathbb{R}_i$ denote the union of two partitions corresponding to the interval $[b_{i-1}, b_i]$ and $[b_{14-i}, b_{14-i+1}]$, for all $i = 1, \cdots, 14$.

In practice, participants often compute a noisy fairness evaluation, due to the ambiguity in their preferences towards diverse fairness notions. This stochasticity in the preferences across fairness notions can be captured using the Mixed-Logit model (McFadden et al., 1973; Train, 2009) from discrete

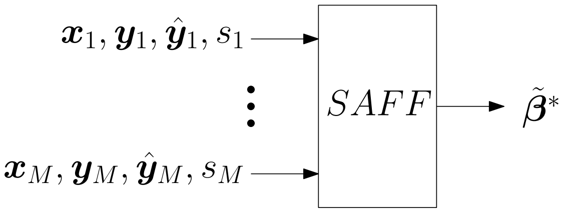


Figure 4: Social Aggregation of Fairness Feedback

choice theory, which is already found to be useful in the field of kidney transplantation (Genie et al., 2020; Howard, 2006). Let the true intrinsic fairness evaluation ψ follow a logit-Normal distribution $F(\cdot|\mu, \sigma)$, where the mean and variance of logit variable $\texttt{Logit}(\psi) = \log \frac{\psi}{1-\psi}$ are given by $\mu = \psi_{n,m}(\boldsymbol{\beta}_n)$ and some known constant σ^2 respectively. Then, the n^{th} participant experiences a utility $u_{n,i}$ as the probability of the true intrinsic fairness evaluation ψ to lie in a specific region $\mathbb{R}_i$. In other words, the utility is formally given by

$$u_{n,m,i}(\boldsymbol{d}_m) = V_i\Big(\psi_{n,m}(\boldsymbol{d}_m)\Big) + V_{14-i+1}\Big(\psi_{n,m}(\boldsymbol{d}_m)\Big), \tag{2}$$

where

$$\begin{aligned} V_i\Big(\psi_{n,m}(\boldsymbol{d}_m)\Big) &= F\left(\frac{1-b_i}{2}; \psi_{n,m}(\boldsymbol{d}_m), \sigma\right) \\ &\quad - F\left(\frac{1-b_{i-1}}{2}; \psi_{n,m}(\boldsymbol{d}_m), \sigma\right) \\ &= \int_{b_{i-1}}^{b_i} f(z; \psi_{n,m}(\boldsymbol{d}_m), \sigma) dz \end{aligned} \tag{3}$$

is the probability that the true intrinsic fairness evaluation ψ lies in the interval $[b_{i-1}, b_i]$, where $f(\cdot; \mu, \sigma)$ is the logit-normal density function with parameters μ and σ. Hence, the fairness feedback score $s_{n,m}$ is modeled as the mixed-logit probability

$$\tilde{s}_{n,m} = \frac{1}{\Delta_{n,m}} \cdot \Big\{ e^{\lambda \cdot u_{n,m,1}}, \cdots, e^{\lambda \cdot u_{n,m,7}} \Big\}, \tag{4}$$

where $\Delta_{n,m} = \sum_{j=1}^{7} e^{\lambda \cdot u_{n,m,j}}$ is the normalizing factor, and λ is the temperature parameter that captures the participant's sensitivity to the utilities.

4.2. Proposed Algorithm

The goal of this approach is to develop a social preference weight $\boldsymbol{\beta}^*$ that minimizes the average feedback

---

Algorithm 1: SAFF

Input: $\mathbf{x}_{1:M}, \mathbf{y}_{1:M}, \hat{\mathbf{y}}_{1:M}, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_N, \delta$
Output: Learned social preference $\tilde{\boldsymbol{\beta}}^*$
Initialize $\boldsymbol{\beta}^{(0)}$ with a random L-dim. weight
for $e = 1$ **to** *num_epochs* **do**
  for $m = 1$ **to** M **do**
    $\boldsymbol{\phi}_m \leftarrow \texttt{FairnessScores}(\mathbf{x}_m, \mathbf{y}_m, \hat{\mathbf{y}}_m)$
    $\tilde{s}^*_m \leftarrow \texttt{EstimateFeedback}(\boldsymbol{\beta}^{(e)}, \boldsymbol{\phi}_m)$
    $\nabla \ell_m(\boldsymbol{\beta}^{(e)}) \leftarrow$
     $\texttt{SRG}(s_{1,m}, \ldots, s_{N,m}, \tilde{s}^*_m, \boldsymbol{\phi}_m, \boldsymbol{\beta}^{(e)})$
  end
  $\nabla \mathcal{L}_F(\boldsymbol{\beta}^{(e)}) = \frac{1}{M} \sum_{m=1}^{M} \nabla \ell_m(\boldsymbol{\beta}^{(e)})$
  $\boldsymbol{\beta}^{(e+1)} \leftarrow \mathbb{P}\left[\boldsymbol{\beta}^{(e)} - \delta \cdot \nabla \mathcal{L}_F(\boldsymbol{\beta})\right]$
end

---

regret $\mathcal{L}_F(\boldsymbol{\beta})$, which is given by

$$\mathcal{L}_F(\boldsymbol{\beta}) \triangleq \frac{1}{M} \sum_{m=1}^{M} \left(\frac{1}{N} \sum_{n=1}^{N} \| s_{n,m} - \tilde{s}^*_m(\boldsymbol{\beta}) \|_2^2 \right), \tag{5}$$

where $\tilde{s}^*_m(\boldsymbol{\beta})$ represents the social fairness evaluation which follows the same definition in Equation (4), but without having the participant index n. For the same reason, the participant index n does not appear in Equations (1), (2), and (3) as well, for the computation of social fairness evaluation $\tilde{s}^*_m(\boldsymbol{\beta})$.

The social preference weight $\boldsymbol{\beta}^*$ can be learned using *Social Aggregation of Fairness Feedback* (SAFF) algorithm as shown in Algorithm 1, which is developed using projected gradient descent. The projection operator $\mathbb{P}$ ensures that $\boldsymbol{\beta}^*$ is a valid preference weight vector that has entries between 0 to 1 and sums to 1. The regret gradient $\nabla \mathcal{L}_F$ with respect to the model parameters $\boldsymbol{\beta}$ is computed using the well-known *backpropagation* algorithm. Since the feedback regret indirectly depends on the model parameter $\boldsymbol{\beta}$, each auxiliary term is expanded until direct dependency is achieved. Hence, the regret gradient $\nabla \mathcal{L}_F$ with respect to the model parameters $\boldsymbol{\beta}$ is computed using the following dependency chain of variables:

$$\mathcal{L}_F \xleftarrow{(5)} \tilde{s}^* \xleftarrow{(4)} \boldsymbol{u} \xleftarrow{(2),(3)} \boldsymbol{\psi} \xleftarrow{(1)} \boldsymbol{\beta}, \tag{6}$$

where the text above the arrows represent the Equation labels corresponding to their respective relationship. Consequently, the gradient of each dependent variable with respect to the model parameter $\boldsymbol{\beta}$ has to be computed. Therefore, the backpropagation of gradients is given by

$$\nabla_{\boldsymbol{\beta}} \mathcal{L}_F = (\nabla_{\tilde{s}^*} \mathcal{L}_F)^T \cdot \nabla_{\boldsymbol{\beta}} \tilde{s}^* \tag{7a}$$

$$\nabla_{\boldsymbol{\beta}} \tilde{s}^* = (\nabla_{\boldsymbol{u}} \tilde{s}^*)^T \cdot \nabla_{\boldsymbol{\beta}} u \tag{7b}$$

$$\nabla_{\boldsymbol{\beta}} u = (\nabla_{\boldsymbol{\psi}} u)^T \cdot \nabla_{\boldsymbol{\beta}} \boldsymbol{\psi} \tag{7c}$$

where the gradient $\nabla_{\boldsymbol{q}} \boldsymbol{p}$ is a $P \times Q$ matrix, where $\boldsymbol{p}$ is a $P \times 1$ vector, and $\boldsymbol{q}$ is a $Q \times 1$ vector, for any general $\boldsymbol{p}$ and $\boldsymbol{q}$. Note that the gradients $\nabla_{\tilde{s}^*} \mathcal{L}_F$, $\nabla_{\boldsymbol{u}} \tilde{s}^*$, $\nabla_{\boldsymbol{\psi}} u$ and $\nabla_{\boldsymbol{\beta}} \boldsymbol{\psi}$ in Equations (7a), (7b) and (7c) can be respectively computed as

$$\nabla_{\tilde{s}^*} \mathcal{L}_F = 2 \left[\frac{1}{M} \sum_{m=1}^{M} \tilde{s}_m^*(\boldsymbol{\beta}) - \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} s_{n,m} \right], \tag{8}$$

$\nabla_{\boldsymbol{u}_m} \tilde{s}_m^*$ is a 7×7 matrix, where the $(i,k)^{th}$ entry $\eta_{i,k}$ is given by

$$\eta_{i,k} = \begin{cases} \dfrac{\lambda}{\Delta_m^2} \cdot e^{\lambda u_{m,i}} \cdot \displaystyle\sum_{j \neq i} e^{\lambda u_{m,j}}, & \text{if } i = k, \\ -\dfrac{\lambda}{\Delta_m^2} \cdot e^{\lambda u_{m,i}} \cdot e^{\lambda u_{m,k}}, & \text{otherwise,} \end{cases} \tag{9}$$

with $\Delta_m = \sum_{j=1}^{7} e^{\lambda \cdot u_{m,j}}$ being the normalizing factor,

$$\begin{aligned} \nabla_{\psi_m} u_{m,i} = \frac{1}{\sigma^2} \Bigg[& \frac{\sigma}{\sqrt{2\pi}} \exp\left\{ -\frac{(z_{i-1} - \psi_m)^2}{2\sigma^2} \right\} \\ & - \frac{\sigma}{\sqrt{2\pi}} \exp\left\{ -\frac{(z_i - \psi_m)^2}{2\sigma^2} \right\} + \frac{\psi_m}{2} \operatorname{erf}\left(\frac{z_i - \psi_m}{\sigma\sqrt{2}} \right) \\ & - \frac{\psi_m}{2} \operatorname{erf}\left(\frac{z_{i-1} - \psi_m}{\sigma\sqrt{2}} \right) - \psi_m u_{m,i} \\ & + \frac{\sigma}{\sqrt{2\pi}} \exp\left\{ -\frac{(z_{14-i} - \psi_m)^2}{2\sigma^2} \right\} \\ & - \frac{\sigma}{\sqrt{2\pi}} \exp\left\{ -\frac{(z_{14-i+1} - \psi_m)^2}{2\sigma^2} \right\} \\ & + \frac{\psi_m}{2} \operatorname{erf}\left(\frac{z_{14-i+1} - \psi_m}{\sigma\sqrt{2}} \right) \\ & - \frac{\psi_m}{2} \operatorname{erf}\left(\frac{z_{14-i} - \psi_m}{\sigma\sqrt{2}} \right) \Bigg], \end{aligned} \tag{10}$$

where $z_i = \texttt{Logit}(b_i)$, and

$$\nabla_{\boldsymbol{\beta}} \psi_m = \boldsymbol{\phi}(\boldsymbol{d}_m). \tag{11}$$

Algorithm 2: SRG

Input: $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N, \tilde{s}^*, \boldsymbol{\phi}, \boldsymbol{\beta}$
Output: Feedback Regret Gradient $\nabla \mathcal{L}_F(\boldsymbol{\beta})$

Compute $\nabla_{\boldsymbol{\beta}} \psi_m$ using the Equation (11)
Compute $\nabla_{\psi_m} u_{m,i}$ using the Equation (10)
Compute $\nabla_{\boldsymbol{u}_m} \tilde{s}_m^*$ using the Equation (9)
Compute $\nabla_{\tilde{s}^*} \mathcal{L}_F$ using the Equation (8)

The method of computing the gradient of social regret is called *Social Regret Gradient* (SRG), which is formally presented in Algorithm 2.

5. Evaluation Methodology

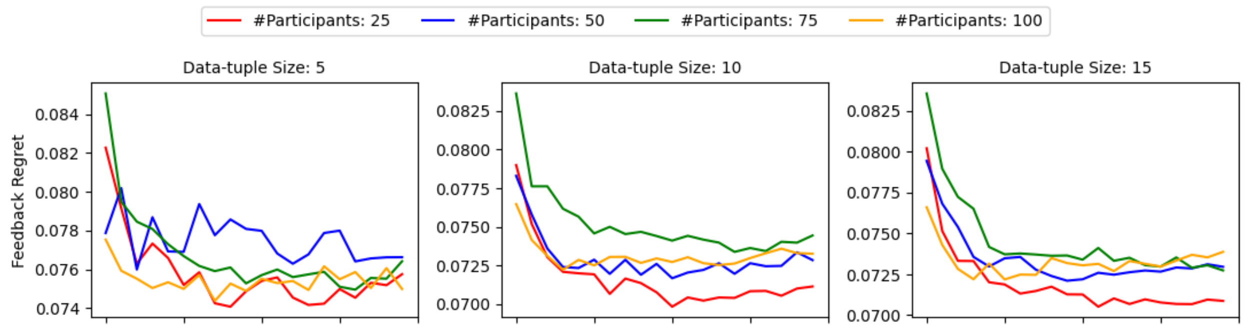
The proposed algorithm SAFF is employed on both simulated data as well as survey responses. This paper considers $L = 6$ group fairness notions (see Table 4) to evaluate the Acceptance Rate Predictor (ARP) with respect to the sensitive attributes *race* ={Black, All Other Races}, *gender* = {Male, Female}, and *age* = {<50, >50}. In addition, the privileged and underprivileged groups are defined as $\mathcal{X}_m$ = {Other, Male, <50} and $\mathcal{X}_{m'}$ = {Black, Female, >50}, respectively. The predicted probability of kindey acceptance from the ARP is discretized into binary, where the probability ≥ 0.5 indicates acceptance ($\hat{y} = 1$), and probability < 0.5 indicates rejection ($\hat{y} = 0$). The computation of various group fairness scores is elaborated in Appendix C.

5.1. Evaluation on Simulated Data

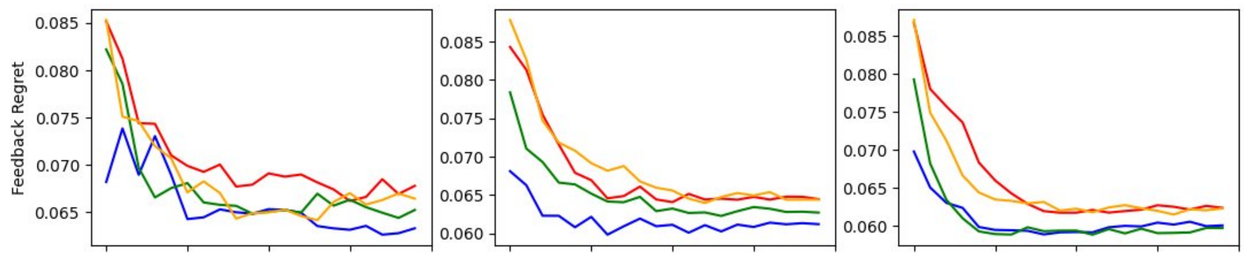
For simulation experiments, the true preferences of the non-expert participants $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_N$ are constructed by randomly assigning preference values for all $L = 6$ fairness notions based on uniform distribution. Similarly, the estimated social preference is also initialized with random values based on uniform distribution. The estimated social preference $\boldsymbol{\beta}^*$ is updated over $M = \{5, 10, 15\}$ data tuples each containing $K = 10$ donor-recipient pairs. The results are averaged across 100 iterations for all $N = \{25, 50, 75, 100\}$ non-expert participants. The learning rate is declared as $\delta = 0.1$ and the number of epochs as 20.

5.2. Evaluation on ARP Survey

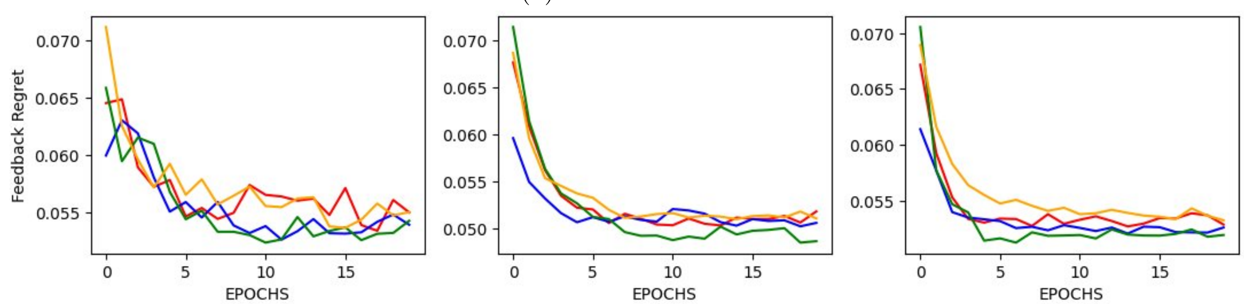
Unlike simulation experiment, the true preferences of the participants are unknown in the survey experi-



(a) Age Attribute



(b) Race Attribute



(c) Gender Attribute

Figure 5: Convergence of Feedback Regret across Different Data-Tuple Sizes

ment. The estimated social preference $\boldsymbol{\beta}^{(0)}$ is initialized randomly based on uniform distribution. Note that the participants rate the fairness of ARP on a Likert scale of 1 to 7, $\boldsymbol{s}_n \in \{1, 2, \cdots, 7\}$. The estimated social preference $\boldsymbol{\beta}^{(0)}$ is updated over $M = 10$ data-tuples each containing $K = 10$ donor-recipient pairs presented to $N = 75$ participants.

6. Results and Discussion

6.1. Simulation Results

Figure 5 illustrates feedback regret for varying numbers of participants, $N = \{25, 50, 75, 100\}$, with each receiving $M = \{5, 10, 15\}$ data-tuples. Figure 5(a) demonstrates the social feedback regret with respect to the age attribute computed using the participants' responses to the question Q2 (refer Figure 2). Similarly, Figure 5(b) depicts the social feedback regret with respect to the race computed using the responses received from question Q3. On the other hand, Figure 5(c) shows the convergence of social feedback regret with respect to the gender computed using the responses from question Q4.

Note that the preference regret converges with increasing number of epochs for any sensitive attribute and any combination of data tuple size and the number of participants. However, the increase in the number of participants and/or data tuple size has little improvement on social feedback regret.

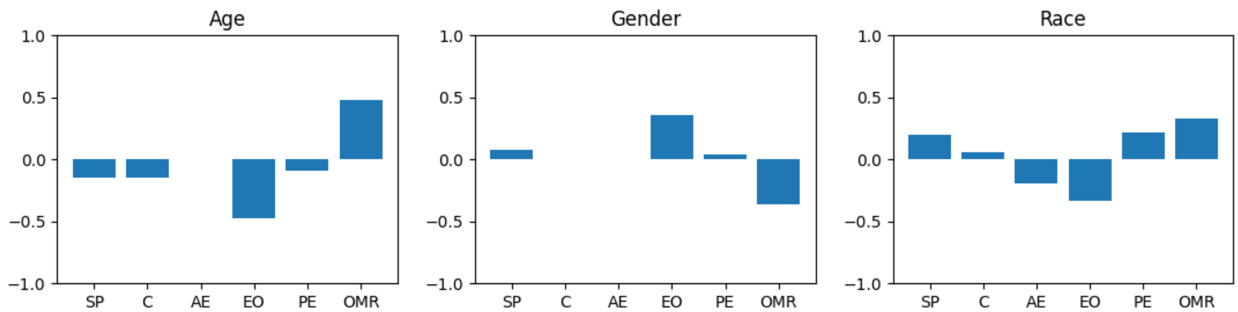


Figure 6: Group Fairness Evaluations of the ARP across Different Sensitive Attributes.

Sensitive Attribute	**Social Fairness Preference**					
	SP	**C**	**AE**	**EO**	**PE**	**OMR**
Age	0.15	0	**0.45**	0.007	0.37	0.01
Gender	0.19	0.02	**0.48**	0	0.24	0.06
Race	0.28	0.10	**0.38**	0	0.19	0.03

Table 3: Social Fairness Preferences of the Recruited Participants over $L = 6$ Group Fairness Notions

Additionally, the simulation results also suggest that our hypothesis on modeling participants' preferences using Mixed-Logit model holds true. Specifically, given that the simulation results yielded low feedback regret (as shown in Figure 5), it indicates that the proposed model is able to estimate a social preference that complies with individual participant preferences.

On the contrary, it is possible that the participants' fairness preferences may change over longer time horizons due to diverse reasons such as changes in societal norms, technological advancements and dynamics in social demographics. Therefore, it is necessary to repeat the proposed experiment on a reasonably regular basis in order to capture updated social fairness preferences regarding kidney placement.

Initialization: The proposed algorithm converges quite well, as demonstrated in Figure 5, when the preference weights in the proposed model are initialized as random weight vectors. However, the same approach does not exhibit the desired convergence when the social preferences are initialized to equal preference, i.e. $\beta_l = 1/6$ for all $l = 1, \cdots, 6$.

6.2. Survey Results

Table 3 shows the estimated social preferences of the recruited participants over $L = 6$ group fairness notions in the Prolific survey experiment. Note that *accuracy equality* (AE) is the preferred group fairness notion across all three sensitive attributes. Note that the ARP is perceived to exhibit less bias in terms of accuracy equality across all three sensitive attributes (as shown in the Figure 6). In the case of age and gender, *predictive equality* (PE) has the second highest preference over the six group fairness notions. Even from the perspective of PE, the ARP exhibits little/no bias wit respect to all the three sensitive attributes. On the contrary, although the ARP is perceived to have no bias in terms of *calibration*, the social fairness preference is close to zero with respect to both age and gender.

At the same time, the ARP seems unfair in terms of *equal opportunity* (EO) with evaluations ranging to -0.5 with respect to age, and 0.46 with respect to gender (as depicted in Figure 6). However, EO is the least preferred fairness notion, with almost negligible preference weight for all the three sensitive attributes, as shown in the Table 3. Similar observations can be made with *overall misclassification rate* (OMR) as well. Although the ARP is unfair in terms of OMR, the non-expert participants clearly do not prefer OMR. Therefore, group fairness notions such as C, EO and OMR have little role in public's fairness evaluation regarding the U.S. kidney placement.

In summary, accuracy equality and predictive equality can be deemed as critical group fairness notions from the public stakeholders' viewpoint. Furthermore, as a follow-up to the above claim, it is also natural to conclude that the non-expert participants' perceive ARP as a *reasonably* fair system when deployed in the kidney placement pipeline. However, policy makers and regulatory authorities can suggest ARP developers to develop fairer predictions that align with estimated social fairness preferences. In the future, we plan to design novel survey experiments that are designed to obtain different types of feedback from each individual stakeholder group depending on their expertise on various tasks within the kidney placement pipeline. For instance, *transplant surgeons* can evaluate ARP's fairness based on medical attributes (e.g. diabetes status, glomerular filtration rate) of donors and recipients, unlike the public survey presented in this paper. Additionally, transplant surgeons can also provide an independent evaluation on the likelihood of recipient's post-transplant survival for a given donor/recipient pair, as feedback in the survey. On the contrary, survey experiments for *Organ Procurement Organizations* (OPOs) can be designed to evaluate the fairness of predicting donor kidney utilization based on donor's creatinine levels, cause of death, gender and race. Furthermore, note that the fairness preferences of any society can change over time due to various factors such as changes in societal norms, technological advancements in predictive analytics and dynamics in social demographics. Therefore, policy makers and regulatory authorities should regularly repeat such experiments to learn the current fairness preferences across diverse stakeholders.

Acknowledgments

This work is supported by the National Science Foundation (Award #2222801). We express our gratitude to Ashiku et al. (2022) for granting access to their proposed machine learning model. Additionally, we extend our appreciation to the United Network of Organ Sharing (UNOS) for support this project and for providing the kidney matching datasets.

References

Lirim Ashiku, Richard Threlkeld, Casey Canfield, and Cihan Dagli. Identifying AI Opportunities in Donor Kidney Acceptance: Incremental Hierarchical Systems Engineering Approach. In *2022 IEEE International Systems Conference (SysCon)*, pages 1–8. IEEE, 2022.

Masoud Barah and Sanjay Mehrotra. Predicting Kidney Discard using Machine Learning. *Transplantation*, 105(9):2054, 2021.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 50(1):3–44, 2018.

U.S. Census Bureau. ACS Demographic and Housing Estimates. U.S. Census Bureau. URL https://data.census.gov/table/ACSDP5Y2021.DP05?g=010XX00US&y=2021&d=ACS5-YearEstimatesDataProfiles.

Simon Caton and Christian Haas. Fairness in Machine Learning: A Survey. *arXiv preprint arXiv:2010.04053*, 2020.

Nick Charalambides. We Recently Went Viral on TikTok - Here's What We Learned. August 2021.

Kelly Chong, Igor Litvinovich, Shan Shan Chen, Yiliang Zhu, Christos Argyropoulos, and Yue-Harn Ng. Reconsidering Donor Race in Predicting Allograft and Patient Survival among Kidney Transplant Recipients. *Kidney360*, 2(11):1831, 2021.

Alexandra Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big data*, 5(2):153–163, 2017.

Alexandra Chouldechova and Aaron Roth. The Frontiers of Fairness in Machine Learning. *arXiv preprint arXiv:1810.08810*, 2018.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic Decision Making and The Cost of Fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness Through Awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

Benjamin Eidelson. Kidney Allocation and The Limits of the Age Discrimination Act. *Yale LJ*, 122: 1635, 2012.

Golnoosh Farnadi, William St-Arnaud, Behrouz Babaki, and Margarida Carvalho. Individual Fairness in Kidney Exchange Programs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11496–11505, 2021.

John J Friedewald, Ciara J Samana, Bertram L Kasiske, Ajay K Israni, Darren Stewart, Wida Cherikh, and Richard N Formica. The Kidney Allocation System. *Surgical Clinics*, 93(6):1395–1406, 2013.

Mesfin G Genie, Antonio Nicoló, and Giacomo Pasini. The Role of Heterogeneity of Patients' Preferences in Kidney Transplantation. *Journal of Health Economics*, 72:102331, 2020.

Nina Grgic-Hlaca, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference*, pages 903–912, 2018.

M. Hardt, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3315–3323. Curran Associates, Inc., 2016.

Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. An Empirical Study on the Perceived Fairness of Realistic, Imperfect Machine Learning Models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 392–402, 2020.

David H Howard. Quality and Consumer Choice in Healthcare: Evidence from Kidney Transplantation. *The BE Journal of Economic Analysis & Policy*, 5(1):0000101515153806531349, 2006.

Maude Lavanchy, Patrick Reichert, Jayanth Narayanan, and Krishna Savani. Applicants' Fairness Perceptions of Algorithm-Driven Hiring Procedures. *Journal of Business Ethics*, pages 1–26, 2023.

Krista L Lentine, Jodi M Smith, Jonathan M Miller, Keighly Bradbrook, Lindsay Larkin, Samantha Weiss, Dzhuliyana K Handarova, Kayla Temple, Ajay K Israni, and Jon J Snyder. OPTN/SRTR 2021 Annual Data Report: Kidney. *American Journal of Transplantation*, 23(2):S21–S120, 2023.

John C Magee, John C Bucuvalas, Douglas G Farmer, William E Harmon, Tempie E Hulbert-Shearon, and Eric N Mendeloff. Pediatric Transplantation. *American Journal of Transplantation*, 4:54–71, 2004.

Nicholas Mattei, Abdallah Saffidine, and Toby Walsh. Fairness in Deceased Organ Matching. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 236–242, 2018.

Ian McCulloh, Darren Stewart, Kevin Kiernan, Ferben Yazicioglu, Heather Patsolic, Christopher Zinner, Sumit Mohan, and Laura Cartwright. An Experiment on the Impact of Predictive Analytics on Kidney Offer Acceptance Decisions. *American Journal of Transplantation*, 23:957–965, 2023.

Daniel McFadden et al. Conditional Logit Analysis of Qualitative Choice Behavior. 1973.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Dana Pessach and Erez Shmueli. Algorithmic Fairness. *arXiv preprint arXiv:2001.09784*, 2020.

Julien Rouzot, Julien Ferry, and Marie-José Huguet. Learning Optimal Fair Scoring Systems for Multi-Class Classification. In *ICTAI 2022-The 34th IEEE International Conference on Tools with Artificial Intelligence*, 2022.

Megha Srivastava, Hoda Heidari, and Andreas Krause. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2459–2468, 2019.

Kenneth E Train. *Discrete choice methods with simulation.* Cambridge university press, 2009.

Anne M Turner, Thomas Engelsma, Jean O Taylor, Rashmi K Sharma, and George Demiris. Recruiting Older Adult Participants through Crowdsourcing Platforms: Mechanical Turk versus Prolific Academic. In *AMIA Annual Symposium Proceedings*, volume 2020, page 1230. American Medical Informatics Association, 2020.

Wenbin Zhang, Tina Hernandez-Boussard, and Jeremy Weiss. Censored Fairness through Awareness. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 14611–14619, 2023.

Appendix A. Kidney Placement in the United States

The term kidney placement refers to the process of procuring kidneys and identifying potential recipients for transplant surgery based on several donor/recipient characteristics, as well as location proximity. In the United States, organ procurement and transplantation are led by the United Network of Organ Sharing (UNOS), where donors in the Organ Procurement Organizations (OPOs) are matched with patients waiting for organs in the Transplant Centers (TXCs). The OPOs are responsible for procuring the organs, evaluating them for quality using Kidney Donor Profile Index (KDPI) score, and maintaining a donor registry. The KDPI score, ranging from 0 to 100, is computed using donor characteristics such as donor's age, height, race, and history of hypertension, where 0 indicates high quality and 100 indicates low quality. On the other hand, the TXCs are responsible for evaluating recipients on the waiting list using Estimated Post Transplant Survival (EPTS) score and performing transplant surgery. The EPTS score, also ranging from 0 to 100, is computed using patient attributes such as patient's age, years on dialysis, and diabetes status, where 0 implies longer life expectancy and 100 implies shorter life expectancy. Once a deceased donor kidney is identified as suitable, it will be matched with the candidates in the waiting list based on scores computed from KDPI and EPTS (Friedewald et al., 2013). Thereafter, the potential recipients for a specific deceased donor kidney are ranked based on geographic location and medical urgency. As of now, a single deceased donor kidney can be matched with thousands of potential recipients and at most two of them will undergo kidney transplantation.

Appendix B. Survey Information

First, the recruited participants are presented with a brief overview of the kidney placement process in the United States which includes information regarding the transplant centers, kidney offers, identifying potential recipient, and transportation of the donor kidney. In the next page, instructions regarding the survey experiment is detailed. Specifically, this page explains how the data-tuple is represented, different donor-recipient attributes involving in a data-tuple, and what is expected from the participants (as shown in Figure 7).

Appendix C. Group Fairness Notions

Over the past decade, several group fairness notions have been proposed to measure the biases in a given system. Such fairness notions seek for parity of some statistical measure (e.g. true positive rate, predictive parity value) be equal across all the sensitive attributes (e.g. race) present in the data. Specifically, group fairness notions measure the difference in a specific statistical measure between protected (e.g. Caucasians) and unprotected (e.g. African-Americans) groups of a sensitive attribute. Different versions of group-conditional metrics led to different statistical definitions of fairness Caton and Haas (2020); Chouldechova and Roth (2018); Mehrabi et al. (2021); Pessach and Shmueli (2020). Let $y \in \mathcal{Y}$ as the true label and $\hat{y} = g(x) \in \mathcal{Y}$ as the predicted label given by the ML-based system for some input $x \in \mathcal{X}$. Furthermore, let $\mathcal{X}_m, \mathcal{X}_{m'} \in \mathcal{X}$ denote the protected and unprotected sensitive groups respectively. The *unfairness* within the acceptance predictor can be evaluated based on several group fairness notions which can be generalized as

$$\phi_f \triangleq \phi_f(m) - \phi_f(m'), \tag{12}$$

for any $\mathcal{X}_m, \mathcal{X}_{m'}$, and $\phi_f(m)$ denotes the groupwise rate with respect to the group $\mathcal{X}_m$. The groupwise rates considered in this work are listed in Table 4.

Brief Overview of the Kidney Transplant Placement Process

1. Information about the kidney is sent to transplant centers. This is based on a ranked list of transplant candidates based on need, proximity, and medical compatibility.
2. Transplant centers accept or decline the kidney. This is based on the kidney's quality, its fit with the candidate, time of day, and ranking of the candidate.
3. The kidney is shipped to the transplant center. The highest ranked candidate that accepts the kidney gets it. The surgeon then transplants the kidney.

Problems: Problems can occur at any of these steps. This kidney might be lower quality and therefore less desirable if it came from someone with health problems. This kidney might not be compatible with the transplant candidate. This information might not be discovered until certain tests are done. Additionally, if it takes too long, the donated kidney will not be able to be used.

Instructions

We need your help in evaluating the outcomes given an ML Predictor. We are interested in your assessments of fairness.

First, we will show you information about a deceased donor which will include their age, race/ethnicity, gender, and kidney quality. The kidney quality is a score that ranges from 0-100 where **0 is high quality** and 100 is low quality.

Then you will see information about 10 potential transplant recipients. They have already been matched with the donor based on basic medical criteria like blood type. This will include their age, race, gender, estimated post-transplant survival score, and distance from the transplant center. To clarify:

- The **estimated post-transplant survival score** is a score that ranges from 0-100 where **0 indicates longer life expectancy** and 100 shorter life expectancy after transplant.
- The **distance from the transplant center** matters because the longer the kidney is outside the body, the greater the risk of damage and reduced transplant success. Moreover, reducing the distance simplifies logistical challenges such as organ transport.
- The **surgeon's decision** (highlighted in purple) is the action taken by the transplant surgeon in reality to either transplant or not transplant.

We will ask for your fairness assessment regarding following ML Predictor:

- **Acceptance Predictor** (highlighted in blue) – this AI model predicts the probability of a recipient accepting the specific deceased donor kidney i.e, Acceptance Rate. This is based on clinical characteristics (such as glomeruli count, creatinine levels, history of diabetes, hypertension, weight, height) as well as demographics (such as gender, race, age).

We will ask for your fairness assessment for 10 such donor-recipient information pairs.

Figure 7: Kidney Placement Overview and the Survey Instructions Presented to the Participants.

Table 4: Group Fairness Notions Used in the Learning of Participants' Fairness Preferences

Index (l)	**Group Fairness Notion (f)**	**Groupwise Rate $\phi_f(m)$**
1	Statistical Parity (SP) (Dwork et al., 2012)	$\phi_{SP}(m) = \mathbb{P}(\hat{y} = 1 \mid x \in \mathcal{X}_m)$
2	Calibration (C) (Chouldechova, 2017)	$\phi_C(m) = \mathbb{P}(y = 1 \mid \hat{y} = 1, x \in \mathcal{X}_m)$
3	Accuracy Equality (AE) (Berk et al., 2018)	$\phi_{AE}(m) = \mathbb{P}(\hat{y} = y \mid x \in \mathcal{X}_m)$
4	Equal Opportunity (EO) (Hardt et al., 2016)	$\phi_{EO}(m) = \mathbb{P}(\hat{y} = 1 \mid y = 1, x \in \mathcal{X}_m)$
5	Predictive Equality (PE) (Corbett-Davies et al., 2017)	$\phi_{PE}(m) = \mathbb{P}(\hat{y} = 1 \mid y = 0, x \in \mathcal{X}_m)$
6	Overall Misclassification Rate (OMR) (Rouzot et al., 2022)	$\phi_{OMR}(m) = \mathbb{P}(\hat{y} = 0 \mid y = 1, x \in \mathcal{X}_m)$