A DATA ENVELOPMENT ANALYSIS APPROACH FOR ASSESSING FAIRNESS IN RESOURCE ALLOCATION: APPLICATION TO KIDNEY EXCHANGE PROGRAMS

ALI KAAZEMPUR-MOFRAD[1,a] AND XIAOWU DAI[1,*,b]

[1]*Department of Statistics and Data Science, University of California, Los Angeles*
[a]*amofrad@ucla.edu* [b]*dai@stat.ucla.edu*

Kidney exchange programs have significantly increased transplantation rates but raise pressing questions about fairness in organ allocation. We present a novel framework leveraging Data Envelopment Analysis (DEA) to evaluate multiple fairness criteria—Priority, Access, and Outcome—within a single model, capturing complexities that may be overlooked in single-metric analyses. Using data from the United Network for Organ Sharing, we analyze these criteria individually, measuring Priority fairness through waitlist durations, Access fairness through Kidney Donor Profile Index scores, and Outcome fairness through graft lifespan. We then apply our DEA model to demonstrate significant disparities in kidney allocation efficiency across ethnic groups. To quantify uncertainty, we employ conformal prediction within the DEA framework, yielding group-conditional prediction intervals with finite sample coverage guarantees. Our findings show notable differences in efficiency distributions between ethnic groups. Our study provides a rigorous framework for evaluating fairness in complex resource allocation systems, where resource scarcity and mutual compatibility constraints exist. All code for using the proposed method and reproducing results is available on GitHub.

1. Introduction. Chronic kidney disease (CKD) represents a significant and growing public health challenge in the United States and globally. According to the Centers for Disease Control and Prevention (2023), CKD affects approximately 14% of U.S. adults, or 35.5 million people. This prevalence has been steadily increasing over the past decades, driven by rising rates of diabetes, hypertension, and an aging population (Coresh et al., 2007). Of particular concern is the subset of CKD patients who progress to end-stage renal disease (ESRD), a life-threatening condition where the kidneys have lost over 85% of their function. In 2020, over 805,000 Americans were living with ESRD, with 130,719 new cases diagnosed that year (United States Renal Data System, 2023). For these patients, kidney transplantation offers the best outcomes in terms of quality of life, long-term survival, and cost-effectiveness compared to dialysis (Wolfe et al., 1999; Axelrod et al., 2018). However, the stark reality is that the demand for donor kidneys far exceeds the available supply. For example, in 2020, while 23,642 kidney transplants were performed in the U.S., nearly 100,000 patients remained on the waiting list (Lentine et al., 2022). This critical shortage has spurred innovations in transplantation strategies, with kidney exchange programs emerging as a particularly promising solution. These programs, first proposed by Rapaport (1986) and later formalized by Roth, Sönmez and Ünver (2004), allow patients with willing but incompatible living donors to exchange donors with other incompatible pairs. Kidney paired donation (KPD) programs have significantly increased the number of living donor transplants performed annually in the United States (Massie et al., 2013), offering new hope to patients who might otherwise face prolonged waiting times or potential ineligibility for transplantation. These programs have been

*Corresponding author.
Keywords and phrases: Data Envelopment Analysis, Fairness, Resource Allocation, Kidney Exchange, Decision Making.

particularly beneficial for hard-to-match patients, including those with high panel reactive antibody levels or blood type O recipients (Gentry, Montgomery and Segev, 2011). Moreover, recent advancements such as global kidney exchange (Rees et al., 2017) and the incorporation of compatible pairs (Gentry et al., 2007) have further enhanced the potential of these programs, expanding the pool of potential donors and recipients.

While kidney exchange programs have undoubtedly improved overall transplantation rates, they raise pressing questions about fairness in organ allocation. These programs, typically designed to maximize the total number of transplants, may inadvertently disadvantage certain subgroups of the population (Agarwal et al., 2019). Examining fairness in such systems poses significant challenges, particularly when multiple, potentially conflicting criteria must be considered simultaneously. Throughout this paper, we focus on group fairness, which aims for parity across demographic groups (Hardt, Price and Srebro, 2016). This concept is particularly relevant in kidney allocation, where disparities in access and outcomes among different ethnic groups have been well-documented. For instance, Malek et al. (2011) found that only 57% of Asian patients received a transplant within five years compared to nearly 70% of White patients, while Fan et al. (2010) found that Asian patients experience the highest graft survival rates among all ethnic groups. Such conflicting outcomes across different metrics underscore the need for a comprehensive framework to evaluate and enhance fairness in kidney allocation systems.

We propose a general framework for evaluating fairness in kidney allocation using Data Envelopment Analysis (DEA). Originally developed by Charnes, Cooper and Rhodes (1978) for measuring the relative efficiency of decision-making units, DEA offers a flexible approach for evaluating multiple inputs and outputs without requiring a priori specification of their relative importance. Our framework employs a hyperbolic graph efficiency DEA model (Färe, Grosskopf and Lovell, 2013), allowing for the simultaneous consideration of multiple fairness criteria. Prior to applying the DEA framework, we conduct individual analyses of three key fairness components—Priority, Access, and Outcome Fairness—using specialized models to show complex and sometimes conflicting patterns of disparity across ethnic groups. For Priority Fairness, we employ mediation analysis, aligned with the work of Imai, Keele and Yamamoto (2010), to disentangle the direct and indirect effects of patient characteristics on waiting times. Access Fairness is evaluated through counterfactual analysis of KDPI score disparities (Kusner et al., 2017). Outcome Fairness evaluates graft survival through competing risks analysis (Pinto-Ramirez et al., 2022) to examine graft rejection risks in the presence of other competing causes of graft failure. Our analysis shows that while Asian patients face longer waitlist times, they demonstrate better graft survival outcomes. Conversely, White patients enjoy shorter waitlist durations but face higher risks of graft failure due to rejection. Black patients, while experiencing moderate waitlist times, face the highest risk of graft rejection. These conflicting results highlight the challenge of achieving fairness across all criteria simultaneously in kidney allocation. Our DEA approach addresses these complexities by concurrently minimizing unfavorable inputs (e.g., long waiting times) while maximizing desirable outputs (e.g., graft survival outcomes). The resulting hyperbolic efficiency measure provides a single, comprehensive score that captures how well the allocation system serves each patient or subgroup, considering multiple fairness criteria simultaneously.

We integrate uncertainty quantification through conformal prediction techniques, following the approach of Gibbs, Cherian and Candès (2023). This novel application within the DEA framework generates group-conditional prediction intervals for efficiency scores, offering finite sample coverage guarantees. Our approach provides a new method for evaluating fairness in kidney allocation and offers several key contributions to the field. By mapping efficiency scores across demographic groups and comparing their distributions, we identify potential disparities and areas for improvement in the allocation system, which may not be

apparent when considering individual fairness criteria in isolation. The comprehensive efficiency score offers an interpretable and unified measure of overall system fairness for each patient or subgroup, while integrating uncertainty quantification enhances the reliability of our findings.

Our fairness framework for kidney allocation expands on existing methods, which often rely on single metrics, such as mechanism design models (Ogryczak et al., 2014; Chorppath and Alpcan, 2011), multi-criteria decision analysis with predetermined weights (Moreno-Calderón, Tong and Thokala, 2020; Thokala et al., 2016), or specific survival comparisons (Cohen et al., 2016; Tonelli et al., 2011). The proposed DEA framework provides a unified efficiency measure that integrates multiple fairness criteria without requiring fixed weights or distributional assumptions. This approach can be applied beyond kidney exchanges to other resource allocation systems that face resource scarcity and compatibility constraints, such as healthcare (Chilingerian, 1995; Ozcan et al., 2008) and education (Shero et al., 2022), where fairness is critical.

The remainder of this paper is structured as follows. Section 2 introduces the method, including the DEA model and its uncertainty quantification. Section 3 shows the data preparation process and studies the different fairness criteria individually. Section 4 presents the results of applying the DEA framework to our data, along with statistical testing and inference. Section 5 concludes the paper with potential future research.

2. Methodology.

2.1. *Fairness Criteria in Kidney Allocation.* Evaluating fairness in kidney allocation systems requires multiple criteria to capture the complexities of the transplantation process (Persad, Wertheimer and Emanuel, 2009). Our analysis focuses on three key criteria: Priority Fairness, Access Fairness, and Outcome Fairness, with each targeting a critical aspect of the allocation system.

Priority Fairness. Priority Fairness reflects the fairness in waitlist duration across different ethnic groups, focusing on how fairly patients are prioritized for transplantation. In our model, let X_{1i} denote the waitlist duration (in days) for patient i. The goal is to evaluate disparities in X_1 across demographic groups, where shorter duration signifies higher priority. Studies by Wolfe et al. (1999) and Gill et al. (2004) emphasize the crucial impact of waitlist duration on fair and timely access to transplantation. We conduct mediation analysis in Section 3.3 to examine the direct effect of ethnicity on X_{1i} while considering mediators like blood type, panel reactive antibody levels, and dialysis status.

Access Fairness. Access Fairness considers the distribution of kidney quality, as indicated by the Kidney Donor Profile Index (KDPI). The KDPI is a numerical measure that combines various donor characteristics to summarize the quality of the donor's kidneys relative to other recovered kidneys (Park et al., 2020). It is expressed as a value between 0 and 1, where lower scores are associated with higher donor quality and expected longevity. In our model, let X_{2i} denote the KDPI score for the kidney allocated to patient i. By studying X_2 across ethnic groups, this criterion suggests whether there are disparities in the quality of organs distributed. In Section 3.3, we conduct a counterfactual analysis to estimate how X_{2i} might change if the ethnic identity of recipients were altered, thereby isolating the effect of ethnicity on organ quality allocation.

Outcome Fairness. Outcome Fairness evaluates the long-term success of kidney transplants by examining graft lifespan across different ethnic groups. Graft lifespan refers to the duration from transplantation to graft failure, which occurs when the transplanted kidney ceases to function (Wolfe et al., 2008). In our model, let Y_{1i} denote the graft lifespan (in days) for patient i. This criterion is critical for determining whether the allocation system results in fair long-term outcomes. It has been shown that graft lifespan is a key indicator of transplant success and patient quality of life (Hariharan, Israni and Danovitch, 2021). In Section 3.3, we conduct a competing risks analysis to model graft failure and rejection risks, allowing us to identify disparities in graft lifespan across different ethnic groups.

2.2. *Data Envelopment Analysis Framework.* Our study employs Data Envelopment Analysis (DEA) to evaluate the fairness of kidney allocation across multiple criteria. This approach, originally developed by Charnes, Cooper and Rhodes (1978), allows for a comprehensive analysis of efficiency in converting multiple inputs into outputs. In our context, we aim to simultaneously minimize certain criteria such as waitlist duration and KDPI score while maximizing other criteria like graft lifespan. This objective motivates our use of the hyperbolic graph efficiency measure in our DEA framework, an extension proposed by Färe, Grosskopf and Lovell (2013), where we treat the criteria to be minimized as inputs and those to be maximized as outputs.

Let $\{(X_i, Y_i)\}_{i=1}^n$ represent the set of input-output pairs for n patients, where $X_i = (x_{1i}, \ldots, x_{pi}) \in \mathbb{R}_+^p$ denotes the p inputs and $Y_i = (y_{1i}, \ldots, y_{qi}) \in \mathbb{R}_+^q$ denotes the q outputs for patient $i = 1, \ldots, n$. In our kidney allocation analysis, we consider two key pre-transplant factors as inputs, waitlist duration and KDPI score, represented by $X_i = (x_{1i}, x_{2i})$ with $p = 2$. The single output $Y_i = y_{1i}$ with $q = 1$ denotes graft lifespan. This structure enables the evaluation of how efficiently the kidney allocation system converts these pre-transplant factors into prolonged graft survival.

Central to our DEA framework is the production possibility set $\mathcal{T}$, which includes all feasible input-output combinations, defined as

$$\mathcal{T} = \left\{ (X, Y) \in \mathbb{R}_+^p \times \mathbb{R}_+^q \;\middle|\; X \geq \sum_{i=1}^n \lambda_i X_i,\ Y \leq \sum_{i=1}^n \lambda_i Y_i,\ \sum_{i=1}^n \lambda_i = 1,\ \lambda \geq 0 \right\}.$$

This set encompasses all feasible input-output combinations within the kidney allocation system, including both observed transplants and those constructed through convex combinations of the observed transplants. The convexity assumption of $\mathcal{T}$ holds in our context as it allows for the interpolation between observed allocation outcomes. Within this set $\mathcal{T}$, we employ the hyperbolic graph efficiency measure to quantify the efficiency of each patient's transplantation process. This measure employs a scaling factor θ, defined as

$$\theta = \min\{\theta > 0 | (\theta X, \theta^{-1} Y) \in \mathcal{T}\},$$

which enables the simultaneous and proportional adjustment of inputs and outputs. The hyperbolic measure is particularly well-suited to the kidney allocation setting, as it allows for the simultaneous reduction of inputs (waitlist time, KDPI) and increase of outputs (graft survival). This approach provides an intuitive interpretation as a single scalar value that facilitates comparisons across patient groups.

The efficiency frontier is a key concept in our DEA framework, representing optimal efficiencies for kidney allocation. It is formed by input-output combinations $(X, Y) \in \mathcal{T}$ for which there exists no other combination $(X', Y') \in \mathcal{T}$ such that $X' \leq X$ and $Y' \geq Y$, with at least one strict inequality. Points lying on this frontier have $\theta = 1$, representing maximum efficiency within our framework. For patients not on the frontier ($\theta < 1$), θ indicates the factor by which inputs must be reduced and output increased to reach the efficiency frontier.

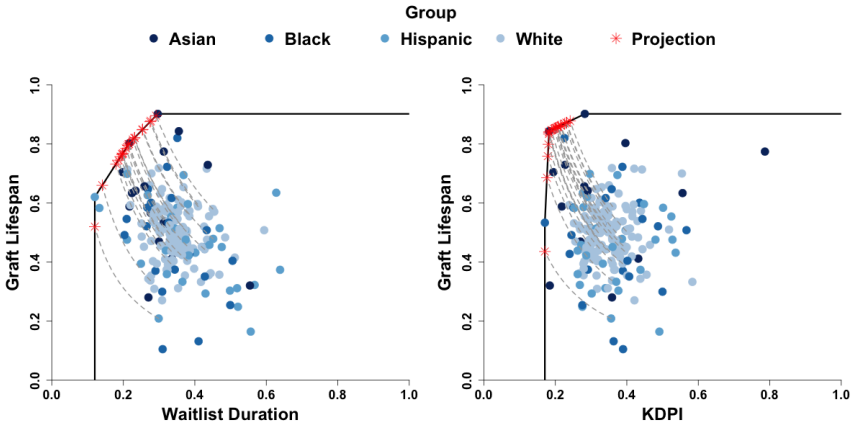


Fig 1. Illustration of the DEA efficiency frontier for a random subset of 200 patients, showing the hyperbolic trajectories and their projections onto the efficiency frontier, demonstrating how the input-output combinations of a patient could be scaled by θ to reach the frontier.

Figure 1 illustrates a simplified representation of our DEA framework using a random subset of 200 patients. Each plot shows one input (waitlist duration or KDPI score) against the output (graft lifespan). To enhance visibility across the full range of values, we apply min-max scaling to the waitlist duration and graft lifespan values to normalize these measures to a range of 0 to 1. In each plot, the solid curve represents the efficiency frontier, which serves as the optimal boundary for patient transplantation processes. The dashed lines, shown for a selected subset of points, represent the hyperbolic trajectories that each patient's input-output combination would follow if scaled by the factor θ until it reaches the frontier. The red stars at the ends of these paths mark the points where each patient would achieve $\theta = 1$, illustrating how the suggested input-output adjustments would project a patient onto the efficiency frontier.

Incorporating the hyperbolic graph efficiency measure into the DEA framework results in the following optimization problem for evaluating the efficiency of patient i:

$$\min_{\theta,\lambda_1,\ldots,\lambda_n} \theta \quad \text{s.t.} \quad \theta X_i \geq \sum_{j=1}^{n} \lambda_j X_j, \quad \frac{1}{\theta} Y_i \leq \sum_{j=1}^{n} \lambda_j Y_j, \tag{1}$$
$$\text{and} \quad \lambda_j \geq 0, \quad \sum_{j=1}^{n} \lambda_j = 1.$$

This model determines the optimal θ for each patient, representing their efficiency score relative to the frontier. The constraints ensure that the composite unit uses no more of each input than the patient being evaluated, produces at least as much as the patient being evaluated, and is a convex combination of observed patients, ensuring realistic benchmarks.

This framework enables us to compare the allocation efficiency across patients from different demographic groups on a common scale. By analyzing the distribution of efficiency scores across groups, we can identify disparities in the efficiency of allocation processes. For instance, consistently lower efficiency scores in one group may indicate less favorable transplantation outcomes relative to the efficient frontier, suggesting potential interventions for improvement.

2.3. *Uncertainty Quantification for the DEA Framework.* Uncertainty quantification provides confidence measures for our DEA efficiency scores in (1), accounting for input-output value measurement and sampling variability. This is essential for evaluating fairness

across ethnic groups by determining the statistical significance of efficiency score differences. We employ conformal predictions for its finite sample guarantees and ability to handle complex, non-linear relationships without distributional assumptions.

Let $\mathcal{G} = \{\text{Asian}, \text{Black}, \text{Hispanic}, \text{White}\}$ denote the set of ethnic groups in our study. To ensure group-conditional coverage in the uncertainty quantification, we define the function class as,

$$\mathcal{F} = \left\{ f(x,y) = \sum_{g \in \mathcal{G}} \beta_g \mathbb{I}\{(x,y) \in g\} \mid \beta_g \in \mathbb{R}, \forall g \in \mathcal{G} \right\}. \tag{2}$$

Here $\mathbb{I}\{(x,y) \in g\}$ is the indicator function for group membership of the input-output pair, and β_g are the coefficients for each group $g \in \mathcal{G}$. The indicator function allows us to isolate effects for each ethnic group, while the coefficients capture group-specific impacts on efficiency scores. We implement a conformal prediction method (Gibbs, Cherian and Candès, 2023), which guarantees that the prediction set satisfies

$$\mathbb{P}(\theta_{n+1} \in \hat{C}(X_{n+1}, Y_{n+1}) | (X_{n+1}, Y_{n+1}) \in g) = 1 - \alpha, \quad \forall g \in \mathcal{G},$$

where α is the desired miscoverage rate. This equation represents a conditional guarantee by specifying that the probability of θ_{n+1} being in the prediction set $\hat{C}(X_{n+1}, Y_{n+1})$, given that the input-output pair (X_{n+1}, Y_{n+1}) belongs to group g, equals $1 - \alpha$ for all $g \in \mathcal{G}$. Conditioning on $(X_{n+1}, Y_{n+1}) \in g$ makes this a group-specific guarantee, ensuring that the coverage probability holds for each ethnic group. We use the conformity score $\mathcal{S}((X,Y),\theta) = |\theta - \hat{\theta}(X,Y)|$, where $\hat{\theta}(X,Y) = \sum_{g \in \mathcal{G}} \beta_g \mathbb{I}\{(X,Y) \in g\}$ estimates the conditional mean efficiency score $\mathbb{E}[\theta|X,Y]$. Our approach involves solving the optimization problem:

$$\max_{\eta \in \mathbb{R}^{n+1}} \sum_{i=1}^{n} \eta_i \mathcal{S}_i + \eta_{n+1} \mathcal{S} - \mathcal{R}^*(\eta),$$

$$\text{subject to } -\alpha \leq \eta_i \leq 1 - \alpha,$$

where $\mathcal{R}^*(\eta)$ is the convex conjugate of the regularization term:

$$\mathcal{R}^*(\eta) = -\min_{h \in \mathcal{F}} \left\{ (n+1)\lambda \|h\|_2^2 - \sum_{i=1}^{n+1} \eta_i h(X_i, Y_i) \right\}.$$

Here $\mathcal{F}$ is defined in (2), and $\lambda > 0$ is the regularization parameter. The constraints $-\alpha \leq \eta_i \leq 1-\alpha$ maintain the desired coverage probability $(1-\alpha)$ for each ethnic group by limiting individual sample influence, preventing extreme adjustments that could lead to overfitting or biased results across groups. Our randomized prediction set is then defined as:

$$\hat{C}_{\text{rand.}}(X_{n+1}, Y_{n+1}) = \{\theta : \eta_{n+1}^{\mathcal{S}((X_{n+1}, Y_{n+1}),\theta)} < U\}, \tag{3}$$

where $U \sim \text{Unif}([-\alpha, 1-\alpha])$ is drawn independently, and $\eta_{n+1}^{\mathcal{S}}$ is the $(n+1)$-th component of the solution to the above optimization for a given $\mathcal{S}$. This approach is implemented using ten-fold cross-validation.

THEOREM 1. *Let $\mathcal{G}$ be a set of groups, and let $\{(X_1, Y_1, \theta_1), \ldots, (X_n, Y_n, \theta_n)\}$ be i.i.d. samples from some distribution P, where (X_i, Y_i) represents the input-output pair and θ_i is the DEA efficiency score from (1) for patient i. Let $\mathcal{F}$ be the function class as defined in (2). Then, for any new sample $(X_{n+1}, Y_{n+1}, \theta_{n+1})$, the randomized prediction set $\hat{C}_{rand.}(X_{n+1}, Y_{n+1})$ in (3) satisfies:*

$$\mathbb{P}(\theta_{n+1} \in \hat{C}_{rand.}(X_{n+1}, Y_{n+1}) | (X_{n+1}, Y_{n+1}) \in g) = 1 - \alpha, \quad \forall g \in \mathcal{G},$$

where α is the miscoverage rate.

This theorem shows that the prediction set (3) has group-conditional coverage guarantees for DEA efficiency scores. Hence, our approach, which integrates DEA efficiency scores with conformal prediction, quantifies uncertainties in a way that is sensitive to potential disparities across ethnic groups.

3. Preliminary Analyses and Fairness Criteria.

3.1. *Exploratory Data Analysis.* The United Network for Organ Sharing (UNOS) operates the Organ Procurement and Transplantation Network (OPTN), which has collected organ transplant data since 1987. Our study used this UNOS data, focusing specifically on kidney transplants conducted through exchange programs, as these represent our primary area of interest in evaluating fairness and efficiency in organ allocation. To provide a more accurate representation of the current landscape, we analyzed kidney transplant records from 2011 to 2021. During this period, 19,692 transplants were performed through kidney exchange programs. The variables include recipient and donor age, ethnicity, blood type, panel reactive antibody (PRA) levels, and waiting time for transplantation. Each case includes both pre-transplant factors (e.g., waitlist duration) and post-transplant outcomes (e.g., graft survival time). Table 1 presents summary statistics for key variables across different ethnic groups.

Table 1
Summary Statistics of Key Variables by Ethnic Group

Variable	Asian	Black	Hispanic	White
Waitlist Duration (days)				
Mean (SD)	1441 (959)	1348 (882)	1353 (929)	1139 (828)
Median [IQR]	1363 [715, 2034]	1282 [618, 1964]	1241 [588, 1988]	1022 [474, 1650]
KDPI Score				
Mean (SD)	0.452 (0.267)	0.416 (0.258)	0.390 (0.265)	0.396 (0.259)
Median [IQR]	0.435 [0.230, 0.670]	0.390 [0.190, 0.620]	0.360 [0.160, 0.600]	0.370 [0.170, 0.600]
Graft Survival (days)				
Mean (SD)	2065 (1010)	1970 (985)	1976 (979)	2120 (971)
Median [IQR]	2095 [1209, 2905]	1971 [1097, 2743]	1853 [1111, 2735]	2166 [1387, 2902]

These summary statistics show differences across ethnic groups in key transplantation metrics. Asian patients, on average, experience the longest waitlist durations compared to other ethnic groups, with a mean of 1441 days, followed closely by Hispanic (1353 days) and Black patients (1348 days). White patients have the shortest mean waitlist duration at 1139 days, substantially lower than other groups. Asian patients receive kidneys with slightly higher KDPI scores (mean 0.452) compared to other groups, potentially indicating lower quality organs, while Hispanic patients receive kidneys with the lowest mean KDPI score (0.390). In terms of graft survival, White patients show the longest mean survival time (2120 days or approximately 5.8 years), followed closely by Asian patients (2065 days or 5.7 years). Black and Hispanic patients have slightly shorter mean graft survival times (1970 and 1976 days respectively, or about 5.4 years). These observations suggest potential disparities in the kidney allocation process, with White patients appearing to have advantages in both waitlist duration and graft survival, while Asian patients face longer wait times but achieve comparable graft survival outcomes. These findings warrant further in-depth analysis to understand the underlying factors contributing to these differences across ethnic groups.

3.2. *Data Preprocessing: Resampling and Debiasing.* Our data preparation has two main steps: first, addressing data imbalance through resampling, and second, mitigating potential biases through a debiasing procedure.

According to the U.S. Census Bureau (2020), the composition of the U.S. population is as follows: Asian (6.1%), Black (12.6%), Hispanic (19.0%), and White (62.4%). However,

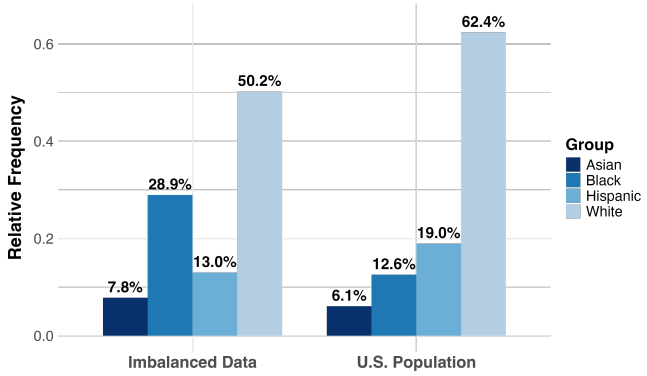


Fig 2. Comparison of the Distribution of Ethnic Groups between the Imbalanced Dataset and the U.S. Population.

our dataset shows substantial deviations from these proportions. Figure 2 illustrates the imbalance, comparing the relative frequency of each ethnic group in our dataset to the desired representation based on the U.S. population. Such imbalances can lead to biased estimates and incorrect inferences. It is shown that class imbalance can significantly impact the performance of learning algorithms, leading to biased predictions favoring the majority class (He and Garcia, 2009). In our case, this could result in efficiency estimates that are disproportionately influenced by the overrepresented groups, resulting in the bias $\mathbb{E}[\hat{\theta}_g] \neq \theta_g^*$, where $\hat{\theta}_g$ is the estimated efficiency scores for group g and θ_g^* is the true population efficiency.

Table 2
Distribution of Patients by Ethnicity - Post-Resampling

Ethnicity	Number of Patients	Proportion of Total (%)
Asian	818	6.1%
Black	1,691	12.6%
Hispanic	2,553	19.0%
White	8,403	62.4%
Total	13,465	

To correct this imbalance, we implemented a stratified sampling method. Let p_g be the desired proportion for group g based on the U.S. population distribution, and n_g be the number of samples in group g in the imbalanced dataset. We determine the total number of samples in the resampled dataset as $n = \min_{g \in \mathcal{G}} \lfloor n_g / p_g \rfloor$, ensuring sufficient samples for each group without oversampling. We then calculate the resampled group sizes as $\tilde{n}_g = \lfloor n \cdot p_g \rfloor, \forall g \in \mathcal{G}$, maintaining proportions that reflect the U.S. population. Applying this to our imbalanced data resulted in a balanced dataset, as summarized in Table 2. Our resampling approach improves the representation of all ethnic groups and provides bias reduction in DEA efficiency estimates, as shown in Theorem 2. The conditions of the theorem are satisfied in our study, as the production possibility set $\mathcal{T}$ is compact, being formed by observed data with bounded input and output variables, and the large sample size and continuous nature of our variables support the smoothness of the efficiency frontier.

THEOREM 2. *Let $\mathcal{T} \subset \mathbb{R}^{p+q}$ be the true production possibility set for our DEA model with fixed number of inputs p and outputs q, and assume $\mathcal{T}$ is compact with a smooth frontier. For each group $g \in \mathcal{G}$, let θ_g^* denote the true efficiency score, $\hat{\theta}_{n,g}$ the empirical efficiency score based on the imbalanced data, and $\tilde{\theta}_{n,g}$ the empirical efficiency score based on the*

resampled data. Let $\delta > 0$ be a constant representing the desired precision in estimating the true efficiency scores, and $C > 0$ be a constant capturing the complexity of the DEA estimation problem in (1), which depends on the dimensionality of inputs p and outputs q that form $\mathcal{T}$. Then, for $n \geq \max\{(\frac{C}{\delta})^{\frac{p+q+1}{2}}, \min_{g\in\mathcal{G}} \frac{1}{p_g}\}$,

$$|\mathit{Bias}(\tilde{\theta}_{n,g})| \leq |\mathit{Bias}(\hat{\theta}_{n,g})| \quad \forall g \in \mathcal{G},$$

where $\mathit{Bias}(\theta_{n,g}) = \mathbb{E}[\theta_{n,g}] - \theta^*_g$.

Following resampling, we implemented a debiasing procedure to isolate the effect of ethnicity on our fairness criteria measures while controlling for potential confounding variables. We employed Neyman orthogonal scores with a five-fold cross-fitting strategy, as described by Chernozhukov et al. (2018) and Dai and Li (2023), due to its robustness and ability to handle high-dimensional confounders. This approach involves estimating propensity scores $p_0(W) = \mathbb{P}(G|W)$ and nuisance functions $\mu_0(G, W) = \mathbb{E}[Z|G, W]$, where G denotes ethnicity, W represents confounders, and Z is the respective outcome (X_1, X_2, or Y_1). For each fold k, we calculated residuals $U_k = Z - \hat{\mu}_{0,k}(G, W)$, orthogonalized treatment $V_k = G - \hat{p}_{0,k}(W)$, and Neyman orthogonal scores $\psi_k = V_k \cdot U_k$. The final adjusted score was computed as the average across folds: $\tilde{Z} = \sum_{k=1}^{5} \psi_k/5$. We applied this procedure to our three fairness metrics—waitlist duration, KDPI score, and graft lifespan—with confounding variables specific to each outcome based on literature. For waitlist duration, we include PRA, ABO blood type, gender, age, geographic region, and insurance status (Hall et al., 2011). For KDPI score, we include donor characteristics such as age, height, weight, cause of death, serum creatinine, history of hypertension and diabetes, and HCV status (OPTN, 2021). For graft lifespan, we include recipient and donor age, PRA, gender, HLA match, cold ischemia time, and previous transplant history (Hariharan et al., 2000). The resulting adjusted scores $\tilde{X}_1, \tilde{X}_2$, and $\tilde{Y}_1$ represent debiased estimates of waitlist duration, KDPI score, and graft lifespan, respectively, after accounting for identified confounding factors including ethnicity. For the remainder of this paper, we will drop the tilde notation and refer to these debiased measures as X_1, X_2, and Y_1, respectively.

3.3. *Fairness Criteria Analysis.* To thoroughly evaluate fairness in kidney allocation, we study three criteria: Priority Fairness (waitlist duration), Access Fairness (KDPI scores), and Outcome Fairness (graft lifespan). These detailed analyses provide a foundation for our subsequent Data Envelopment Analysis.

Priority Fairness: Mediation Analysis. Priority fairness aims to identify and quantify disparities in waitlist durations across ethnic groups that cannot be explained by medical factors alone. To achieve this, we use mediation analysis to decompose the total effect of ethnicity into direct and indirect effects (MacKinnon, 2012). Specifically, we study how ethnicity influences waitlist duration (denoted by X_1) while accounting for key mediators such as blood type, panel reactive antibody (PRA) levels, and dialysis status. These mediators were chosen due to their roles in organ compatibility (Shin and Kim, 2011) and the severity of patient conditions (Meier-Kriesche and Kaplan, 2002).

Our mediation model is formulated as a series of regression equations capturing the interactions between ethnicity, mediators, and waitlist time. Consider the following model,

$$X_1 = i_1 + \sum_{g\in\mathcal{G}\setminus\{\text{White}\}} \beta_g \mathbb{I}(G = g) + \tau^\top M + \psi_1 A + \zeta_1 \mathbb{I}(S = \text{Male}) + \epsilon_1,$$

where $\mathcal{G} = \{\text{Asian}, \text{Black}, \text{Hispanic}, \text{White}\}$ is the set of ethnic groups, G is the categorical variable indicating the ethnic group for each observation, M is the vector of mediators composed of M_b (blood type mediators), M_d (dialysis status mediator), and M_{PRA} (PRA level

mediator), A represents Age as a continuous variable, S represents Sex, and τ, β_g, ψ_1, and ζ_1 are the corresponding coefficient vectors. For each mediator, we construct specific regression models. Let $\mathcal{B} = \{A, B, AB, O\}$ be the set of blood types, with O as the reference level. The blood type mediators (M_b) for each $b \in \mathcal{B} \setminus \text{O}$ and the dialysis status mediator (M_d) are modeled by logistic regression expressions of the form,

$$\log\left(\frac{P(M_\xi = 1)}{1 - P(M_\xi = 1)}\right) = i_\xi + \sum_{g \in \mathcal{G} \setminus \{\text{White}\}} \gamma_{\xi,g} \mathbb{I}(G = g) + \epsilon_\xi,$$

where ξ represents either a blood type b or dialysis status d, i_ξ is the intercept term with reference levels blood type O and not being on dialysis respectively, and $\gamma_{\xi,g}$ is the effect of ethnicity g on the mediator M_ξ. The PRA level mediator (M_{PRA}) is modeled by the linear function,

$$M_{\text{PRA}} = i_{\text{PRA}} + \sum_{g \in \mathcal{G} \setminus \{\text{White}\}} \gamma_{\text{PRA},g} \mathbb{I}(G = g) + \epsilon_{\text{PRA}},$$

where i_{PRA} similarly represents the intercept for White patients, and $\gamma_{\text{PRA},g}$ represents the effect of ethnicity g on the PRA level mediator M_{PRA}.

To estimate the direct effects of ethnicity on waitlist duration, adjusting for mediators, we can express $\hat{\beta}_g$ as

$$\hat{\beta}_g = \text{TE}_g - \text{IE}_g,$$

where TE_g is the total effect and IE_g is the indirect effect for ethnic group g, defined as

$$\text{TE}_g = \mathbb{E}[X_1 \mid G = g] - \mathbb{E}[X_1 \mid G = \text{White}],$$

$$\text{IE}_g = \sum_m \tau_m \left(\mathbb{E}[M_m \mid G = g] - \mathbb{E}[M_m \mid G = \text{White}]\right).$$

Here, $\mathbb{E}[X_1|G = g]$ represents the expected waitlist duration for ethnic group g, $\mathbb{E}[M_m|G = g]$ is the expected value of mediator m for group g, and τ_m is the effect of mediator m on waitlist duration. TE_g quantifies the overall difference in expected waitlist duration for ethnicity g relative to the White reference group, while IE_g captures how this difference is mediated through changes in blood type, PRA levels, and dialysis status. By subtracting the indirect effects from the total effect, we isolate $\hat{\beta}_g$, which represents the direct effect of ethnicity on waitlist duration not explained by the medical factors in our model. A significant $\hat{\beta}_g$ indicates that, even after accounting for the medical factors in our model, there remain substantial differences in waitlist duration for ethnic group g compared to the White reference group.

Table 3 shows significant disparities in waitlist durations across ethnic groups. Substantial direct effects for all minority groups, especially Asian patients, indicate unfairness not fully explained by the chosen mediators. The varying proportions of mediated effects suggest non-uniform mechanisms underlying these disparities, with Asian patients showing minimal mediation compared to Black and Hispanic patients.

Access Fairness: Counterfactual Analysis with Random Forest. Access fairness investigates potential inequalities in the quality of kidneys allocated across different ethnic groups, using the KDPI score as a key metric. We employed a counterfactual analysis using a random forest algorithm to isolate the effect of ethnicity on KDPI scores, while accounting for other influential factors. To implement our methodology with five-fold cross-validation, consider the framework,

$$\hat{X}_2 = \frac{1}{N} \sum_{l=1}^{N} T_l(\mathcal{Z}; \Theta_l),$$

Table 3

Summary of direct, indirect, and total effects for each ethnic group. Analysis was conducted using the `mma` *package in* `R` *with 1000 bootstraps to estimate standard errors and confidence intervals. *** indicates p-value* < 0.001.

Patient Ethnicity	Effect Type	Estimate	SE	95% CI
Asian				
	Direct Effect ***	311.298	32.324	[218.658, 399.285]
	Indirect Effect	1.158	8.232	[-20.904, 23.602]
	Total Effect ***	312.456	33.469	[217.151, 400.359]
	Proportion Mediated	0.004		
Black				
	Direct Effect ***	187.400	22.087	[124.373, 252.079]
	Indirect Effect ***	55.236	6.027	[38.694, 71.187]
	Total Effect ***	242.636	22.362	[180.515, 307.940]
	Proportion Mediated	0.228		
Hispanic				
	Direct Effect ***	221.340	20.331	[170.366, 280.672]
	Indirect Effect ***	49.481	4.492	[37.946, 62.550]
	Total Effect ***	270.821	20.692	[216.884, 333.033]
	Proportion Mediated	0.183		

where $\hat{X}_2$ denotes the predicted KDPI score, $N = 500$ is the number of trees, T_l is the l^{th} tree's prediction, and $\mathcal{Z}$ represents the feature set encompassing 14 predictors consisting of both recipient and donor characteristics. The random parameters Θ_l for each tree l govern the tree construction process by determining the subset of features randomly selected for evaluating splits at each node, with cross-validation determining that 12 features at each split yielded optimal model performance. Each tree is constructed by recursively splitting nodes, maximizing information gain at each split. Trees are grown to their maximum depth, with splitting continuing until each terminal node contains no fewer than 5 samples, ensuring a balance between model complexity and predictive power.

To understand the factors influencing KDPI scores, we conducted a variable importance analysis using permutation importance based on out-of-bag estimates (Breiman, 2001). This analysis shows that donor attributes in $\mathcal{Z}$ are the most significant factors influencing KDPI scores. Specifically, donor age emerged as the most important feature, followed by donor creatinine levels and donor ethnicity (particularly for Black donors). Other influential donor factors included non-heart beating donor status, hepatitis C status, and history of hypertension or diabetes. Notably, recipient characteristics in $\mathcal{Z}$, including age, PRA, and ethnicity, played a relatively minor role in determining KDPI scores. Our model demonstrated strong performance with high predictive accuracy, yielding a coefficient of determination $R^2 = 0.9964$. The model's performance is further quantified by the mean absolute error, $\text{MAE} = n^{-1} \sum_{i=1}^{n} |\hat{X}_{2,i} - X_{2,i}| = 0.0101$, and the mean squared error, $\text{MSE} = n^{-1} \sum_{i=1}^{n} (\hat{X}_{2,i} - X_{2,i})^2 = 0.00025$, where $\hat{X}_{2,i}$ is the predicted KDPI score and $X_{2,i}$ is the actual KDPI score for the ith observation. To isolate the effect of patient ethnicity, we created a counterfactual scenario where the ethnicity of all recipients was set to 'White', enabling the estimation of the effect of ethnicity on KDPI scores.

The results of this analysis are summarized in Table 4, which shows minimal disparities in KDPI scores across ethnic groups, with the largest difference observed in the Asian group (a median difference of -0.000397 between actual and counterfactual scenarios). While these differences are statistically significant (Kruskal-Wallis test, p-value < 0.001), their practical significance is negligible. Our variable importance analysis indicates that recipient ethnicity does not directly influence KDPI scores; instead, donor characteristics are the primary factors. This suggests that disparities could arise if certain ethnic groups consistently receive kidneys from donors with specific attributes.

Table 4
Summary of KDPI Score Disparities Across Ethnicities

Ethnicity	Median KDPI (Actual)	Median KDPI (Counterfactual)	Median Difference
White	0.367	-	-
Asian	0.431	0.431	-0.000397
Black	0.391	0.391	-0.000262
Hispanic	0.362	0.363	-0.000215

Outcome Fairness: Competing Risks Analysis. Outcome fairness considers identifying disparities in graft lifespan. To analyze the effect of ethnicity on graft rejection while accounting for other causes of graft failure, we employ the Competing Risks Framework (Fine and Gray, 1999), which offers a comprehensive understanding of the risks associated with graft failure.

In our framework, the subdistribution hazard function for cause k is expressed by

$$h_k(y_1|G) = h_{0k}(y_1)\exp(\delta' G),$$

where $h_k(y_1|G)$ is the subdistribution hazard for cause k at time y_1 since transplantation, $h_{0k}(y_1)$ is the baseline subdistribution hazard for cause k, and G is the vector of indicator variables for Asian, Black, and Hispanic ethnic groups, with White serving as the reference group. The coefficient vector δ contains the effects of these ethnic groups on the risk of graft rejection, which is the primary event of interest, while other causes of graft failure are treated as competing risks. The cause-specific cumulative incidence function for graft rejection is defined as

$$F_1(y_1) = P(Y_1 \leq y_1, D = 1) = \int_0^{y_1} h_1(u)\exp\left(-\int_0^u h_1(s) + h_2(s)ds\right)du,$$

where Y_1 represents the time to graft failure, D is the event type (1 for rejection, 2 for other causes), and h_1 and h_2 are the cause-specific hazard functions for rejection and other causes, respectively. Applying this framework to our data shows significant disparities in graft rejection risk across ethnic groups, with Table 5 detailing substantial variations, where some groups face significantly higher risks compared to others.

Table 5
Subdistribution Hazard Ratios for Graft Rejection

Ethnicity	δ	exp(δ)	SE(δ)	z	p-value	95% CI for exp(δ)
Asian	-0.223	0.800	0.209	-1.069	0.290	[0.532, 1.20]
Black	0.858	2.359	0.101	8.498	0.000	[1.935, 2.88]
Hispanic	0.273	1.313	0.108	2.525	0.012	[1.063, 1.62]

Figure 3 illustrates the cumulative incidence functions for graft rejection and competing risks across ethnic groups. For graft rejection, Black recipients exhibit the highest cumulative incidence, followed by Hispanic, White, and Asian recipients, aligning with the hazard ratios in Table 5. For competing risks, Black recipients demonstrate the highest cumulative incidence of graft failure due to other causes, while White, Hispanic, and Asian recipients show lower and similar trajectories. These patterns highlight the complex nature of disparities in kidney transplant outcomes, with Black recipients facing elevated risks in both categories. In contrast, Asian recipients consistently show lower risk in both categories, while Hispanic recipients show intermediate levels of risk for graft rejection but lower risk for competing causes of graft failure.

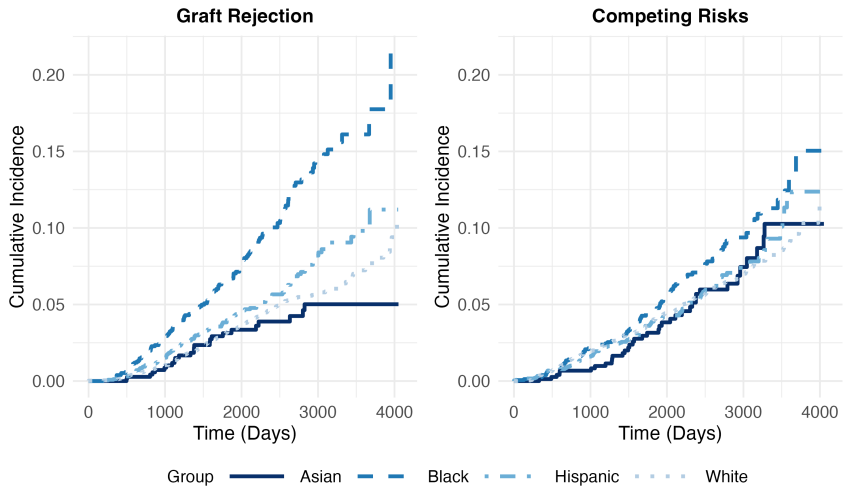


Fig 3. Cumulative Incidence Functions for Graft Rejection and Competing Risks by Ethnicity

4. DEA Estimation and Inference. Given the complex and sometimes conflicting results from our individual fairness criteria analyses, we use Data Envelopment Analysis (DEA) to gain deeper insights into fairness in kidney allocation. DEA enables the simultaneous consideration of multiple inputs and outputs, helping to identify overall disparities while accounting for the multifaceted nature of fairness in the allocation process.

4.1. *DEA Results.* In applying the Data Envelopment Analysis (1), we structured our model with waitlist duration and KDPI score as inputs ($p = 2$, $X_i \in \mathbb{R}^2_+$), and graft lifespan as the output ($q = 1$, $Y_i \in \mathbb{R}_+$) for each patient i. The DEA efficiency scores show notable patterns in kidney allocation efficiency across ethnic groups. With a target coverage level of 95% ($\alpha = 0.05$), we obtained mean efficiency scores and group-conditional prediction intervals for each ethnic group, as presented in Table 6, with the intervals derived from the randomized prediction set defined in (3), ensuring group-conditional coverage guarantees. In this context, group-conditional coverage guarantees ensure that the prediction intervals contain the true efficiency score with 95% probability for each ethnic group, yielding reliable predictions across all groups and revealing differences in the consistency of efficiency scores.

Table 6

DEA efficiency scores by ethnic group, estimated using the `Benchmarking` *package in* `R`*. Group-conditional prediction intervals and coverage were computed using the* `conditionalconformal` *package in Python.*

Group	Mean Efficiency	Interval	Coverage
Asian	0.577	[0.227, 0.926]	0.950
Black	0.561	[0.233, 0.885]	0.952
Hispanic	0.554	[0.258, 0.854]	0.950
White	0.546	[0.398, 0.689]	0.951

The results in Table 6 provide a comprehensive view of efficiency across ethnic groups. Mean efficiency scores range from 0.546 to 0.577, with Asian patients showing the highest average efficiency and White patients the lowest. However, these differences are relatively small, suggesting overall consistency in the kidney allocation system's efficiency across ethnic groups. Notably, the prediction intervals show substantial differences in efficiency score variability. The White patient group exhibits the narrowest interval [0.398, 0.689], indicating

more consistent outcomes. In contrast, Asian, Black, and Hispanic groups show much wider intervals, suggesting greater variability in allocation efficiency within these populations. Figure 4 illustrates these distributions, providing insights into both the central tendencies and the spread and shape of efficiency scores for each group.

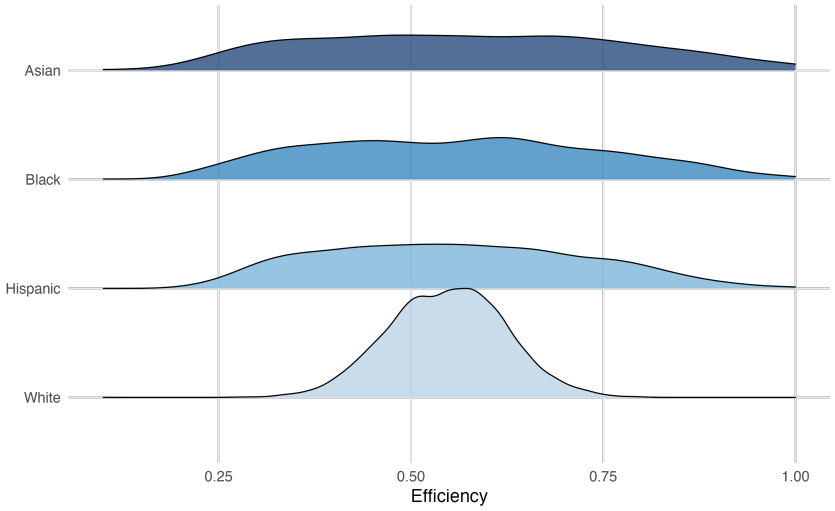


Fig 4. Distribution of Efficiency Scores by Ethnic Group.

4.2. *Analysis of Efficiency Distributions.* To rigorously evaluate the statistical significance of these observed differences, we employed a kernel-based hypothesis testing approach using the Maximum Mean Discrepancy (MMD, Gretton et al., 2012). We conducted pairwise MMD tests for all combinations of ethnic groups. For each pair of groups A and B, we defined the null hypothesis as $H_0 : F_A = F_B$ and the alternative hypothesis as $H_1 : F_A \neq F_B$, where F_A and F_B are the underlying distributions of efficiency scores for groups A and B, respectively. The MMD between these distributions is defined as

$$\text{MMD}^2[F_A, F_B] = \mathbb{E}_{\theta_A,\theta'_A}[K(\theta_A,\theta'_A)] + \mathbb{E}_{\theta_B,\theta'_B}[K(\theta_B,\theta'_B)] - 2\mathbb{E}_{\theta_A,\theta_B}[K(\theta_A,\theta_B)],$$

where $\theta_A, \theta'_A \overset{iid}{\sim} F_A$ and $\theta_B, \theta'_B \overset{iid}{\sim} F_B$. Here, θ represents the DEA efficiency scores corresponding to the inputs X_1, X_2 and output Y_1, as derived in (1). For the kernel function K, we use a Gaussian kernel $K(\theta_A, \theta_B) = \exp(-|\theta_A - \theta_B|^2/2\sigma^2)$, with σ set to the median distance between points in the aggregate sample. We then use this kernel matrix to compute the MMD test statistics and determined their significance via permutation tests with 1000 Monte Carlo iterations. To account for multiple pairwise comparisons, we applied the Bonferonni correction to control the family-wise error rate. With six pairwise tests, the adjusted significance level became $\alpha_{adj} = 0.05/6 \approx 0.008$. We compared the p-values from the permutation tests to this adjusted significance level to determine statistical significance, with the results of the pairwise MMD tests presented in Table 7.

The MMD test results show significant disparities in efficiency score distributions across ethnic group comparisons. The largest disparity was observed between Asian and White patients, with the highest test statistic (0.267) and a highly significant p-value (0.001). Notable differences were also found between Black and White patients (0.207, p-value = 0.001). The comparison between Asian and Hispanic patients (0.011, p-value = 0.001) also showed significant differences. However, the comparisons between Asian and Black patients (0.003,

Table 7
Pairwise MMD Test Results for Efficiency Score Distributions

Group Comparison	MMD Statistic	p-value
Asian vs. Black	0.003	0.011
Asian vs. Hispanic	0.011	0.001
Asian vs. White	0.267	0.001
Black vs. Hispanic	0.003	0.001
Black vs. White	0.207	0.001
Hispanic vs. White	0.156	0.001

p-value = 0.011) and between Black and Hispanic patients (0.003, p-value = 0.001) showed smaller differences, with only the latter being statistically significant after correction. This suggests that while there are significant differences in efficiency distributions between most ethnic groups, the distributions for Asian and Black patients are not significantly different from each other after accounting for multiple comparisons. These findings indicate that the current kidney allocation system does not achieve fair outcomes across ethnic groups, making it necessary to address both average efficiency and within-group variability to improve fairness.

4.3. *Discussion and Implication for Practice.* Our DEA-based fairness analysis has several implications for kidney exchange practices. First, the efficiency score distributions shown in Figure 4 and Table 6 show larger variability for Asian, Black, and Hispanic patients compared to White patients, who exhibit more consistent outcomes. This narrower distribution for White patients could be attributed to better access to healthcare resources or more effective pre-transplant care (Kucirka et al., 2012). These findings suggest the need for tailored interventions to improve outcome consistency for underrepresented groups (Gordon et al., 2014).

Second, we propose that incorporating multiple efficiency measures, as in the DEA model (1), could lead to a fairer kidney allocation system across ethnic groups. While recent developments in kidney paired donation programs, such as integrating deceased donors as nondirected donors and chain initiators (Wang et al., 2021, 2022), have improved matching opportunities, these initiatives primarily focus on enhancing overall efficiency. However, our DEA framework emphasizes the trade-offs between various fairness criteria. For example, reducing waitlist times for Asian patients might negatively impact graft survival, while prioritizing graft survival for White patients could increase their wait times. To address these trade-offs and improve allocation strategies, we recommend using the DEA framework to assess the impact of policy changes on the overall fairness across ethnic groups. This approach ensures that both fairness and efficiency are considered in kidney allocation decisions.

5. Conclusion. We propose a new method for evaluating fairness in kidney allocation using Data Envelopment Analysis (DEA), which can be broadly applied to assess fairness in complex healthcare systems. Our DEA framework considers multiple fairness criteria—priority, access, and outcome—providing a more comprehensive view of disparities compared to traditional single-metric approaches. The analysis shows that White patients tend to have more consistent outcomes, while Asian, Black, and Hispanic groups exhibit wider efficiency distributions. The advantage of handling multiple inputs and outputs without predefined weights in DEA allows for an objective assessment of fairness. Our approach highlights disparities and offers actionable targets for policymakers to improve fairness. The code for reproducing the numerical results is available at https://github.com/amofrad/Kidney-Exchange-DEA.

There are several research directions that could build on this DEA-based framework. For example, it is of interest to develop dynamic DEA models that could capture changes in fairness over time. Beyond healthcare, this approach has potential applications in evaluating fairness in diverse resource allocation contexts where resource scarcity and mutual compatibility constraints are present. For instance, in college admissions, the framework could evaluate fairness in student selection processes, considering factors such as academic performance, extracurricular activities, and institutional diversity goals. In job candidate selection, it could evaluate the fairness of hiring practices by simultaneously considering applicant qualifications, company needs, and diversity initiatives.

SUPPLEMENTARY MATERIAL

Supplement to "A Data Envelopment Analysis Approach for Assessing Fairness in Resource Allocation: Application to Kidney Exchange Programs". Technical proofs for theorems and a simulation study to study the impact of resampling on efficiency scores are included in this file.

REFERENCES

AGARWAL, N., ASHLAGI, I., AZEVEDO, E., FEATHERSTONE, C. R. and KARADUMAN, Ö. (2019). Market failure in kidney exchange. *American Economic Review* **109** 4026–4070.

AXELROD, D. A., SCHNITZLER, M. A., XIAO, H., IRISH, W., TUTTLE-NEWHALL, E., CHANG, S.-H., KASISKE, B. L., ALHAMAD, T. and LENTINE, K. L. (2018). An economic assessment of contemporary kidney transplant practice. *American Journal of Transplantation* **18** 1168–1176.

BREIMAN, L. (2001). Random forests. *Machine Learning* **45** 5–32.

U. S. CENSUS BUREAU (2020). RACE. U.S. Census Bureau.

CHARNES, A., COOPER, W. W. and RHODES, E. (1978). Measuring the efficiency of decision making units. *European journal of operational research* **2** 429–444.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21** C1-C68.

CHILINGERIAN, J. A. (1995). Evaluating physician efficiency in hospitals: A multivariate analysis of best practices. *European Journal of Operational Research* **80** 548–574.

CHORPPATH, A. K. and ALPCAN, T. (2011). Learning user preferences in mechanism design. In *2011 50th IEEE Conference on Decision and Control and European Control Conference* 5349–5355. IEEE.

COHEN, J. B., BLOOM, R. D., REESE, P. P., PORRETT, P. M., FORDE, K. A. and SAWINSKI, D. L. (2016). National outcomes of kidney transplantation from deceased diabetic donors. *Kidney International* **89** 636–647.

CORESH, J., SELVIN, E., STEVENS, L. A., MANZI, J., KUSEK, J. W., EGGERS, P., VAN LENTE, F. and LEVEY, A. S. (2007). Prevalence of chronic kidney disease in the United States. *Journal of the American Medical Association* **298** 2038–2047.

DAI, X. and LI, L. (2023). Orthogonalized kernel debiased machine learning for multimodal data analysis. *Journal of the American Statistical Association* **118** 1796–1810.

FAN, P.-Y., ASHBY, V. B., FULLER, D., BOULWARE, L., KAO, A., NORMAN, S. P., RANDALL, H., YOUNG, C., KALBFLEISCH, J. D. and LEICHTMAN, A. B. (2010). Access and outcomes among minority transplant patients, 1999–2008, with a focus on determinants of kidney graft survival. *American Journal of Transplantation* **10** 1090–1107.

FÄRE, R., GROSSKOPF, S. and LOVELL, C. K. (2013). *The Measurement of Efficiency of Production* **6**. Springer Science & Business Media.

FINE, J. P. and GRAY, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94** 496–509.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2023). Chronic kidney disease in the United States, 2023. *Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention*.

GENTRY, S. E., MONTGOMERY, R. A. and SEGEV, D. L. (2011). Kidney paired donation: fundamentals, limitations, and expansions. *American Journal of Kidney Diseases* **57** 144–151.

GENTRY, S. E., SEGEV, D. L., SIMMERLING, M. and MONTGOMERY, R. A. (2007). Expanding kidney paired donation through participation by compatible pairs. *American Journal of Transplantation* **7** 2361–2370.

GIBBS, I., CHERIAN, J. J. and CANDÈS, E. J. (2023). Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*.

GILL, J. S., TONELLI, M., JOHNSON, N. and PEREIRA, B. J. (2004). Why do preemptive kidney transplant recipients have an allograft survival advantage? *Transplantation* **78** 873–879.

GORDON, E. J., MULLEE, J. O., RAMIREZ, D. I., MACLEAN, J., OLIVERO, M., FEINGLASS, J., CARNEY, P., O'CONNOR, K. and CAICEDO, J. C. (2014). Hispanic/Latino concerns about living kidney donation: a focus group study. *Progress in Transplantation* **24** 152–162.

GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research* **13** 723–773.

HALL, Y. N., CHOI, A. I., XU, P., O'HARE, A. M. and CHERTOW, G. M. (2011). Racial ethnic differences in rates and determinants of deceased donor kidney transplantation. *Journal of the American Society of Nephrology* **22** 743–751.

HARDT, M., PRICE, E. and SREBRO, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* **29**.

HARIHARAN, S., ISRANI, A. K. and DANOVITCH, G. (2021). Long-term survival after kidney transplantation. *New England Journal of Medicine* **385** 729–743.

HARIHARAN, S., JOHNSON, C. P., BRESNAHAN, B. A., TARANTO, S. E., MCINTOSH, M. J. and STABLEIN, D. (2000). Improved graft survival after renal transplantation in the United States, 1988 to 1996. *New England Journal of Medicine* **342** 605–612.

HE, H. and GARCIA, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21** 1263–1284.

IMAI, K., KEELE, L. and YAMAMOTO, T. (2010). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science* **25** 51 – 71.

KUCIRKA, L. M., GRAMS, M. E., BALHARA, K. S., JAAR, B. G. and SEGEV, D. L. (2012). Disparities in Provision of Transplant Information Affect Access to Kidney Transplantation. *American Journal of Transplantation* **12** 351-357. https://doi.org/10.1111/j.1600-6143.2011.03865.x

KUSNER, M. J., LOFTUS, J., RUSSELL, C. and SILVA, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems* **30**.

LENTINE, K., SMITH, J., HART, A., MILLER, J., SKEANS, M., LARKIN, L., ROBINSON, A., GAUNTT, K., ISRANI, A., HIROSE, R. et al. (2022). OPTN/SRTR 2020 annual data report: kidney. *American Journal of Transplantation* **22** 21–136.

MACKINNON, D. (2012). *Introduction to Statistical Mediation Analysis*. Routledge.

MALEK, S. K., KEYS, B. J., KUMAR, S., MILFORD, E. and TULLIUS, S. G. (2011). Racial and ethnic disparities in kidney transplantation. *Transplant International* **24** 419–424.

MASSIE, A. B., GENTRY, S. E., MONTGOMERY, R. A., BINGAMAN, A. A. and SEGEV, D. L. (2013). Center-Level Utilization of Kidney Paired Donation. *American Journal of Transplantation* **13** 1317-1322.

MEIER-KRIESCHE, H.-U. and KAPLAN, B. (2002). Waiting time on dialysis as the strongest modifiable risk factor for renal transplant outcomes: A paired donor kidney analysis: 1. *Transplantation* **74** 1377–1381.

MORENO-CALDERÓN, A., TONG, T. S. and THOKALA, P. (2020). Multi-criteria decision analysis software in healthcare priority setting: a systematic review. *Pharmacoeconomics* **38** 269–283.

OGRYCZAK, W., LUSS, H., PIÓRO, M., NACE, D. and TOMASZEWSKI, A. (2014). Fair optimization and networks: A survey. *Journal of Applied Mathematics* **2014** 612018.

OPTN (2021). A Guide to Calculating and Interpreting the Kidney Donor Profile Index (KDPI).

OZCAN, Y. A. et al. (2008). *Health Care Benchmarking and Performance Evaluation*. Springer.

PARK, W. Y., KIM, J. H., KO, E. J., MIN, J.-W., BAN, T. H., YOON, H.-E., KIM, Y. S., JIN, K., YANG, C. W., HAN, S. et al. (2020). Impact of kidney donor profile index scores on post-transplant clinical outcomes between elderly and young recipients, a multicenter cohort study. *Scientific Reports* **10** 7009.

PERSAD, G., WERTHEIMER, A. and EMANUEL, E. J. (2009). Principles for allocation of scarce medical interventions. *The Lancet* **373** 423–431.

PINTO-RAMIREZ, J., GARCIA-LOPEZ, A., SALCEDO-HERRERA, S., PATINO-JARAMILLO, N., GARCIA-LOPEZ, J., BARBOSA-SALINAS, J., RIVEROS-ENRIQUEZ, S., HERNANDEZ-HERRERA, G. and GIRON-LUQUE, F. (2022). Risk factors for graft loss and death among kidney transplant recipients: A competing risk analysis. *PLoS One* **17** e0269990.

RAPAPORT, F. T. (1986). The case for a living emotionally related international kidney donor exchange registry. In *Transplantation Proceedings* **18** 5–9.

REES, M. A., DUNN, T. B., KUHR, C. S., MARSH, C. L., ROGERS, J., REES, S. E., CICERO, A., REECE, L. J., ROTH, A. E., EKWENNA, O. et al. (2017). Kidney exchange to overcome financial barriers to kidney transplantation. *American Journal of Transplantation* **17** 782–790.

ROTH, A. E., SÖNMEZ, T. and ÜNVER, M. U. (2004). Kidney exchange. *The Quarterly Journal of Economics* **119** 457–488.

SHERO, J. A., AL OTAIBA, S., SCHATSCHNEIDER, C. and HART, S. A. (2022). Data envelopment analysis (DEA) in the educational sciences. *The Journal of Experimental Education* **90** 1021–1040.

SHIN, M. and KIM, S.-J. (2011). ABO incompatible kidney transplantation—current status and uncertainties. *Journal of Transplantation* **2011** 970421.

UNITED STATES RENAL DATA SYSTEM (2023). 2023 USRDS annual data report: epidemiology of kidney disease in the United States.

THOKALA, P., DEVLIN, N., MARSH, K., BALTUSSEN, R., BOYSEN, M., KALO, Z., LONGRENN, T., MUSSEN, F., PEACOCK, S., WATKINS, J. et al. (2016). Multiple criteria decision analysis for health care decision making—an introduction: report 1 of the ISPOR MCDA Emerging Good Practices Task Force. *Value in Health* **19** 1–13.

TONELLI, M., WIEBE, N., KNOLL, G., BELLO, A., BROWNE, S., JADHAV, D., KLARENBACH, S. and GILL, J. (2011). Systematic review: kidney transplantation compared with dialysis in clinically relevant outcomes. *American Journal of Transplantation* **11** 2093–2109.

WANG, W., REES, M. A., LEICHTMAN, A. B., SONG, P. X.-K., BRAY, M., ASHBY, V. B., SHEARON, T., WHITEMAN, A. and KALBFLEISCH, J. D. (2021). Deceased donors as nondirected donors in kidney paired donation. *American Journal of Transplantation* **21** 103–113.

WANG, W., LEICHTMAN, A. B., REES, M. A., SONG, P. X.-K., ASHBY, V. B., SHEARON, T. and KALBFLEISCH, J. D. (2022). Kidney paired donation chains initiated by deceased donors. *Kidney International Reports* **7** 1278–1288.

WOLFE, R. A., ASHBY, V. B., MILFORD, E. L., OJO, A. O., ETTENGER, R. E., AGODOA, L. Y., HELD, P. J. and PORT, F. K. (1999). Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *New England Journal of Medicine* **341** 1725–1730.

WOLFE, R., MCCULLOUGH, K., SCHAUBEL, D., KALBFLEISCH, J., MURRAY, S., STEGALL, M. and LEICHTMAN, A. (2008). Calculating Life Years from Transplant (LYFT): Methods for Kidney and Kidney-Pancreas Candidates. *American Journal of Transplantation* **8**.