

UCLA

UCLA Previously Published Works

Title

A data envelopment analysis approach for assessing fairness in resource allocation: Application to kidney exchange programs

Permalink

<https://escholarship.org/uc/item/3t66v9nt>

Journal

The Annals of Applied Statistics, 20(1)

ISSN

1932-6157

Authors

Kaazempur-Mofrad, Ali
Dai, Xiaowu

Publication Date

2026-03-01

DOI

10.1214/25-aoas2128

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at

<https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

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

BY ALI KAAZEMPUR-MOFRAD^a  AND XIAOWU DAI^b 

¹*Department of Statistics and Data Science, University of California, Los Angeles, ^aamofrad@ucla.edu, ^bdaix@ucla.edu*

Kidney exchange programs have substantially increased transplantation rates but also raise critical concerns about fairness in organ allocation. We propose a novel framework leveraging Data Envelopment Analysis (DEA) to evaluate multiple dimensions of fairness—Priority, Access, and Outcome—within a unified model. This approach captures complexities often missed in single-metric analyses. Using data from the United Network for Organ Sharing, we separately quantify fairness across these dimensions: Priority Fairness through waitlist durations, Access Fairness via the Living Kidney Donor Profile Index (LKDPI) scores, and Outcome Fairness based on graft lifespan. We then apply our conditional DEA model with covariate adjustment to demonstrate significant disparities in kidney allocation efficiency across ethnic groups. To quantify uncertainty, we employ conformal prediction within a novel reference frontier mapping (RFM) 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 with resource scarcity and mutual compatibility constraints.

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 \(2024\)](#)). 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

Received June 2025; revised November 2025.

Key words and phrases. Conformal prediction, data envelopment analysis, fairness, kidney exchange, resource allocation.

have been 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, thereby 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 KPD allocation using conditional 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)) extended to account for exogenous covariates via conditional frontiers (Bădin, Daraio and Simar (2012)). Prior to applying this conditional 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 LKDPI 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 conditional DEA approach addresses these complexities by 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 score that captures how well KPD allocations serve each patient or subgroup, considering multiple fairness criteria simultaneously.

To quantify uncertainty around these efficiency scores, we integrate conformal prediction techniques proposed by Gibbs, Cherian and Candès (2025) within the conditional DEA framework, using our proposed reference frontier mapping (RFM) procedure to decouple frontier estimation from evaluation. This procedure generates group-conditional prediction intervals for efficiency scores with finite-sample coverage guarantees, ensuring equitable uncertainty quantification across ethnic groups. 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 provides an interpretable, unified measure of overall system fairness for each patient or subgroup, and 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)), multicriteria 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 framework provides a unified efficiency measure that integrates multiple fairness criteria without requiring fixed weights or distributional assumptions. This approach can be applied more broadly to other resource allocation systems facing 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 methods, including the conditional DEA model and RFM-based uncertainty quantification. Section 3 details the data preparation process and the individual fairness analyses. Section 4 presents the results of applying our conditional 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 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, with shorter durations indicating 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 a mediation analysis in Section 3.3 to examine the direct effect of ethnicity on X_{1i} while accounting for mediators, such as recipient blood type and dialysis status.

Access fairness. Access Fairness considers the distribution of kidney quality, as indicated by the Living Kidney Donor Profile Index (LKDPI). The LKDPI score is a numerical measure that combines both donor and recipient characteristics to summarize the relative risk associated with a living donor transplant (Massie et al. (2016)). The LKDPI can theoretically take values below 0 (indicating lower risk than any deceased donor kidney) or above 100 (indicating higher risk than any deceased donor kidney). In our model, let X_{2i} denote the LKDPI 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 kidney allocation fairness 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 LKDPI score, while maximizing others, such as 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 $X_i = (x_{1i}, \dots, x_{pi}) \in \mathbb{R}_+^p$ denotes the p inputs and $Y_i = (y_{1i}, \dots, y_{qi}) \in \mathbb{R}_+^q$ denotes the q outputs for patient $i = 1, \dots, n$. In our kidney allocation analysis, we consider two key pretransplant factors as inputs, waitlist duration, and LKDPI 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 pretransplant factors into prolonged graft survival.

Central to the DEA framework is the production possibility set

$$\mathcal{T} = \left\{ (X, Y) \in \mathbb{R}_+^p \times \mathbb{R}_+^q \mid X \geq \sum_{j=1}^n \lambda_j X_j, Y \leq \sum_{j=1}^n \lambda_j Y_j, \sum_{j=1}^n \lambda_j = 1, \lambda_j \geq 0 \right\}$$

which represents the convex hull of all observed transplants. To account for and control structural differences in patient background characteristics, we extend the classical DEA framework using the conditional efficiency formulation of Bădin, Daraio and Simar (2012). Let $Z_i \in \mathbb{R}^r$ be a vector of exogenous covariates for i , such as education level, UNOS region, and citizenship status. To incorporate these covariates into the DEA framework, we define the covariate-conditional production set for each patient i as

$$(1) \quad \mathcal{T}_{Z_i} = \left\{ (X, Y) \in \mathbb{R}_+^p \times \mathbb{R}_+^q \mid X \geq \sum_{j \in \mathcal{I}_i} \lambda_j X_j, Y \leq \sum_{j \in \mathcal{I}_i} \lambda_j Y_j, \sum_{j \in \mathcal{I}_i} \lambda_j = 1, \lambda_j \geq 0 \right\},$$

where $K_h(Z_j, Z_i)$ is a smoothing kernel that measures similarity between the covariates of patients j and i and $\mathcal{I}_i = \{j \in \{1, \dots, n\} : K_h(Z_j, Z_i) > 0\}$. The kernel is constructed as

$$(2) \quad K_h(Z_j, Z_i) = \prod_{k \in \mathcal{C}} [\mathbb{I}(Z_{jk} = Z_{ik}) + (1 - \mathbb{I}(Z_{jk} = Z_{ik}))e^{-1/h^2}] \cdot \exp\left(-\frac{\|Z_j^{\text{cont}} - Z_i^{\text{cont}}\|^2}{2h^2}\right),$$

where $\mathcal{C}$ is the index set of categorical variables and Z^{cont} denotes the continuous subset of Z . The bandwidth parameter h controls the degree of localization in the conditional efficiency estimation and is selected as $h = n^{-1/(r+4)}$, which represents an optimal bandwidth choice that achieves the best convergence rates for conditional DEA estimators (Bădin, Daraio and Simar (2012)). This construction softly localizes the frontier to account for environmental constraints, enabling a fairer comparison across patients with differing backgrounds.

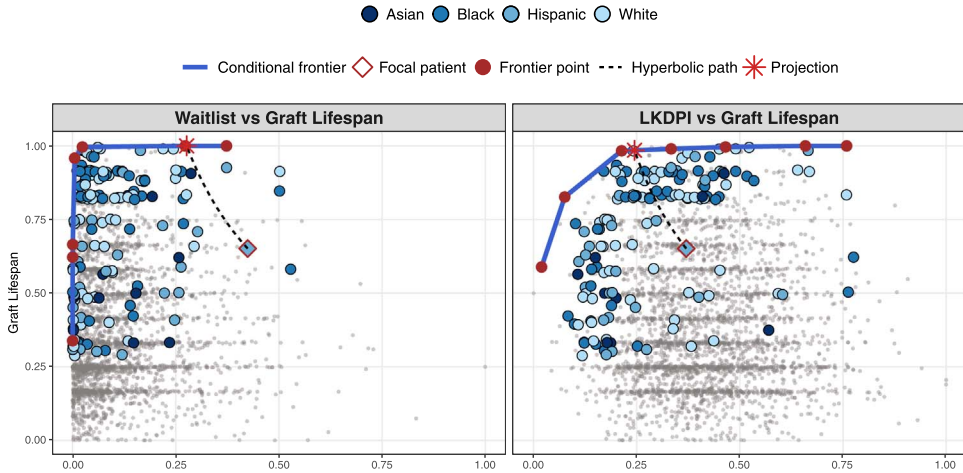


FIG. 1. Illustration of the conditional DEA framework for a single focal patient. Each panel plots one input (waitlist duration or LKDPI) against the output (graft lifespan). Small points represent all observations in the sample, and the focal patient is marked with a diamond. Filled circular points indicate the focal patient's local reference set, consisting of patients with similar covariate profiles Z . The solid curve represents the covariate-conditional efficiency frontier estimated from this local reference set; points marked along this curve correspond to observations lying on the frontier. The dashed curve shows the hyperbolic trajectory along which the focal patient's input–output pair is scaled by θ until it intersects the frontier at the conditional efficiency target $\theta_i(Z_i) = 1$.

The efficiency of a KPD transplant is quantified by the conditional hyperbolic efficiency score

$$(3) \quad \theta_i(Z_i) = \min\{\theta > 0 \mid (\theta X_i, \theta^{-1} Y_i) \in \mathcal{T}_{Z_i}\},$$

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

The efficiency frontier is a key component of our DEA framework, representing the set of input–output combinations $(X, Y) \in \mathcal{T}_{Z_i}$ for which no other feasible point satisfies $X' \leq X$ and $Y' \geq Y$ with at least one strict inequality. For any patient not on this frontier ($\theta_i(Z_i) < 1$), the value of $\theta_i(Z_i)$ quantifies the proportional decrease in inputs and increase in output required for the observed transplant to reach the frontier.

Figure 1 illustrates how the conditional frontier and hyperbolic efficiency score are constructed for a single focal patient. To make the geometry visually interpretable, the inputs and output in these panels have been rescaled to the unit interval via min–max normalization, though this scaling does not affect the DEA estimates. Each panel displays all transplants as background points, highlighting one representative patient and the subset of reference units identified using covariate similarity. The solid curve depicts the patient-specific covariate-conditional frontier estimated from this local reference set. The dashed curve represents the hyperbolic path the patient's input–output combination would follow if proportionally scaled by θ until it reaches the conditional frontier at $\theta = 1$, marked by a star. Although every patient has an individualized conditional frontier, we present a single example to maintain clarity and to demonstrate the geometric interpretation of $\theta_i(Z_i)$.

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

$$(4) \quad \min_{\theta, \lambda_j} \theta \quad \text{s.t.} \quad \theta X_i \geq \sum_{j \in \mathcal{I}_i} \lambda_j X_j, \quad \frac{1}{\theta} Y_i \leq \sum_{j \in \mathcal{I}_i} \lambda_j Y_j, \\ \sum_{j \in \mathcal{I}_i} \lambda_j = 1, \quad \lambda_j \geq 0,$$

where $\mathcal{I}_i = \{j \in \{1, \dots, n\} : K_h(Z_j, Z_i) > 0\}$ denotes the locally weighted set of comparison units for patient i . These constraints ensure that each patient is only compared to peers with similar covariate profiles.

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 to improve outcomes.

2.3. Group-conditional uncertainty quantification under reference frontier mapping. Uncertainty quantification provides confidence measures for the conditional efficiency scores computed in (4), accounting for variability in input-output measurements and sampling. This is essential for evaluating fairness across ethnic groups by determining the statistical significance of efficiency score differences. To this end, we employ conformal prediction methods, which provide finite-sample coverage guarantees without relying on distributional assumptions and are well-suited for complex, nonlinear relationships.

A central challenge in applying conformal prediction to DEA is that efficiency scores $\{\theta_i\}_{i=1}^n$ are computed relative to a frontier estimated from the full dataset. This induces statistical dependence between the construction of the frontier and the observations, violating the i.i.d. assumptions required for conformal inference. To resolve this, we propose a data-splitting framework, *reference frontier mapping* (RFM), which enables valid and equitable uncertainty quantification for conditional DEA efficiency scores. RFM separates the roles of frontier construction and score evaluation by partitioning the data into two disjoint subsets: a reference set $\mathcal{D}_{\text{ref}}$, used to estimate a covariate-conditional production frontier, and an evaluation set $\mathcal{D}_{\text{eval}}$ on which prediction intervals are computed.

The reference set is chosen to be informative about the boundary of the production possibility set, while the remaining observations form $\mathcal{D}_{\text{eval}}$. The targeted sampling and frontier-relevance scoring procedure used to construct $\mathcal{D}_{\text{ref}}$ is detailed in Supplementary Material Section S1.1 (Kaaempur-Mofrad and Dai (2026)). For each $i \in \mathcal{D}_{\text{eval}}$, we estimate a covariate-conditional production set $\mathcal{T}_{Z_i}$ using only reference units weighted by the kernel $K_h(Z_j, Z_i)$ in (1). This yields a locally adaptive frontier reflecting the structural and demographic context of KPD transplant patient i . The corresponding conditional efficiency score is then computed by solving (4), which can be expressed as the functional mapping $\theta_i = \varphi_{\mathcal{T}_{Z_i}}(X_i, Y_i)$, where $\varphi_{\mathcal{T}_{Z_i}} : \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$ denotes the DEA operator defined by the covariate-conditional frontier. Once $\mathcal{D}_{\text{ref}}$ is fixed, the evaluation triplets (X_i, Y_i, θ_i) are conditionally independent of frontier construction and may be treated as conditionally i.i.d. for conformal inference.

To obtain group-conditional uncertainty guarantees, we apply the method of Gibbs, Cherian and Candès (2025), which ensures that prediction intervals achieve the target coverage level within each ethnic group. In brief, we estimate group-specific conditional means of θ_i , compute conformity scores based on deviations from these estimates, and then invert a conformal calibration step to obtain prediction intervals satisfying

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

This is a group-specific coverage guarantee: conditional on belonging to group g , the interval contains the efficiency score with probability $1 - \alpha$. The detailed optimization formulation, function-class specification, and randomized conformal prediction-set construction, together with the proof of the corresponding coverage theorem stated below, are provided in Sections S1.2 and S2.1 of the Supplementary Material (Kaazempur-Mofrad and Dai (2026)).

THEOREM 1. *Let $\mathcal{G}$ denote a set of groups, and let $(X_i, Y_i, Z_i)_{i=1}^n$ be i.i.d. samples from a joint distribution $\mathbb{P}_{X,Y,Z}$, where $X_i \in \mathbb{R}^p$, $Y_i \in \mathbb{R}^q$, and $Z_i \in \mathbb{R}^r$. Let $\mathcal{D}_{\text{ref}}$ be a fixed reference set used to compute conditional efficiency scores via the mapping $\theta_i = \varphi_{\mathcal{T}_{Z_i}}(X_i, Y_i)$, where $\varphi_{\mathcal{T}_{Z_i}} : \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$ denotes the conditional DEA operator defined by $\mathcal{D}_{\text{ref}}$ and kernel smoothing. Then for any new sample $(X_{n+1}, Y_{n+1}, Z_{n+1}) \sim \mathbb{P}_{X,Y,Z}$, the randomized conformal prediction set $\hat{C}_{\text{rand.}}(X_{n+1}, Y_{n+1})$ satisfies*

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

where $\theta_{n+1} = \varphi_{\mathcal{T}_{Z_{n+1}}}(X_{n+1}, Y_{n+1})$ and α is the target miscoverage rate.

The efficiency score θ_{n+1} in Theorem 1 represents the patient's conditional DEA efficiency relative to the fixed reference frontier $\mathcal{D}_{\text{ref}}$. This formulation does not require (X_{n+1}, Y_{n+1}) to correspond to future arrivals in a dynamic KPD system; instead, it provides uncertainty quantification for any new or held-out observation relative to a clearly defined historical benchmark. Each patient's efficiency score reflects performance relative to what was empirically achievable during their treatment period using completed transplants with similar covariate profiles. Although our analysis focuses on completed transplants within fixed time windows, the RFM framework can be extended to dynamic settings by periodically updating $\mathcal{D}_{\text{ref}}$ using rolling historical data. The resulting prediction set $\hat{C}(X_{n+1}, Y_{n+1})$, therefore, provides rigorous, group-conditional uncertainty quantification for any patient-donor pair evaluated against the RFM frontier, enabling valid assessment of efficiency disparities across ethnic groups in both retrospective and prospective settings.

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 and serves as one of the most comprehensive national repositories for transplant data in the United States. Focusing specifically on kidney transplants conducted through KPD programs, we treat the UNOS registry as a representative sample of historical KPD transplants in the United States for our analysis. To provide a more accurate representation of the current landscape, we analyzed kidney transplant records from 2010 to 2019, focusing exclusively on the 6365 patients who successfully received KPD transplants during this period, ensuring complete observation of all time-to-event outcomes. For each transplant, the UNOS data provides comprehensive information on a wide range of clinical and demographic variables, including recipient and donor age, ethnicity, blood type, panel reactive antibody (PRA) levels, and waitlist duration (measured as the time from initial waitlist registration to KPD transplant date). This enables analysis of both pretransplant factors (e.g., waitlist duration) and posttransplant outcomes (e.g., graft survival time). Table 1 presents summary statistics for key variables across different ethnic groups.

These summary statistics show differences across ethnic groups in key transplantation metrics. Black patients, on average, experience the longest waitlist durations compared to other ethnic groups, with a mean of 691.5 days, followed closely by Asian (685.4 days) and Hispanic patients (646.1 days). White patients have the shortest mean waitlist duration of 514.3

TABLE 1
Summary statistics of key variables by ethnic group

Variable	Asian	Black	Hispanic	White
Waitlist Duration (days)				
Mean (SD)	685.4 (611.3)	691.5 (691.4)	646.1 (698.5)	514.3 (528.8)
Median [IQR]	465 [228, 1064]	460 [203, 944]	398.5 [140, 892]	348 [166, 678]
LKDPI				
Mean (SD)	−8.30 (21.2)	−2.65 (20.8)	−7.67 (19.4)	−4.56 (20.6)
Median [IQR]	−9.83 [−23.43, 4.11]	−4.37 [−17.69, 10.01]	−8.97 [−21.63, 5.09]	−6.45 [−19.05, 8.55]
Graft Survival (days)				
Mean (SD)	1684.2 (881.2)	1744.0 (990.3)	1610.5 (895.5)	1782.7 (978.6)
Median [IQR]	1459.0 [1090, 2197]	1462.5 [1045.75, 2430.5]	1418.5 [935.75, 2192.5]	1486 [1071, 2513]

days, substantially lower than other groups. Black patients receive kidneys with higher mean LKDPI scores (−2.65) in comparison to other groups. In terms of graft survival, White patients have the longest mean survival time (1782.7 days or approximately 4.9 years), followed closely by Black patients (1744.0 days or 4.8 years). Asian and Hispanic patients have shorter mean graft survival times (1684.2 and 1610.5 days or about 4.6 and 4.4 years, respectively). These observations suggest potential disparities in the kidney allocation process, with White patients appearing to have shorter waitlist durations and better graft survival, while other patients face longer wait times and shorter graft survival but have comparable LKDPI scores. 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 relative measures.* Our data preparation proceeds in two stages: first, addressing group representation imbalance through stratified resampling, and second, defining fairness measures in relative terms rather than absolute terms.

To illustrate the magnitude of group imbalance in our data, we compare the ethnic composition of patients who entered the KPD system in 2019 with the national ESRD prevalence for that year. According to the United States Renal Data System (USRDS) Annual Report (United States Renal Data System (2024)), the 2019 ESRD prevalence across major ethnic groups in the U.S. was: Asian (5.4%), Black (30.0%), Hispanic (19.1%), and White (45.5%). As shown in Figure 2, the empirical distribution of our 2019 KPD data deviates substantially from these population benchmarks. Similar imbalances are observed in other years as well,

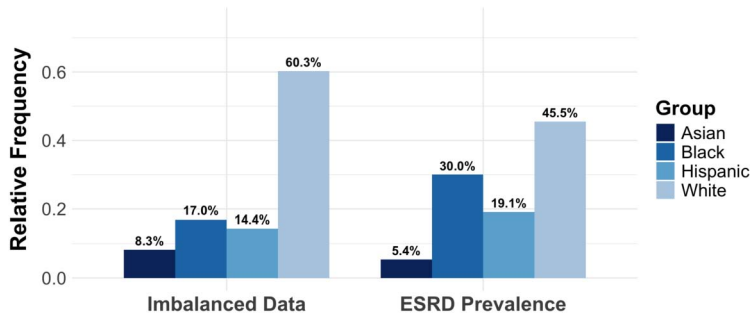


FIG. 2. Comparison of the Distribution of Ethnic Groups in the 2019 Subset of the Imbalanced Data and the 2019 ESRD Prevalence in the U.S.

motivating the need for debiasing. Group representation disparities can compromise the validity of efficiency comparisons by introducing systematic bias into the DEA estimates. It has been shown that class imbalance can significantly affect the performance of learning algorithms, leading to biased predictions favoring the majority class (He and Garcia (2009)). In our setting, this results in biased empirical efficiency estimates $\hat{\theta}_g$ for group g , relative to the population-level quantity of interest

$$\theta_g^* := \mathbb{E}_{(X,Y,Z) \sim P_g}[\theta(X, Y|Z)],$$

where $\theta(X, Y|Z)$ denotes the conditional DEA efficiency score of a patient with inputs X , outputs Y , and covariates Z , evaluated relative to the covariate-specific production possibility set $\mathcal{T}_Z \subset \mathbb{R}^{p+q}$. The distribution P_g denotes the joint distribution of (X, Y, Z) , given group membership $\mathcal{G} = g$, and is supported on the union of the covariate-conditional production sets. That is, $P_g(A) := \mathbb{P}((X, Y, Z) \in A | \mathcal{G} = g)$, for all measurable $A \subseteq \mathbb{R}_+^p \times \mathbb{R}_+^q \times \mathbb{R}^r$.

To assess disparities in efficiency scores across ethnic groups in a manner that accurately reflects the demographic composition of patients affected by ESRD, we treat the U.S. ESRD patient population as the reference population to ensure fairness. Consequently, our resampling strategy aims to align the group proportions in the data with year-specific ESRD prevalence statistics from the USRDS by performing stratified resampling separately for each calendar year. For patients who entered the KPD system in year t , we resample according to the population distribution of ESRD patients in that year. For each group $g \in \mathcal{G}$ and year t , let $p_{g,t}$ be the ESRD prevalence proportion and $n_{g,t}$ the number of observed patients in the original data. To avoid oversampling, we first compute the maximum feasible resampling size for year t as $N_t = \min_{g \in \mathcal{G}} \lfloor n_{g,t} / p_{g,t} \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function. Then within each group g , we sample $n'_{g,t} = \lfloor p_{g,t} \cdot N_t \rfloor$ observations without replacement, where $\lfloor \cdot \rfloor$ denotes rounding to the nearest integer. This procedure yields a resampled dataset whose ethnic composition more closely reflects the year-specific ESRD population structure. Let $n = \sum_t \sum_{g \in \mathcal{G}} n_{g,t}$ denote the total number of observations in the original dataset, and let $n' = \sum_t \sum_{g \in \mathcal{G}} n'_{g,t} \leq n$ denote the total sample size after resampling. The resulting post-resampling group distribution is summarized in Table 2.

Let $\hat{\theta}_g$ and $\hat{\theta}'_g$ denote the empirical mean efficiency scores for group g computed from the original and resampled datasets, respectively, obtained by the DEA estimation problem in (4) and averaging the resulting scores across individuals in group g . Theorem 2 below formalizes the bias reduction achieved through the resampling procedure.

THEOREM 2. *Let $\mathcal{T}_Z \subset \mathbb{R}^{p+q}$ denote the covariate-conditional production set defined by the DEA kernel-based estimator for each patient with covariate profile $Z \in \mathbb{R}^r$. Assume that $\mathcal{T}_Z$ is compact and possesses a C^2 -smooth frontier for all Z , and that the smoothing kernel in (2) is Lipschitz continuous in Z . For each group $g \in \mathcal{G}$, let θ_g^* denote the true conditional efficiency, $\hat{\theta}_g$ the empirical mean efficiency score from the imbalanced dataset of size n , and*

TABLE 2
Distribution of patients by ethnicity—Post-resampling

Ethnicity	Number of patients	Proportion of total (%)
Asian	161	4.97%
Black	1004	30.99%
Hispanic	582	17.96%
White	1493	46.08%
Total	3240	

$\hat{\theta}_g'$ the corresponding estimate from the resampled dataset of size n' , where $n' \leq n$. Let $\delta > 0$ be a constant representing the desired precision in estimating the true efficiency scores, and define $K := \sup_{g \in \mathcal{G}} \{ \limsup_{n \rightarrow \infty} n^{4/((r+4)(p+q))} \mathbb{E}[|\hat{\theta}_g - \theta_g^*|] \}$ as a constant characterizing the convergence rate of the estimators. Then for $n' \geq \max\{(\frac{K}{\delta})^{\frac{(r+4)(p+q)}{4}}, \min_{g \in \mathcal{G}} \frac{1}{p_g}\}$, we have

$$|\text{Bias}(\hat{\theta}_g')| \leq |\text{Bias}(\hat{\theta}_g)| \quad \forall g \in \mathcal{G},$$

where $\text{Bias}(\hat{\theta}_g) := \mathbb{E}[\hat{\theta}_g] - \theta_g^*$.

The constant K in Theorem 2 depends on the dimensionality (p, q, r) of the input-output-covariate space and reflects both the curvature of the conditional frontier $\partial \mathcal{T}_Z$ and the smoothness of the covariate kernel used to define the local neighborhood around each observation. Under the assumptions that the conditional production set $\mathcal{T}_Z$ is compact with a twice continuously differentiable boundary and that the kernel function $K_h(Z_j, Z)$ is Lipschitz continuous, conditional DEA estimators are known to be consistent with convergence rates of order $O(n^{-4/((r+4)(p+q))})$, which accounts for the effective sample size nh^r (Bădin, Daraio and Simar (2012)). The true group-level efficiency score θ_g^* is defined with respect to the U.S. ESRD patient population; thus, the empirical data used in our analysis must reflect the demographic composition of this population to provide unbiased and representative estimates.

Following the resampling step, we define the fairness criteria used in our analysis. For each KPD transplant i performed in year t , we denote $X_{1i}^{(t)}$ as the observed waitlist duration, $X_{2i}^{(t)}$ as the observed LKDPI score, and $Y_{1i}^{(t)}$ as the observed graft lifespan.

Our analysis employs two distinct but complementary approaches that require different data transformations. For the individual fairness analyses described in Section 3.3, we center each criterion around its year-specific sample mean to evaluate fairness relative to expected outcomes and account for temporal changes in the transplant landscape. Specifically, we define the centered fairness measures as

$$\tilde{X}_{1i}^{(t)} := X_{1i}^{(t)} - \bar{X}_1^{(t)}, \quad \tilde{X}_{2i}^{(t)} := X_{2i}^{(t)} - \bar{X}_2^{(t)}, \quad \tilde{Y}_{1i}^{(t)} := Y_{1i}^{(t)} - \bar{Y}_1^{(t)},$$

where $\bar{X}_1^{(t)}$, $\bar{X}_2^{(t)}$, and $\bar{Y}_1^{(t)}$ denote the sample means of waitlist duration, LKDPI, and graft lifespan, respectively, computed over all patients transplanted in year t within the resampled dataset. This year-specific centering procedure ensures that fairness comparisons account for temporal changes in transplant practices, allocation policies, and patient populations. The centered fairness measures $\tilde{X}_{1i}^{(t)}$, $\tilde{X}_{2i}^{(t)}$, and $\tilde{Y}_{1i}^{(t)}$ quantify whether each KPD transplant achieved better or worse outcomes relative to the yearly sample mean, thereby enabling a relative assessment of priority fairness, access fairness, and outcome fairness in the individual analyses presented in Section 3.3. Throughout the analyses in Section 3.3, we use X_1 , X_2 , and Y_1 to refer to these year-centered fairness measures. Table 3 presents summary statistics for these centered fairness criteria across ethnic groups, providing an overview of the magnitude and direction of group-specific deviations from the year-specific sample means.

For the DEA framework described in Section 2.2, we preserve the natural input-output relationships essential for meaningful efficiency measurement by using the original observed values with minimal transformation. Specifically, we use the raw waitlist duration $X_{1i}^{(t)}$ and graft lifespan $Y_{1i}^{(t)}$ directly, while applying a constant shift to LKDPI scores equal to the absolute value of the minimum observed LKDPI value to ensure nonnegativity, as required by DEA. We then apply year-specific centering to the resulting DEA efficiency scores themselves, as described in Section 4.1, to account for temporal changes in the efficiency frontier while preserving the interpretability of relative performance measures.

TABLE 3
Summary statistics of centered fairness measures by ethnic group

Variable	Asian	Black	Hispanic	White
Priority Fairness				
Mean (SD)	83.6 (612.8)	90.2 (690.3)	44.5 (697.4)	−87.0 (528.4)
Median [IQR]	−125.79 [−386.94, 448.17]	−142.69 [−392.80, 334.00]	−212.67 [−462.61, 294.13]	−251.35 [−434.83, 71.60]
Access Fairness				
Mean (SD)	−3.6 (21.1)	2.1 (20.8)	−3.0 (19.4)	0.2 (20.6)
Median [IQR]	−4.48 [−18.84, 9.81]	0.50 [−12.94, 14.57]	−4.50 [−16.76, 9.53]	−1.55 [−14.57, 13.40]
Outcome Fairness				
Mean (SD)	−0.8 (692.1)	−5.5 (712.8)	−94.0 (648.4)	40.4 (684.4)
Median [IQR]	150.78 [−108.53, 332.06]	99.06 [−167.64, 368.42]	−8.09 [−293.96, 291.70]	158.92 [−129.53, 427.56]

3.3. *Fairness criteria analysis.* To thoroughly evaluate fairness in kidney allocation, we study three criteria: Priority Fairness (waitlist duration), Access Fairness (LKDPI score), 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 solely by medical factors. To achieve this, we use mediation analysis to decompose the total effect of ethnicity into direct and indirect effect pathways (MacKinnon (2012)). This approach enables us to isolate disparities that persist even after accounting for observable patient characteristics.

To identify which variables to include as mediators or covariates in the mediation model, we implemented a structured two-stage hypothesis testing procedure. In the first stage, we assessed the association between each candidate variable $M_j \in M$ and the outcome X_1 (waitlist duration) using likelihood ratio tests within a generalized linear modeling framework. Each test compared a full model including M_j to a reduced model excluding it, thereby evaluating the incremental explanatory power of M_j conditional on the other variables. In the second stage, we tested whether each candidate variable M_j was statistically associated with the treatment variable G (ethnicity). Here, the choice of model depended on the structure of the mediator. We used linear models for continuous mediators, logistic regression for binary mediators, and multinomial logistic regression for categorical mediators with more than two levels. In each case, we compared a full model that included ethnicity as a predictor to a null model that excluded it and computed likelihood ratio statistics to assess significance. To control for false discoveries across multiple comparisons, we applied the Benjamini–Hochberg procedure (Benjamini and Hochberg (1995)) in both stages to adjust the resulting p -values. Candidate variables with BH-adjusted p -values less than 0.05 in both stages were classified as mediators, as they showed evidence of being influenced by ethnicity and of affecting waitlist duration. Variables significantly associated only with the outcome were classified as covariates. Variables not significantly associated with waitlist duration were excluded from the mediation model, as they did not contribute to explaining the variation in transplant priority.

Based on this classification procedure, we include the following variables as mediators in the final model: Recipient Age, ABO Blood Type, UNOS Transplant Region, Recipient Education, Recipient Citizenship Status, Recipient Employment Status, and Recipient Dialysis Status. These variables were significantly associated with both ethnicity and waitlist duration,

indicating that they may lie along a causal pathway connecting the two. Additionally, we include the PRA Score as a covariate, as it is significantly associated with waitlist duration but not with ethnicity. All other candidate variables were excluded because they did not exhibit a significant association in the initial testing stage. Detailed results from both stages of the testing procedure, including test statistics and adjusted p -values for each candidate variable, are provided in Section S3.1 of the Supplementary Material (Kaaempur-Mofrad and Dai (2026)).

We formulate the mediation model as a system of regression equations to estimate both the direct effects of ethnicity and the mediated (indirect) pathways through the selected variables. Let G denote the categorical variable for ethnicity, and define X_1 as the observed waitlist duration. The set of mediators is denoted $M = (M_1, \dots, M_7)$, where each M_m corresponds to one of the seven mediator variables listed above. Let C denote the PRA covariate. The total, direct, and indirect effects are estimated using the following linear model:

$$X_1 = \beta_0 + \sum_{g \in \mathcal{G} \setminus \{\text{White}\}} \beta_g \mathbb{I}(G = g) + \tau^\top M + \zeta C + \epsilon,$$

where β_g represents the direct effect of ethnicity group g on waitlist duration, τ is the vector of coefficients for the mediators M , and ζ is the coefficient for the covariate C .

Each mediator M_m is regressed individually on the group indicator G using the appropriate regression model, depending on its type (logistic for binary, multinomial for categorical, and linear for continuous mediators). For each mediator M_m , the fitted model takes the general form,

$$M_m = \gamma_0 + \sum_{g \in \mathcal{G} \setminus \{\text{White}\}} \gamma_{g,m} \mathbb{I}(G = g) + \nu_m,$$

where $\gamma_{g,m}$ captures the effect of ethnicity g on mediator m , and ν_m is the residual error term capturing variation in M_m not explained by ethnicity.

The indirect effect for each ethnic group g is expressed as

$$\text{IE}_g = \sum_m \tau_m (\mathbb{E}[M_m | G = g] - \mathbb{E}[M_m | G = \text{White}]),$$

and the direct effect is then defined as

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

where $\text{TE}_g = \mathbb{E}[X_1 | G = g] - \mathbb{E}[X_1 | G = \text{White}]$ is the total effect. This formulation allows us to isolate disparities in waitlist duration attributable to ethnicity that remain after adjusting for all mediators and covariates included in the model. A significant $\hat{\beta}_g$ suggests that, even after accounting for the factors in our model, there remain substantial differences in waitlist duration for ethnic group g compared to the White patient reference group.

We estimate these effects using the `mma` package in R with 1000 bootstrap samples to obtain standard errors and confidence intervals. Results from this analysis are presented in Table 4, showing significant disparities in waitlist durations across ethnic groups. Substantial direct effects for all minority groups indicate unfairness not fully explained by the chosen mediators. The varying proportions of mediated effects suggest nonuniform mechanisms underlying these disparities.

Access fairness: Counterfactual analysis with random forest. Access Fairness investigates potential inequalities in the quality of kidneys allocated across different ethnic groups, using the LKDPI score as a key metric. We employed a counterfactual analysis using a random forest algorithm to isolate the effect of ethnicity on LKDPI scores, while accounting

TABLE 4
Summary of direct, indirect, and total effects for each ethnic group. *** indicates $p\text{-value} < 0.001$

Patient ethnicity	Effect type	Estimate	SE	95% CI
Asian	Direct Effect**	132.433	50.391	[37.006, 228.527]
	Indirect Effect*	52.847	24.671	[7.679, 104.372]
	Total Effect***	185.280	49.967	[96.275, 285.998]
	Proportion Mediated	0.285		
Black	Direct Effect***	187.601	26.442	[135.300, 240.471]
	Indirect Effect	−10.157	10.039	[−29.608, 9.627]
	Total Effect***	177.445	26.102	[128.225, 226.646]
	Proportion Mediated	0.057		
Hispanic	Direct Effect***	151.760	34.889	[88.457, 221.877]
	Indirect Effect	−23.345	19.396	[−61.872, 12.478]
	Total Effect***	128.415	31.323	[66.728, 192.739]
	Proportion Mediated	0.182		

for other influential factors. To implement our methodology with five-fold cross-validation, consider the framework,

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

where $\hat{X}_2$ denotes the predicted LKDPI 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 donor clinical and demographic characteristics involved in the computation of LKDPI scores (age, eGFR, BMI, SBP, ethnicity, smoking history, hypertension, diabetes, hepatitis C status, HLA mismatches, donor-recipient gender pair, ABO incompatibility, and donor-recipient weight ratio), along with recipient socioeconomic covariates such as education, insurance type, employment status, and UNOS transplant region. 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 at least five samples, ensuring a balance between model complexity and predictive power.

To understand the factors influencing LKDPI 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 LKDPI scores. Specifically, donor age emerged as the most important feature, followed by donor sex, eGFR, Black ethnicity, smoking history, and systolic blood pressure. Other relevant features include donor-recipient gender match, HLA mismatches, ABO incompatibility, and the donor-to-recipient weight ratio. Among recipient-related and socioeconomic features, only marginal effects were observed, with variables such as education, insurance type, and citizenship status contributing little to overall predictive accuracy. Our model demonstrated strong performance with high predictive accuracy, yielding a coefficient of determination $R^2 = 0.985$. The model's performance is further quantified by the mean absolute error, $\text{MAE} = n^{-1} \sum_{i=1}^n |\hat{X}_{2,i} - X_{2,i}| = 1.793$, and the mean squared error, $\text{MSE} = n^{-1} \sum_{i=1}^n (\hat{X}_{2,i} - X_{2,i})^2 = 6.371$, where $\hat{X}_{2,i}$ is the predicted LKDPI score and $X_{2,i}$ is the actual LKDPI score for the i th observation. To isolate the effect of patient ethnicity, we

TABLE 5
Summary of LKDPI score disparities across ethnicities

Ethnicity	Mean LKDPI (actual)	Mean LKDPI (counterfactual)	Mean diff.
White	−0.0041	-	-
Asian	−2.974	−2.8871	0.087
Black	1.782	1.697	−0.0849
Hispanic	−3.0295	−2.947	0.0823

created a counterfactual scenario in which the ethnicity of all recipients was set to “White,” enabling the estimation of the effect of ethnicity on LKDPI scores.

The results of this analysis are summarized in Table 5, which shows minimal disparities in LKDPI scores across ethnic groups, with the largest difference observed in the Asian group (a mean difference of 0.087 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 LKDPI 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.

Outcome fairness: Competing risks analysis. Outcome Fairness focuses on 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

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

where $h_k(y_1|G, W)$ 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 , G is a vector of effect-coded variables for ethnic groups (Asian, Black, Hispanic) using sum-to-zero contrasts, and W is a vector of observed covariates including employment status at time of transplant and whether the patient received posttransplant rejection treatment. The coefficient vector δ captures the estimated deviations of each ethnic group from the overall sample mean subdistribution hazard for graft rejection, the primary event of interest, while treating all other causes of graft failure as competing risks and adjusting for clinical and socioeconomic confounders. The cause-specific cumulative incidence function for graft rejection is

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

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.

Figure 3 illustrates the cumulative incidence functions for graft rejection and competing risks, stratified by ethnicity and adjusted for observed covariates. Hispanic recipients exhibit the highest adjusted cumulative incidence of graft rejection throughout the posttransplant period, followed by White, Asian, and Black recipients. These differences are consistent with the subdistribution hazard ratios reported in Table 6, which indicate that Hispanic patients have a statistically significant elevation in rejection risk relative to the overall sample mean (HR = 2.07, 95% CI: 1.37–3.13, p < 0.001). In contrast, Black recipients demonstrate the

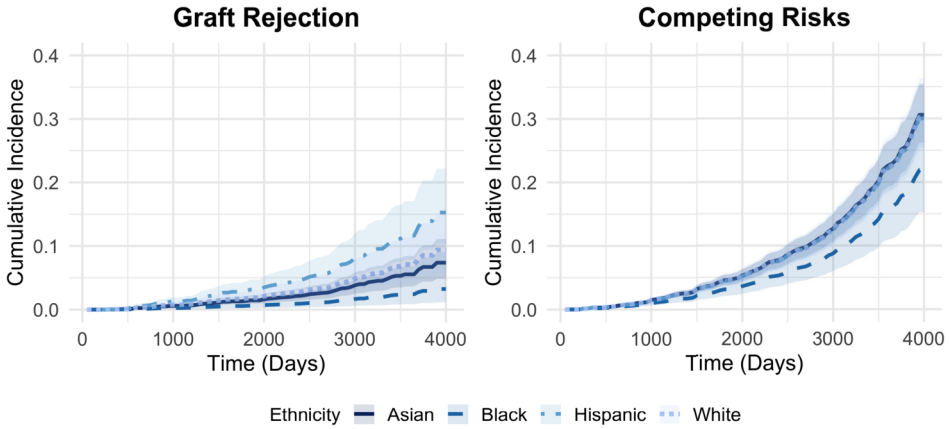


FIG. 3. Cumulative Incidence Functions for Graft Rejection and Competing Risks by Ethnicity.

lowest risk of graft rejection relative to the sample average, though this difference does not reach statistical significance ($HR = 0.41$, 95% CI: 0.14–1.18, $p = 0.098$). Asian and White recipients show risk profiles closer to the sample average, with neither group demonstrating statistically significant deviations from the population mean. In contrast, for graft failure attributable to causes other than rejection, the risk patterns differ considerably from those observed for immunologic failure. This divergence in risk profiles across failure types emphasizes the multifaceted nature of transplant outcomes. The findings suggest that Hispanic recipients are disproportionately affected by graft rejection, even after adjustment for socioeconomic and posttransplant clinical factors, while Black recipients exhibit a lower risk of rejection-related graft loss compared to the study population average. Asian and White recipients demonstrate risk profiles that approximate the population mean for rejection-related graft failure.

4. DEA estimation and inference. Given the complex and sometimes conflicting results from our individual fairness criteria analyses, we use 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 conditional DEA model defined in (4), we incorporate waitlist duration and LKDPI 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 production frontier for each patient is localized using the covariate-conditional set $\mathcal{T}_{Z_i}$, as described in Section 2.2, enabling the resulting efficiency score $\theta_i(Z_i)$ to reflect performance relative to patients with similar socioeconomic and clinical backgrounds.

TABLE 6
Subdistribution hazard ratios for graft rejection relative to overall sample mean, adjusted for covariates W .
Coefficients use sum-to-zero contrasts; the White coefficient is the negative sum of the other groups

Ethnicity	δ	$\exp(\delta)$	$SE(\delta)$	z	p -value	95% CI for $\exp(\delta)$
Asian	−0.046	0.955	0.220	−0.208	0.835	[0.621, 1.470]
Black	−0.888	0.412	0.536	−1.657	0.098	[0.144, 1.177]
Hispanic	0.727	2.068	0.212	3.428	0.001	[1.365, 3.133]
White	0.207	1.230	0.252	0.821	0.412	[0.750, 2.016]

TABLE 7

Relative conditional DEA scores and group-conditional prediction intervals by ethnic group across cumulative time periods. Efficiency scores were estimated using the Benchmarking package in R

Period	Group	Mean relative efficiency	Prediction interval (95%)	Coverage
2010–2013	Asian	−0.053	[−0.6581, 0.3767]	0.956
	Black	−0.0024	[−0.5655, 0.3609]	0.953
	Hispanic	−0.0113	[−0.5709, 0.3827]	0.954
	White	0.0074	[−0.5593, 0.3754]	0.952
2010–2016	Asian	−0.0073	[−0.5300, 0.3773]	0.952
	Black	−0.0060	[−0.5136, 0.3459]	0.949
	Hispanic	−0.0152	[−0.4877, 0.3364]	0.950
	White	0.0104	[−0.4911, 0.3236]	0.950
2010–2019	Asian	0.0055	[−0.4879, 0.3551]	0.956
	Black	−0.0071	[−0.4874, 0.3363]	0.950
	Hispanic	−0.0047	[−0.4377, 0.3279]	0.950
	White	0.0066	[−0.4466, 0.3189]	0.950

To assess disparities over time and adjust for temporal changes in the transplant landscape, we compute relative efficiency scores by centering each patient’s score against the mean efficiency of transplants performed in the same calendar year. Specifically, for each patient i transplanted in year t , we define the centered efficiency score as $\tilde{\theta}_i := \theta_i - \bar{\theta}_t$, where $\bar{\theta}_t$ is the mean conditional DEA score among all KPD recipients in year t . This centering procedure accounts for year-to-year variation and enables comparison of relative efficiency levels across groups.

The results in Table 7, together with the distributions shown in Figure 4, provide a comprehensive view of efficiency across three cumulative time periods (2010–2013, 2010–2016, and 2010–2019). For each ethnic group and period, we report the mean centered efficiency score $\tilde{\theta}_i$, along with 95% prediction intervals and empirical coverage rates, all computed using group-conditional conformal prediction under the Reference Frontier Mapping (RFM) procedure introduced in Section 2.3 and averaged over 100 random train–calibration–test

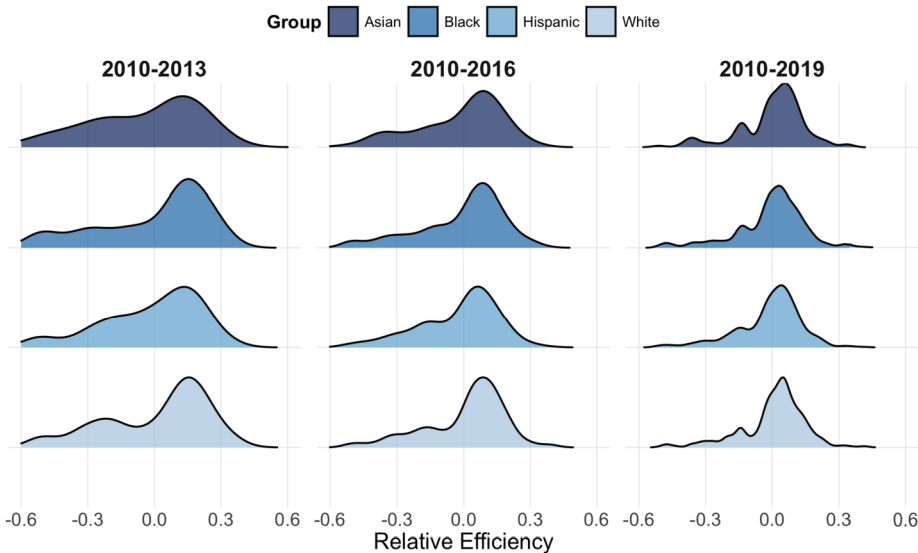


FIG. 4. Distribution of Relative Efficiency Scores by Ethnic Group.

splits to account for sampling variability. These prediction intervals ensure that, for each group, the interval contains the true efficiency score with 95% probability, providing robust and equitable uncertainty quantification across groups. The mean relative efficiency scores are generally close to zero across all groups, as expected from the year-specific centering. White patients tend to show slightly positive relative efficiency scores across all periods, while other groups tend to fall marginally below their respective yearly averages. However, the differences are modest, indicating broad similarity in average efficiency across groups when conditioning on year and covariate structure. A more pronounced distinction emerges in the width of the prediction intervals. White patients exhibit the narrowest uncertainty intervals, indicating more homogeneous efficiency outcomes, whereas Asian, Black, and Hispanic patients have noticeably wider intervals, suggesting greater within-group heterogeneity in KPD transplant efficiency. It is important to emphasize that these wide group-conditional prediction intervals limit the strength of any claims about equality of efficiency across ethnic groups. By design, the conformal procedure is calibrated to achieve 95% within-group coverage, and the interval width reflects both model uncertainty and the intrinsic variability of efficiency scores in a small, highly selected KPD population. As a result, the substantial overlap between intervals should not be interpreted as evidence that the allocation system is fair or that groupwise efficiency levels are identical. Rather, the table indicates that, given current sample sizes and noise levels, we do not detect large average differences in overall efficiency once priority, access, and outcome are aggregated and conditioned on covariates. In this sense the DEA-based summaries are best viewed as diagnostic tools that complement the more pronounced criterion-specific disparities documented in Section 3.3, rather than as definitive proof of group-level fairness.

4.2. Analysis of efficiency distributions. To rigorously evaluate the statistical significance of the observed differences in efficiency distributions, we employed a kernel-based hypothesis-testing approach based on the maximum mean discrepancy (MMD, [Gretton et al. \(2012\)](#)). We conducted both pairwise comparisons and “group-vs.-rest” analyses across three cumulative time periods (2010–2013, 2010–2016, and 2010–2019) to assess how distributional differences evolve with sample size and temporal scope. For each comparison between 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 relative efficiency scores.

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 $\theta_A, \theta'_A \stackrel{i.i.d.}{\sim} F_A$ and $\theta_B, \theta'_B \stackrel{i.i.d.}{\sim} F_B$ and $K(\cdot, \cdot)$ is a Laplacian kernel chosen to effectively detect variations in the distributions by measuring similarity between observations based on their absolute differences. Specifically, for efficiency scores θ_i and θ_j , the Laplacian kernel is $K(\theta_i, \theta_j) = \exp(-\sigma \|\theta_i - \theta_j\|)$, where σ is set to the median Euclidean distance between points in the aggregate sample for each comparison. This kernel achieves maximum similarity (value of 1) when two observations are identical and decreases exponentially as their absolute difference increases. The Laplacian kernel is particularly well suited for our context because it is sensitive to local distributional differences, making it effective at detecting variations in distributional spread and heterogeneity ([Gretton et al. \(2012\)](#)). This choice aligns with our finding that White patients exhibit narrower prediction intervals (indicating more homogeneous outcomes) while Asian, Black, and Hispanic patients show wider intervals (suggesting greater within-group heterogeneity). The Laplacian kernel effectively captures these differences in distributional shape and spread, which are central to understanding equity in transplant outcomes. We determined statistical significance via permutation tests

TABLE 8
*MMD Test results across time periods. * indicates $p < 0.05$ after Benjamini–Hochberg correction*

Period	Comparison	MMD statistic	p -value	Adjusted p -value
2010–2013	Asian vs. Black	−0.0002	0.427	0.577
	Asian vs. Hispanic	−0.001	0.792	0.792
	Asian vs. White	−0.00002	0.319	0.577
	Black vs. Hispanic	0.0004	0.154	0.533
	Black vs. White	−0.0001	0.481	0.577
	Hispanic vs. White	0.0003	0.178	0.533
	Asian vs. Rest	−0.0003	0.457	0.457
	Black vs. Rest	−0.000008	0.351	0.457
	Hispanic vs. Rest	0.0003	0.176	0.457
	White vs. Rest	−0.00002	0.377	0.457
2010–2016	Asian vs. Black	−0.0004	0.992	0.992
	Asian vs. Hispanic	−0.0002	0.637	0.802
	Asian vs. White	−0.0002	0.669	0.802
	Black vs. Hispanic	0.00008	0.202	0.403
	Black vs. White	0.00004	0.172	0.403
	Hispanic vs. White	0.0004	0.004	0.024*
	Asian vs. Rest	−0.0003	0.946	0.946
	Black vs. Rest	−0.00003	0.559	0.745
	Hispanic vs. Rest	0.0002	0.048	0.096
	White vs. Rest	0.0001	0.028	0.096
2010–2019	Asian vs. Black	−0.000002	0.355	0.563
	Asian vs. Hispanic	−0.000003	0.375	0.563
	Asian vs. White	−0.00008	0.928	0.928
	Black vs. Hispanic	−0.00004	0.910	0.928
	Black vs. White	0.00007	0.020	0.120
	Hispanic vs. White	0.00004	0.120	0.359
	Asian vs. Rest	−0.00006	0.677	0.677
	Black vs. Rest	0.00003	0.080	0.160
	Hispanic vs. Rest	−0.000002	0.389	0.519
	White vs. Rest	0.000052	0.006	0.024*

with 500 Monte Carlo iterations and applied Benjamini–Hochberg correction to control the false discovery rate across multiple comparisons.

The results in Table 8 reveal a consistent pattern of distributional differences that becomes more detectable as sample sizes increase across time periods. In the earliest period (2010–2013), no significant differences were detected between any group pairs, likely due to limited statistical power with smaller sample sizes. As the sample expanded to the 2010–2016 period, the Hispanic vs. White comparison became statistically significant (adjusted p -value = 0.024), suggesting these groups have detectably different efficiency distributions. With the largest sample (2010–2019), while the pairwise Hispanic vs. White comparison loses significance, the “group-vs.-rest” analysis reveals that White patients show a significant difference from all other groups combined (adjusted p -value = 0.024). The “group-vs.-rest” analyses provide critical insights into how individual groups differ from the broader transplant population. The pattern shows that White patients consistently exhibit distributional characteristics that distinguish them from the broader transplant population, with this difference becoming statistically significant in the largest sample. This finding provides statistical evidence supporting the descriptive findings from Section 4.1 regarding differential within-group variability across ethnic groups.

At the same time, the general lack of statistically significant pairwise differences across many comparisons should not be overinterpreted. Failure to reject $H_0 : F_A = F_B$ does not imply that the underlying efficiency distributions are equal or that allocation outcomes are fair across ethnic groups. Rather, it reflects the combined effect of sample size, substantial within-group variability, and the fact that efficiency aggregates multiple fairness criteria that already exhibit clear disparities when analyzed individually. Within these limitations the MMD tests indicate only that we do not detect large distributional shifts in *overall* efficiency between most group pairs. These findings suggest that while average efficiency scores are similar across groups after accounting for temporal trends, there are meaningful differences in the distributional properties of efficiency scores that may reflect disparities in access to optimal transplant outcomes within the kidney paired donation system. Taken together, the MMD and descriptive results highlight DEA's role as a synthesis and monitoring tool rather than a standalone determinant of 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 7 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 pretransplant care (Kucirka et al. (2012)). From a practical standpoint, the combination of centered efficiency scores and group-conditional prediction intervals provides a summary diagnostic for each ethnic group that goes beyond average outcomes: programs can monitor whether a group's efficiency distribution remains systematically wider or more skewed than that of others, even when mean performance is comparable. Taken together, these findings suggest the need for tailored interventions to improve outcome consistency for underrepresented groups (Gordon et al. (2014)).

Second, incorporating multiple efficiency measures, as in the DEA model (4), can 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 conditional DEA framework to assess the impact of policy changes on overall fairness across ethnic groups. In practice, this does not certify that a given allocation rule is fair, but it provides a structured way to quantify how proposed changes shift groupwise efficiency distributions while jointly accounting for multiple fairness-relevant criteria of the allocation process.

A concrete way to use our framework in practice is via policy simulation. Modern KPD platforms, such as StanfordKPD (Sakshuwong, Ashlagi and Roth (2023)), which support intelligent matching, prioritization of urgent patients, adjustable chain-length rules, and fine-grained immunologic constraints, allow programs to rerun historical pools or generate counterfactual match-runs under modified allocation policies. For example, a KPD program considering changes aimed at reducing long wait times among Asian and Black patients, such as increasing the priority weights for these groups in chain initiation or bridge-donor selection, can implement these adjustments within a platform like StanfordKPD and obtain the resulting set of simulated match outcomes. Using these simulated transplants, the program can then apply predictive models to estimate the priority, access, and outcome criteria scores for each match and compute conditional DEA scores via (4) under the proposed mechanism.

Comparing the resulting efficiency distributions and their group-conditional prediction intervals to those under the status quo provides a principled diagnostic for assessing whether the policy improves groupwise fairness without inducing large efficiency losses or excessive variability for other groups. In this way the conditional DEA scores serve as a common currency for evaluating competing allocation proposals that alter priority, access, and outcome in different ways. More broadly, the conditional DEA and reference frontier mapping (RFM) framework, introduced in Section 2.3, can be embedded into a routine monitoring pipeline. Programs can periodically reconstruct the RFM reference set, recompute group-specific conditional efficiency scores via (4), and update group-conditional prediction intervals using the conformal procedure of Section 2.3 as new transplants accrue. Because covariates localize the production frontier to $\mathcal{T}_{Z_i}$, observed disparities in these diagnostics can be interpreted as differences in how effectively similar patients are treated rather than as artifacts of case-mix variation. This enables a feedback-driven workflow in which clinicians and policymakers: (i) use standard clinical tools to design candidate changes to matching rules or eligibility criteria, (ii) screen those changes with DEA and RFM-based fairness diagnostics, and (iii) implement only policies that improve groupwise efficiency distributions while maintaining, at minimum, current outcomes for all other groups.

Despite these advantages, several practical barriers limit the immediate implementation of DEA-based fairness monitoring in real KPD programs. First, our framework requires detailed, high-quality covariate data to construct conditional frontiers and RFM reference sets; in practice, centers differ in how completely they record socioeconomic variables, post-transplant care, and follow-up outcomes, which can constrain the reliability of groupwise efficiency estimates for smaller subpopulations. Additionally, translating fairness diagnostics into policy requires agreement among clinicians, administrators, and regulators on acceptable trade-offs between efficiency and equity; interventions that narrow the efficiency distribution for one group may be perceived as limiting short-term matching flexibility or complicating existing allocation rules. Finally, routine monitoring can only be sustained if the resulting flags are interpretable and actionable; without clear governance about how to respond when a particular group's efficiency distribution degrades, there is a risk that fairness reports become merely procedural rather than a driver of meaningful change. These barriers suggest that, at least initially, our framework is most realistically deployed as an offline audit tool that is used periodically to evaluate the implications of proposed policy changes and to benchmark centers, prior to being integrated into real-time decision support as data infrastructure and institutional capacity improve.

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 pre-defined weights in DEA is that it allows for an objective assessment of fairness. Our approach highlights where disparities emerge across different fairness criteria and offers a diagnostic framework for policymakers to evaluate how candidate policies might improve groupwise efficiency, while also making explicit the data, modeling, and governance infrastructure required for routine deployment.

Several research directions could build on this conditional DEA-based framework. For example, dynamic conditional DEA models could capture how fairness evolves as KPD programs and matching algorithms change. 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, accounting for 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.

Acknowledgment. The authors express their gratitude to the Editor, Associate Editor, and reviewers for their valuable and constructive comments, which have greatly enhanced the quality of the paper.

Address for correspondence: Xiaowu Dai, Department of Statistics and Data Science, University of California, Los Angeles, 8125 Math Sciences Bldg #951554, Los Angeles, CA 90095, USA.

Funding. This work was partially supported by NIH Grant R01DK142026, NIH dkNET AI Pilot Program, a Merck Research Award, and a Hellman Fellowship Award.

SUPPLEMENTARY MATERIAL

Supplement to “A data envelopment analysis approach for assessing fairness in resource allocation: Application to kidney exchange programs” (DOI: [10.1214/25-AOAS2128SUPPA](https://doi.org/10.1214/25-AOAS2128SUPPA); .pdf). Extended methodological details for the uncertainty quantification procedure, technical proofs for theorems, a simulation study to examine the impact of resampling on efficiency scores, and detailed variable selection procedures for the fairness analyses are included in this file.

R code (DOI: [10.1214/25-AOAS2128SUPPB](https://doi.org/10.1214/25-AOAS2128SUPPB); .zip). R code implementing the proposed method is available both in the Supplementary Material and at <https://github.com/amofrad/Kidney-Exchange-DEA>.

REFERENCES

- AGARWAL, N., ASHLAGI, I., AZEVEDO, E., FEATHERSTONE, C. R. and KARADUMAN, Ö. (2019). Market failure in kidney exchange. *Amer. Econ. Rev.* **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. *Amer. J. Transplant.* **18** 1168–1176.
- BÄDIN, L., DARAIO, C. and SIMAR, L. (2012). How to measure the impact of environmental factors in a non-parametric production model. *European J. Oper. Res.* **223** 818–833.
- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B, Methodol.* **57** 289–300.
- BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.
- CENTERS FOR DISEASE CONTROL AND PREVENTION (2023). Chronic kidney disease in the United States, 2023. US Dept. Health and Human Services, Centers for Disease Control and Prevention, Atlanta, GA.
- CHARNES, A., COOPER, W. W. and RHODES, E. (1978). Measuring the efficiency of decision making units. *European J. Oper. Res.* **2** 429–444.
- CHILINGERIAN, J. A. (1995). Evaluating physician efficiency in hospitals: A multivariate analysis of best practices. *European J. Oper. Res.* **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 Int.* **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. *J. Amer. Med. Assoc.* **298** 2038–2047.

- FAN, P.-Y., ASHBY, V. B., FULLER, D., BOULWARE, L., KAO, A., NORMAN, S. P., RANDALL, H., YOUNG, C., KALBFLEISCH, J. D. et al. (2010). Access and outcomes among minority transplant patients, 1999–2008, with a focus on determinants of kidney graft survival. *Amer. J. Transplant.* **10** 1090–1107.
- FÄRE, R., GROSSKOPF, S. and LOVELL, C. K. (2013). *The Measurement of Efficiency of Production* 6. Springer, Berlin.
- FINE, J. P. and GRAY, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *J. Amer. Statist. Assoc.* **94** 496–509.
- GENTRY, S. E., MONTGOMERY, R. A. and SEGEV, D. L. (2011). Kidney paired donation: Fundamentals, limitations, and expansions. *Amer. J. Kidney Dis.* **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. *Amer. J. Transplant.* **7** 2361–2370.
- GIBBS, I., CHERIAN, J. J. and CANDÈS, E. J. (2025). Conformal prediction with conditional guarantees. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **87** 1100–1126.
- 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. *Prog. Transplant.* **24** 152–162.
- GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773.
- HARDT, M., PRICE, E. and SREBRO, N. (2016). Equality of opportunity in supervised learning. *Adv. Neural Inf. Process. Syst.* **29**.
- HARIHARAN, S., ISRANI, A. K. and DANOVIATCH, G. (2021). Long-term survival after kidney transplantation. *N. Engl. J. Med.* **385** 729–743.
- HE, H. and GARCIA, E. A. (2009). Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21** 1263–1284.
- IMAI, K., KEELE, L. and YAMAMOTO, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.* **25** 51–71.
- KAAZEMPUR-MOFRAD, A. and DAI, X. (2026). Supplement to “A Data Envelopment Analysis Approach for Assessing Fairness in Resource Allocation: Application to Kidney Exchange Programs.” <https://doi.org/10.1214/25-AOAS2128SUPPA>, <https://doi.org/10.1214/25-AOAS2128SUPPB>
- KUCIRKA, L. M., GRAMS, M., BALHARA, K. S., JAAR, B. G. and SEGEV, D. L. (2012). Disparities in provision of transplant information affect access to kidney transplantation. *Amer. J. Transplant.* **12** 351–357.
- KUSNER, M. J., LOFTUS, J., RUSSELL, C. and SILVA, R. (2017). Counterfactual fairness. *Adv. Neural Inf. Process. Syst.* **30**.
- LENTINE, K., SMITH, J., HART, A., MILLER, J., SKEANS, M., LARKIN, L., ROBINSON, A., GAUNTT, K., ISRANI, A. et al. (2022). OPTN/SRTR 2020 annual data report: Kidney. *Amer. J. Transplant.* **22** 21–136.
- MACKINNON, D. (2012). *Introduction to Statistical Mediation Analysis*. Routledge, London.
- MALEK, S. K., KEYS, B. J., KUMAR, S., MILFORD, E. and TULLIUS, S. G. (2011). Racial and ethnic disparities in kidney transplantation. *Transpl. Int.* **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. *Amer. J. Transplant.* **13** 1317–1322.
- MASSIE, A. B., LEANZA, J., FAHMY, L. M., CHOW, E. K., DESAI, N. M., LUO, X., KING, E. A., BOWRING, M. G. and SEGEV, D. L. (2016). A risk index for living donor kidney transplantation. *Amer. J. Transplant.* **16** 2077–2084.
- 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. *J. Appl. Math.* **2014** Art. ID 612018, 25.
- OZCAN, Y. A. et al. (2008). *Health Care Benchmarking and Performance Evaluation*. Springer, Berlin.
- PERSAD, G., WERTHEIMER, A. and EMANUEL, E. J. (2009). Principles for allocation of scarce medical interventions. *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. *Transplant. Proc.* **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. et al. (2017). Kidney exchange to overcome financial barriers to kidney transplantation. *Amer. J. Transplant.* **17** 782–790.

- ROTH, A. E., SÖNMEZ, T. and ÜNVER, M. U. (2004). Kidney exchange. *Q. J. Econ.* **119** 457–488.
- SAKSHUWONG, S., ASHLAGI, I. and ROTH, A. E. (2023). StanfordKPD: a Kidney Exchange Platform for Kidney Paired Donation.
- SHERO, J. A., AL OTAIBA, S., SCHATSCHEIDER, C. and HART, S. A. (2022). Data envelopment analysis (DEA) in the educational sciences. *J. Exp. Educ.* **90** 1021–1040.
- THOKALA, P., DEVLIN, N., MARSH, K., BALTUSSEN, R., BOYSEN, M., KALO, Z., LONGRENN, T., MUSSEN, F., PEACOCK, S. 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 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. *Amer. J. Transplant.* **11** 2093–2109.
- UNITED STATES RENAL DATA SYSTEM (2024). 2024 USRDS Annual Data Report: Epidemiology of kidney disease in the United States.
- 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 Int. Rep.* **7** 1278–1288.
- WANG, W., REES, M. A., LEICHTMAN, A. B., SONG, P. X.-K., BRAY, M., ASHBY, V. B., SHEARON, T., WHITE-MAN, A. and KALBFLEISCH, J. D. (2021). Deceased donors as nondirected donors in kidney paired donation. *Amer. J. Transplant.* **21** 103–113.
- WOLFE, R. A., ASHBY, V. B., MILFORD, E. L., OJO, A. O., ETTINGER, 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. *N. Engl. J. Med.* **341** 1725–1730.
- WOLFE, R. A., MCCULLOUGH, K. P., SCHAUBEL, D. E., KALBFLEISCH, J. D., MURRAY, S., STEGALL, M. D. and LEICHTMAN, A. B. (2008). Calculating life years from transplant (LYFT): Methods for kidney and kidney-pancreas candidates. *Amer. J. Transplant.* **8** 997–1011.