Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Quality Benchmarking and Rank Stability of U.S. Kidney Transplant Centers Using Donor-Type-Specific Data Envelopment Analysis

**Feyza GURER**

feyza.gurer@yuksekihtisas.edu.tr

Yüksek İhtisas Üniversitesi

**Research Article**

**Keywords:** kidney transplantation, transplant center performance, data envelopment analysis, healthcare benchmarking, quality improvement, rank stability

**Posted Date:** November 13th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11180335/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Abstract

Publicly reported kidney transplant outcomes are used for center evaluation, but limited case-mix information and closely clustered outcomes can make comparative rankings appear more precise than the available data justify. This study evaluated multidimensional center performance and the stability of relative rankings using public aggregate data. Data included 423 donor-type records from 227 U.S. kidney transplant centers, representing 148,415 transplants. Expected one-year patient survival and functioning graft proportions were estimated separately for deceased and living donor transplantation using fractional response models with three aggregate measures of recipient composition. Observed-to-expected indices were jointly evaluated using output-oriented data envelopment analysis under variable returns to scale. Stability was assessed by perturbing the reference set and by repeating the full adjustment and benchmarking pipeline within subsamples. Mean quality scores were 0.956 for deceased donor records and 0.979 for living donor records, with one and two strong frontier records, respectively. Unadjusted outcomes produced substantially more frontier records. Among living donor records, 26.9% had adjusted scores of at least 0.990, indicating limited separation near the frontier. Most sensitivity analyses preserved the broader score pattern. Full-pipeline subsampling largely preserved overall ordering but produced greater variation in individual rank percentiles than reference-set perturbation alone. Public aggregate data can support multidimensional comparative assessment of kidney transplant centers, but closely clustered scores and instability in exact rank position limit the interpretation of precise rankings. These measures are better suited to screening and comparative review than to definitive center ranking.


# 1. Introduction

Kidney transplant center performance is publicly reported in the United States and is used for quality assessment and program evaluation (Dickinson et al. 2006; Schold et al. 2019). Because observed outcomes depend partly on recipient, donor, and transplant characteristics, meaningful center comparisons require adjustment for differences in case mix. Current Scientific Registry of Transplant Recipients (SRTR) models therefore estimate expected outcomes using detailed clinical information and distinguish between living and deceased donor kidney transplantation (Scientific Registry of Transplant Recipients 2026). Risk-adjustment choices can materially affect center classification (Weinhandl et al. 2009), and unfavorable performance evaluations have been associated with subsequent reductions in transplant volume (Schold et al. 2013). Center classifications may also vary substantially across repeated evaluations (Schold et al. 2018).

Public aggregate data provide broad access to center-level outcomes but do not contain the detailed clinical information needed to reproduce official patient-level risk adjustment. Residual differences between centers may therefore reflect unmeasured risk factors as well as differences in quality (Hartman et al. 2024). In addition, when activity and outcome data cover different reporting periods, transplant volumes cannot reliably be used to reconstruct outcome counts. These limitations require approaches that preserve the reported outcome structure while acknowledging incomplete adjustment.

Data envelopment analysis (DEA) can evaluate multiple outcomes jointly without specifying a parametric relationship among them (Charnes et al. 1978; Banker et al. 1984). Although DEA is commonly used to study healthcare efficiency in relation to resources (Hollingsworth 2008; Kohl et al. 2019), transplantation applications have addressed both service-level and patient-level performance (Siqueira and Araujo 2018; Santos Arteaga et al. 2021). The present setting differs because the available data contain no direct measures of staffing, expenditure, or infrastructure. Accordingly, the resulting frontier is interpreted as a multidimensional quality benchmark rather than a measure of technical or operational efficiency.

Two additional considerations are important. First, one-year patient survival and graft function are bounded outcomes that cluster near their upper limits, so small differences may substantially affect frontier membership and ranking. Second, living and deceased donor transplantation differ in recipient characteristics, donor selection, clinical pathways, and outcome risk and are therefore more appropriately evaluated against separate frontiers. Because DEA frontiers also depend on the observations used to construct them, closely spaced scores may yield apparently precise but unstable rankings (Shwartz et al. 2016). Formal bootstrap approaches are available for nonparametric frontier inference (Simar and Wilson 1998, 2007), but the present study focuses on descriptive stability of the benchmarking results.

This study therefore evaluated deceased and living donor kidney transplantation separately using public aggregate data. The objectives were to characterize multidimensional quality profiles after limited adjustment for recipient composition, assess sensitivity to alternative outcome and frontier specifications, and compare rank stability when only the DEA reference set was perturbed with stability when the full adjustment and benchmarking pipeline was re-estimated.

## 2. Methods

# 2.1 Study design and data source

This methodological benchmarking study used publicly available aggregate data for U.S. kidney transplant centers from the Health Resources and Services Administration Organ Donation and Transplantation resource archived through DataLumos/ICPSR (United States Department of Health and Human Services, Health Resources and Services Administration 2025). Transplant activity and recipient composition data were obtained from “OTC Transplants by Demographics,” and one-year post-transplant outcomes from “OTC Outcome Measures.” Activity and covariate data covered January 1, 2019, through November 30, 2024, whereas outcome measures covered January 1, 2019, through November 30, 2023; both worksheets were last refreshed on March 17, 2025. Because the reporting periods did not represent identical transplant cohorts, outcome percentages were analyzed as reported proportions and were not converted to case counts using transplant volume. The dataset contained no direct measures of staffing, expenditure, infrastructure, or operating capacity; therefore, DEA was used for multidimensional quality benchmarking rather than for estimating technical or operational efficiency.

## 2.2 Record linkage, exclusions, and unit of analysis

The analysis was restricted to deceased and living donor kidney transplant records; aggregate “All Centers” rows were excluded. Activity and outcome records were linked by normalized center name, donor type, and organ type. Records were excluded if the center was inactive, no corresponding outcome record was identified, transplant volume was missing, suppressed, or nonpositive, or either one-year patient survival or functioning graft outcome was missing, suppressed, or nonpositive. Data checks verified unique linkage keys, absence of many-to-many matches, and valid analytic ranges. The unit of analysis was the center record within each donor type. A center could therefore contribute a deceased donor record, a living donor record, or both. Deceased and living donor records remained separate throughout adjustment and DEA estimation and were never combined into a common ranking.

## 2.3 Outcomes and aggregate covariates

The two outcomes were the published proportions of patients alive at one year and transplants with a functioning graft at one year. Percentages were converted to proportions on the unit interval. Because aligned outcome denominators were unavailable, no patient-level outcomes, reconstructed case counts, or artificial binomial denominators were created.

Recipient demographic and payer counts were converted to donor-type-specific shares using the corresponding transplant activity volume. Candidate variables included age, pediatric status, race and ethnicity, and payer categories. The primary adjustment used three measures selected to capture recipient composition while limiting missingness and model complexity: the proportions of recipients aged 61 years or older, with private insurance, and covered by Medicare fee for service. These variables were intended as limited aggregate adjustment rather than comprehensive clinical risk adjustment. Transplant volume was retained for descriptive purposes only and was not used as a model predictor, DEA input, or outcome multiplier because activity and outcome periods were not aligned.

## 2.4 Aggregate adjustment by donor type

Expected one-year outcomes were estimated separately for deceased and living donor records. Within each donor group, separate quasibinomial logit fractional response models were fitted for patient survival and functioning graft proportions (Wedderburn 1974; Papke and Wooldridge 1996). The models treated the published proportions directly and did not assume individual-level binomial outcomes or use transplant-volume weights.

Missing covariate values were replaced with the donor-type-specific median, with a corresponding missingness indicator included for each variable. Predictors with no variation within a fitted model were omitted. The fitted conditional means were used as expected outcomes, and observed-to-expected (O/E) indices were calculated for patient survival and functioning graft status. Values above 1 indicated observed outcomes above model-based expectation and values below 1 indicated outcomes below expectation. These two O/E indices were used as the outputs in the primary DEA model. Because the

models relied on only three aggregate measures of recipient composition, the resulting expected values should not be interpreted as equivalent to official SRTR risk-adjusted expectations.

# 2.5 Primary DEA model

Deceased and living donor records were evaluated against separate empirical frontiers using output-oriented Banker-Charnes-Cooper models under variable returns to scale (Banker et al. 1984). Each record was assigned a constant input of 1, and the patient-survival and functioning-graft O/E indices served as the two outputs. With a constant input, the variable-returns-to-scale formulation was used to define a convex output technology rather than to imply an economic interpretation of returns to scale.

The output expansion factor ( $\varphi$ ) was converted to a bounded BCC quality score as $S = \frac{1}{\varphi}$. Scores within the prespecified numerical tolerance of 1 were classified as radial frontier observations; strong frontier membership additionally required no positive Phase II input or output slack. No assurance-region or fixed relative-weight restrictions were imposed, preserving endogenous DEA weighting flexibility. Scores were rounded to 10 decimal places before ranking to avoid distinctions caused only by numerical precision.

Within each donor group, scores were ranked from highest to lowest, with ties assigned the same minimum rank and then converted to rank percentiles, where lower percentiles indicated positions nearer the top of the ordering. Ties were included in the top decile. For descriptive reporting, the BCC score deficit was defined as 1 minus the bounded quality score. This complement was not interpreted as the proportional radial output expansion, a predicted clinical improvement, or a causal effect.

### 2.6 Reference set subsampling

Sensitivity to changes in the DEA comparison set was examined using 1,000 repetitions of approximately 90% subsampling without replacement within each donor group. Each repetition included 203 deceased donor and 177 living donor records. The O/E values from the primary adjustment were held fixed, and only the DEA reference frontier was reconstructed.

For each included record, BCC score, rank percentile, radial and strong frontier membership, slack status, and top-decile membership were retained. Record-level summaries included medians, 5th and 95th percentiles, range widths, and frequencies of strong frontier and top-decile membership across repetitions in which the record was sampled. Because the adjusted outputs remained fixed, this analysis isolated sensitivity to the empirical reference set.

# 2.7 Full analysis pipeline

A second subsampling analysis used the same sampled records in each paired repetition but rebuilt the preceding adjustment steps before reconstructing the DEA frontier. Within each sample, donor-type-specific medians were recalculated, missing values and missingness indicators were reconstructed,

fractional response models were refitted, expected outcomes and O/E indices were recalculated, and the DEA frontier was re-estimated.

The same record-level stability measures were retained. Paired reference-set-only and full-pipeline results were compared using Spearman correlations, absolute differences in BCC scores and rank percentiles, agreement in strong frontier and top-decile membership, top-decile overlap, and differences in 5th–95th percentile range widths. This analysis was designed to describe the additional variation introduced by rebuilding the full analytic pipeline. It did not capture patient-level sampling uncertainty in the published outcome proportions, and the resulting percentile ranges were treated as descriptive stability ranges rather than inferential confidence intervals.

## 2.8 Sensitivity and robustness analyses

Several sensitivity analyses examined dependence on outcome representation, missing data, extreme O/E values, frontier geometry, and numerical resolution. First, the published one-year patient survival and functioning graft proportions were entered directly as DEA outputs without aggregate adjustment. Second, a complete-case analysis retained only records with observed values for all three adjustment covariates and refitted both the adjustment models and DEA frontier. Third, patient-survival and functioning-graft O/E values were winsorized separately within each donor group at the 1st and 99th percentiles.

Sensitivity to the convexity assumption was assessed using a free disposal hull (FDH) with the same outputs and constant input. Concentration near the frontier was described using BCC score thresholds of 0.999, 0.995, and 0.990. Numerical resolution was evaluated by repeating the BCC models after rounding copies of the O/E values to four and three decimal places. Finally, an intercept-only adjustment model was used as a scale-invariance check, because division by a common positive expected value within a donor group should not alter DEA scores or ordering.

Agreement between the primary and alternative specifications was summarized descriptively using Spearman correlations, absolute score and rank-percentile differences, frontier counts and overlaps, and top-decile overlap. No inferential p values or confidence intervals were calculated for these comparisons.

## 2.9 Statistical analysis and reproducibility

DEA scores were not analyzed using second-stage regression because such models would need to account for dependence induced by the shared estimated frontier and uncertainty introduced by the preceding adjustment steps. The subsampling analyses were therefore used for descriptive stability assessment rather than formal second-stage inference.

All analyses were performed in R version 4.5.2 using a fully scripted workflow (R Core Team 2025). DEA models were estimated with the Benchmarking package, version 0.33, using its linear-programming implementation based on lpSolveAPI (Bogetoft and Otto 2025). Radial scores and Phase II slacks were retained separately, and fixed random seeds were used for paired subsampling.

Automated checks verified input-file integrity, record counts, donor-group separation, uniqueness of linkage keys, valid analytic and DEA ranges, completion of all subsampling repetitions, identity of paired samples, and exported-result integrity. Coded center identifiers were used in figures displaying individual records to avoid presentation as a league table. The analysis-ready data, derived adjustment and DEA outputs, sensitivity and stability results, figure mapping tables, and complete R script were retained as reproducibility materials. Because the study used only publicly available aggregate center-level data and involved no identifiable patient information or participant interaction, institutional ethics approval and informed consent were not required.

## 3. Results

# 3.1 Analytic sample and data completeness

Of 484 candidate records, 54 were excluded because of missing or suppressed transplant totals (n = 42), missing or suppressed one-year outcomes (n = 9), or absence of a corresponding outcome record (n = 3); seven additional matched records were excluded because the centers were inactive. The final sample comprised 423 donor-type records from 227 U.S. kidney transplant centers, including 226 deceased donor and 197 living donor records. Reported activity totaled 148,415 kidney transplants, of which 112,689 were deceased donor and 35,726 were living donor transplants [Supplementary Table S1].

Among the three covariates used in the primary adjustment, missingness was 14.7% for private insurance, 19.1% for age 61 years or older, and 21.5% for Medicare fee for service. Missingness in other candidate covariates ranged from 31.7% to 63.6% [Supplementary Fig. S1].

# 3.2 Aggregate adjustment and O/E profiles

All primary fractional response models converged successfully. Mean observed and expected proportions were approximately 0.962 for patient survival and 0.936 for functioning grafts among deceased donor records, and 0.987 and 0.975, respectively, among living donor records.

Patient-survival O/E values ranged from 0.943 to 1.047 for deceased donor records and from 0.943 to 1.022 for living donor records; corresponding functioning-graft O/E ranges were 0.896–1.079 and 0.851–1.045. Spearman correlations between the two O/E indices were 0.798 and 0.696, respectively [Supplementary Fig. S2].

# 3.3 Primary DEA results and scores near the frontier

For deceased donor records, the mean BCC quality score was 0.956, the median was 0.958, and the range was 0.901–1.000. One record was on the radial frontier and had no positive Phase II slack. For living donor records, the mean score was 0.979, the median was 0.982, and the range was 0.926–1.000; two records were on the radial frontier and both were slack free [Fig. 1; Table 1].

Scores were more concentrated near the frontier among living donor records. Seven deceased donor records (3.1%) and 53 living donor records (26.9%) had scores of at least 0.990. The highest nonfrontier scores were 0.9995 and 0.9982, respectively [Table 1]. Because donor types were evaluated against separate frontiers, BCC scores represent relative position only within each donor group.

## 3.4 Agreement with alternative BCC specifications

Using the published outcome proportions without aggregate adjustment changed frontier membership substantially. Correlations with the primary BCC scores were 0.859 for deceased donor and 0.936 for living donor records. The unadjusted models identified 16 and 42 strong frontier records, respectively, compared with one and two in the primary models. Top-group overlap was 4 of 23 deceased donor records and 19 of 20 living donor records [Supplementary Table S2; Supplementary Figs. S3–S5].

Complete-case analyses included 162 deceased donor and 126 living donor records and remained highly correlated with the primary models ($\rho = 0.989$ and 0.980). Primary top-group overlap was 17 of 19 and 13 of 15 records, respectively. Winsorized models were similarly concordant ($\rho = 0.992$ and 0.994), with top-group overlap of 20 of 23 and 18 of 20 records [Supplementary Table S2; Supplementary Fig. S5].

Across the primary, unadjusted, complete-case, and winsorized specifications, median BCC-score ranges were 0.0047 for deceased donor and 0.0077 for living donor records, whereas median rank-percentile ranges were 0.089 and 0.107. Maximum rank-percentile ranges reached 0.560 and 0.607, respectively [Supplementary Table S2; Supplementary Figs. S6 and S7]. The intercept-only scale-invariance check reproduced the unadjusted rank ordering in both donor groups.

## 3.5 Sensitivity to frontier geometry and numerical resolution

BCC and FDH results were nearly identical for deceased donor records, with score correlation > 0.99999, identical rank ordering, frontier membership, and top-decile composition. For living donor records, BCC and FDH score and rank correlations were approximately 0.990; FDH identified three strong frontier records compared with two under BCC, retained both BCC strong frontier records, and preserved 17 of 20 BCC top-group records [Supplementary Table S2].

Rounding O/E values to four or three decimal places produced minimal changes. Score correlations with the primary analysis exceeded 0.9999 at four decimals and remained 0.9995 for deceased donor and 0.9994 for living donor records at three decimals. Radial and strong frontier membership was unchanged, and all primary top-group records were retained, although three-decimal rounding added one tied living donor record.

## 3.6 Stability under reference-set and full-pipeline analyses

All 2,000 planned subsampling repetitions were completed. When O/E values were held fixed and only the reference set changed, median 5th–95th percentile BCC-score range widths were 0.0005 for deceased donor and 0.0027 for living donor records; corresponding median rank-percentile widths were

0.035 and 0.045. Maximum rank-percentile widths were 0.347 and 0.244, respectively [Supplementary Figs. S8 and S9].

Rebuilding the full analysis pipeline widened these ranges despite high overall agreement with the reference-set-only analysis. Median paired Spearman correlations for BCC scores were 0.9976 for deceased donor and 0.9933 for living donor records. Median BCC-score range widths increased to 0.0051 in both groups, while median rank-percentile widths increased to 0.059 and 0.085; maximum rank-percentile widths reached 0.436 and 0.494, respectively [Fig. 2; Table 2].

Across paired repetitions, median absolute BCC-score differences were approximately 0.0009 in both donor groups, and median absolute rank-percentile differences were 0.010 for deceased donor and 0.011 for living donor records. Median top-group overlap was 95.2% and 88.9%, respectively. The primary deceased donor strong frontier record retained that status in all full-pipeline repetitions in which it was included, whereas the two primary living donor strong frontier records did so in approximately 93.3% and 98.5% of included repetitions.

# 4. Discussion

The adjusted BCC models produced selective frontiers, although many records, particularly living donor records, remained close to them. Using published outcome proportions without adjustment changed frontier membership substantially, whereas complete-case, winsorized, FDH, and numerical-resolution analyses preserved the broader score pattern. A second consistent finding was that score stability and rank stability were not equivalent: overall ordering remained highly correlated across analyses, but individual rank positions varied more when the full adjustment and DEA pipeline was rebuilt.

The contrast between unadjusted and adjusted models highlights the importance of outcome representation in DEA. One-year patient survival and graft function were concentrated near their upper limits, and direct use of the published proportions placed many records on the radial frontier, including some with positive slack. Expressing outcomes relative to expected values produced a more selective frontier, particularly for deceased donor records. This should not be interpreted as identifying a clinically true frontier, because the expected values were based on only three aggregate measures of recipient composition. Rather, the two specifications represent different benchmarking perspectives. The sensitivity of center classification to adjustment is consistent with previous kidney transplant research showing that risk-adjustment choices can materially alter assessments of center performance (Weinhandl et al. 2009).

Continuous BCC scores were generally more stable than ordinal rank positions. Across alternative specifications, most score differences were small, whereas rank percentiles showed greater variation, particularly where scores were densely clustered near the frontier. Reference-set-only subsampling produced relatively narrow stability ranges, while rebuilding the full pipeline widened both score and rank-percentile ranges despite very high correlations between paired results. Strong frontier records nevertheless remained highly persistent. These findings are consistent with previous work showing that

uncertainty in DEA-based provider quality measures can affect conclusions about high performance (Shwartz et al. 2016).

These results support using DEA as a screening and comparative review tool rather than as a source of definitive institutional rankings. Center position depends on the outcome representation, comparator set, and frontier specification, and small score differences can translate into appreciable changes in rank. Continuous scores, proximity to the frontier, and stability across alternative specifications should therefore be considered together. With the aggregate data used here, the resulting measures may help identify centers for contextual review or peer comparison but are not sufficient on their own to support punitive decisions, resource allocation, or public league tables.

# 5. Limitations

This study has several limitations. First, the analysis relied on publicly available aggregate data and could not reproduce patient-level clinical risk adjustment. The three covariates used in the primary models captured only limited variation in recipient composition and did not account for many donor, transplant, allocation, or center characteristics. In addition, activity and covariate data extended through November 2024, whereas outcome measures extended through November 2023. The resulting O/E indices should therefore be interpreted as limited aggregate adjustments rather than clinically risk-adjusted expectations derived from fully aligned transplant cohorts.

Second, aligned outcome denominators and direct measures of resource use were unavailable. Outcomes were therefore analyzed as reported proportions without volume weighting, and differences in sampling precision across centers could not be modeled. For the same reason, the DEA scores represent multidimensional quality benchmarks rather than measures of technical or operational efficiency. Missingness was also substantial for several candidate covariates. Median imputation with missingness indicators was used pragmatically, and although complete-case analyses produced broadly similar results, they do not establish that the imputation procedure or underlying missing-data mechanism was unbiased.

Finally, DEA results remain conditional on the selected outcomes, comparator set, weighting flexibility, and frontier specification. The subsampling analyses described empirical stability under changes in the reference set and rebuilding of the analytic pipeline but did not capture patient-level sampling uncertainty and should not be interpreted as formal confidence intervals. Because deceased and living donor records were evaluated against separate frontiers, their BCC scores are not directly comparable across donor groups. External validation using clinically richer and temporally aligned data would be required before this framework could support higher-stakes institutional or regulatory decisions.

# 6. Conclusions

This study developed a DEA framework for benchmarking publicly reported kidney transplant outcomes using aggregate data. Separate analyses for deceased and living donor transplantation combined limited

adjustment for recipient composition with joint assessment of patient survival and graft function. The overall score pattern remained broadly consistent across most sensitivity analyses, whereas frontier membership and rank position were more sensitive to model specification and to rebuilding the full analysis pipeline within subsamples. These findings support using the resulting measures for screening and comparative review rather than definitive institutional ranking. Clinically richer, temporally aligned data and external validation would be needed before higher stakes applications.

# Declarations

The author thanks the providers of the publicly available HRSA Organ Donation and Transplantation data resource archived through DataLumos/ICPSR.Acknowledgements

Funding

No funding was received for this study or for the preparation of the manuscript.

Competing interests

The author declares no relevant financial or non-financial competing interests.

Peer Review

I acknowledge participation in the Transplant Peer Review Network and complied with the journal's author guidelines and policies.

Ethics approval

This study used publicly available aggregate center level secondary data. No patient level records, identifiable personal information, protected health information, biological materials, or direct interactions with human participants were involved. Institutional ethics approval was therefore not required.

Clinical trial number: not applicable.

Data availability

The original data are publicly available through the HRSA Organ Donation and Transplantation data resource archived by DataLumos/ICPSR (doi:10.3886/E227006V2). The analysis ready donor type dataset, derived adjustment and DEA outputs, supplementary results workbook, supplementary figures, figure code mapping tables, and complete R analysis script are provided as supplementary materials.

Code availability

The complete R script used for data processing, aggregate covariate adjustment, DEA estimation, subsampling analyses, sensitivity analyses, validation, and figure generation is provided as supplementary material.

Generative AI use statement

During preparation of the manuscript, ChatGPT (OpenAI) and Claude (Anthropic) were used to assist with language editing, and review of portions of the R code. No AI tool was used to generate or alter study data or to make independent analytical decisions. All AI-assisted text and code suggestions were independently reviewed and verified by the author, who takes full responsibility for the content of the manuscript.

Author contributions

Feyza GURER: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, and Writing – review and editing.

# References


1. Banker RD, Charnes A, Cooper WW. Anal Manage Sci. 1984;30(9):1078–92. 10.1287/mnsc.30.9.1078. Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment.
2. Bogetoft P, Otto L. 2025. Benchmarking: Benchmark and Frontier Analysis Using DEA and SFA. R package version 0.33. 10.32614/CRAN.package.Benchmarking
3. Charnes A, Cooper WW, Rhodes E. Measuring the Efficiency of Decision Making Units. Eur J Oper Res. 1978;2(6):429–44. 10.1016/0377-2217(78)90138-8.
4. Dickinson DM, Shearon TH, O'Keefe J, Wong H-H, Berg CL, Rosendale JD, Delmonico FL, Webb RL, Wolfe RA. SRTR Center-Specific Reporting Tools: Posttransplant Outcomes. Am J Transplant. 2006;6(5 Pt 2):1198–211. 10.1111/j.1600-6143.2006.01275.x.
5. Hartman N, Messana JM, Kang J, Naik AS, Shearon TH, He K. Composite Scores for Transplant Center Evaluation: A New Individualized Empirical Null Method. Annals Appl Stat. 2024;18(1):729–48. 10.1214/23-AOAS1809.
6. Hollingsworth B. The Measurement of Efficiency and Productivity of Health Care Delivery. Health Econ. 2008;17(10):1107–28. 10.1002/hec.1391.
7. Kohl S, Schoenfelder J, Fügener A, Brunner JO. The Use of Data Envelopment Analysis (DEA) in Healthcare with a Focus on Hospitals. Health Care Manag Sci. 2019;22(2):245–86. 10.1007/s10729-018-9436-8.
8. Papke LE, Wooldridge JM. Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates. J Appl Econom. 1996;11(6):619–32. 10.1002/(SICI)1099-1255(199611)11:6%3C619::AID-JAE418%3E3.0.CO;2-1.
9. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2025.
10. Santos Arteaga FJ, Di Caprio D, Cucchiari D, Campistol JM, Oppenheimer F, Diekmann F, Revuelta I. Health Care Manag Sci. 2021;24(1):55–71. 10.1007/s10729-020-09516-2. Modeling Patients as

Decision Making Units: Evaluating the Efficiency of Kidney Transplantation Through Data Envelopment Analysis.

11. Schold JD, Buccini LD, Srinivas TR, Srinivas RT, Poggio ED, Flechner SM, Soria C, Segev DL, Fung J, Goldfarb DA. The Association of Center Performance Evaluations and Kidney Transplant Volume in the United States. Am J Transplant. 2013;13(1):67–75. 10.1111/j.1600-6143.2012.04345.x.
12. Schold JD, Andreoni KA, Chandraker AK, Gaston RS, Locke JE, Mathur AK, Pruett TL, Rana A, Ratner LE, Buccini LD. Expanding Clarity or Confusion? Volatility of the 5-Tier Ratings Assessing Quality of Transplant Centers in the United States. Am J Transplant. 2018;18(6):1494–501. 10.1111/ajt.14659.
13. Schold JD, Patzer RE, Pruett TL, Mohan S. Quality Metrics in Kidney Transplantation: Current Landscape, Trials and Tribulations, Lessons Learned, and a Call for Reform. Am J Kidney Dis. 2019;74(3):382–9. 10.1053/j.ajkd.2019.02.020.
14. Scientific Registry of Transplant Recipients (SRTR). 2026. Technical Methods for the Program-Specific Reports. Accessed August 18, 2026. https://srtr.hrsa.gov/transplant-professionals/program-specific-report/technical-methods-for-the-program-specific-reports/
15. Shwartz M, Burgess JF, Zhu J. Eur J Oper Res. 2016;253(2):489–502. 10.1016/j.ejor.2016.02.049. A DEA Based Composite Measure of Quality and Its Associated Data Uncertainty Interval for Health Care Provider Profiling and Pay-for-Performance.
16. Simar L, Wilson PW. Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models. Manage Sci. 1998;44(1):49–61. 10.1287/mnsc.44.1.49.
17. Simar L, Wilson PW. Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes. J Econ. 2007;136(1):31–64. 10.1016/j.jeconom.2005.07.009.
18. Siqueira MM, Araujo CAS. Anal Int J Health Plann Manage. 2018;33(4):e1067–87. 10.1002/hpm.2588. Efficiency of Brazilian Public Services of Kidney Transplantation: Benchmarking Brazilian States via Data Envelopment.
19. United States Department of Health and Human Services, Health Resources and Services Administration. 2025. Health Resources & Services Administration. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], April 21, 2025. Version V2. 10.3886/E227006V2
20. Wedderburn RWM. Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss–Newton Method. Biometrika. 1974;61(3):439–47. 10.1093/biomet/61.3.439.
21. Weinhandl ED, Snyder JJ, Israni AK, Kasiske BL. Effect of Comorbidity Adjustment on CMS Criteria for Kidney Transplant Center Performance. Am J Transplant. 2009;9(3):506–16. 10.1111/j.1600-6143.2008.02527.x.
22. Online. Resource 1 All Tables and all Figures.
23. Online Resource 2 Results File. Supplementary results workbook containing the processed analytic datasets, donor-type-specific aggregate-covariate adjustment outputs, primary and sensitivity DEA results, subsampling stability summaries, missingness diagnostics, validation outputs, and center-code mapping tables (kidney_dea_results.xlsx).

24. Online Resource 3 R analysis script. Complete R analysis script used for data processing, aggregate-covariate adjustment, DEA estimation, subsampling analyses, sensitivity analyses, validation, and figure generation (*kidney_dea.R*).

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineResource.docx