UNC CHARLOTTE ECONOMICS WORKING PAPER SERIES

CLINICIAN DECISION-MAKING IN DECEASED DONOR KIDNEY OFFERS AND THE IMPLICATIONS FOR A NATIONAL ACCEPTANCE MODEL

E. Glenn Dutcher
Ellen P. Green
Jesse D. Schold
Darren E. Stewart

Working Paper No. 2025-001

THE UNIVERSITY OF NORTH CAROLINA AT CHARLOTTE
BELK COLLEGE OF BUSINESS
DEPARTMENT OF ECONOMICS
9201 University City Blvd
Charlotte, NC 28223-0001
October, 2025

ABSTRACT

Case-mix–adjusted report cards summarizing kidney transplant program offer acceptance decisions aim to raise the quality of healthcare. The models used for risk adjustment, however, impose a one-size fits all, signal-to-choice mapping from clinical variables to acceptance, leading to potentially erroneous models of clinician behavior and ineffective organ allocation policies. We test this premise by comparing real-world deceased donor kidney acceptance decisions made by 59 clinicians (2016–2020; 9,000 decisions per clinician) with their decisions on hypothetical kidney offers in a controlled experiment. Some support for the signal-to-choice mapping is verified from the experiment, which reproduces heterogeneity in cross-clinician willingness-to-accept ranking for about half of the clinicians whose choices are better aligned with a model restricted to registry-observed factors (e.g., KDPI, donor age). For the remainder "off-model" clinicians, choices depart; however, they accepted higher-risk kidneys and achieved similar one-year graft survival, implying uniform case-mix adjustment can misincentivize donor-side risky transplants.

Clinician Decision-Making in Deceased Donor Kidney Offers and the Implications for a National Acceptance Model

E. Glenn Dutcher[a], Ellen P. Green[b], Jesse D. Schold[c,d], Darren E. Stewart[e]

[a] *Department of Economics, University of North Carolina Charlotte, Charlotte, North Carolina, USA*
[b] *College of Health Solutions, Arizona State University, Tempe, Arizona, USA*
[c] *Department of Surgery, University of Colorado–Anschutz, Aurora, Colorado, USA*
[d] *Department of Epidemiology, University of Colorado–Anschutz, Aurora, Colorado, USA*
[e] *Department of Surgery, NYU Grossman School of Medicine, New York, New York, USA*

Abstract

Case-mix–adjusted report cards summarizing kidney transplant program offer acceptance decisions aim to raise the quality of healthcare. The models used for risk adjustment, however, impose a one-size-fits-all, signal-to-choice mapping from clinical variables to acceptance, leading to potentially erroneous models of clinician behavior and ineffective organ allocation policies. We test this premise by comparing real-world deceased donor kidney acceptance decisions made by 59 clinicians (2016–2020; 9,000 decisions per clinician) with their decisions on hypothetical kidney offers in a controlled experiment. Some support for the signal-to-choice mapping is verified from the experiment, which reproduces heterogeneity in cross-clinician willingness-to-accept ranking for about half of the clinicians whose choices are better aligned with a model restricted to registry-observed factors (e.g., KDPI, donor age). For the remainder "off-model" clinicians, choices depart; however, they accepted higher-risk kidneys and achieved similar one-year graft survival, implying uniform case-mix adjustment can misincentivize donor-side risky transplants.

Keywords: medical decision making, surgeons, nephrologists, deceased donor kidney, hypothetical choices, risk and uncertainty, economic experiments

JEL Codes: C9 D10 I10 I110 I120

Acknowledgments: We appreciate feedback received from David Cooper, Kelly Vosters, and Bethany Lemont. We are also grateful for the participant feedback at the 2025 Cutting Edge of Transplantation Conference, the 2024 Advances in Field Experiments conference, the 2024 Carolina Regional Empirical Economics Day conference, and the 2025 European Society of Transplantation Congress. This would not have been possible without the research assistance of Alicia Lewis. The data reported here have been supplied by UNOS as the contractor for the Organ Procurement and Transplantation Network (OPTN). The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the OPTN or the U.S. Government.

Email addresses: glenn.dutcher@charlotte.edu (E. Glenn Dutcher), egreen7@asu.edu (Ellen P. Green), jesse.schold@cuanschutz.edu (Jesse D. Schold), Darren.Stewart@nyulangone.org (Darren E. Stewart)

We are grateful for support from the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under award number R21DK128773. This funding made it possible to collect and analyze the data used in this study. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

1. Introduction

Economic models of medical decision-making typically assume a stable mapping from clinical information to choices— i.e., that clinicians interpret the same medical factors in similar ways. In practice, decisions are made quickly and under uncertainty, and clinicians may weigh observed medical and contextual factors differently. Such heterogeneity can cause uniform models to mispredict behavior. The stakes are especially high in deceased-donor kidney (DDK) allocation: in 2024, 27,332 kidney transplants were performed in the United States, while 87,125 candidates remained on the waitlist.[1] When a DDK is offered, an on-call surgeon or nephrologist has no more than an hour to accept or decline. We ask whether clinicians interpret medical factors consistently and, by extension, whether reduced-form and experimental models that rely on those factors can predict real-world acceptance decisions—specifically, whether they recover the cross-clinician willingness-to-accept ranking observed in practice.

Regulators have operationalized the common-mapping assumption in the Scientific Registry of Transplant Recipients (SRTR) Expected Acceptance Probability (EAP) model, which compresses offer, donor, and recipient covariates into a predicted acceptance rate based on national decisions and is embedded in the risk-adjusted scorecard used to evaluate transplant centers (Kizer et al., 2022). The usefulness of this benchmark hinges on the mapping from covariates to choices being sufficiently similar across clinicians so that a common index can proxy for "effort." If surgeons differ in unobserved risk tolerance, institutional culture, or perceptions of donor quality, EAP may systematically misclassify both overly cautious and appropriately selective behavior, thereby distorting incentives to accept viable higher-risk kidneys.

Our empirical tests therefore evaluate whether reduced-form and experimental models built on EAP-style medical factors replicate the cross-clinician dispersion in willingness to accept observed in practice and, by implication, whether a one-size-fits-all benchmark is adequate for steering acceptance behavior.

We evaluate these claims using a clinician-identified, matched lab–field dataset. Our data comprises a unique clinician-level dataset that utilizes multiple years of real-world acceptance behavior linked to kidney offer choices from a controlled experiment. To our knowledge, we are the first to construct clinician-level data for DDK transplant offer decisions and the first to link the real-world data with experimental choices made by the same clinicians. To do so, we match Organ Procurement and Transplantation Network (OPTN) logs to on-call records at transplant centers. Clinicians in our study also made decisions in SimUNet^{SM}, an experimental interface that mirrors DonorNet - the decision tool

[1] OPTN, 2024 (U.S. Government, 2024).

used by clinicians to field real-world offers.[2] In total, we were able to utilize the real-world decisions of 59 clinicians to explore the relevance of EAP and 44 of those same clinicians in SimUNet^{SM}. These 44 clinicians logged approximately 440,000 DDK accept/decline decisions between 2016 – 2020 (mean 9,753 per clinician). This provides us with a rich within-clinician panel for testing whether clinicians consistently apply medical information across settings.

Our empirical strategy proceeds in two steps: we quantify how much clinician-level variation in acceptance is explained by EAP, then link choices in SimUNet^{SM} to real-world decisions, with emphasis on higher-risk offers. This yields three policy-relevant payoffs. First, we identify where EAP's predictions align with or diverge from observed choices, highlighting opportunities for improving benchmarking. Second, because high-KDPI kidneys account for nearly 60% of unused but transplantable organs (Wilk et al., 2017), even modest increases in their acceptance could generate sizable welfare gains through longer survival and better quality of life (Bae et al., 2019). Third, because SimUNet^{SM} uses only information available at the time of decision, we can isolate the association between observed medical factors and acceptance in a standardized environment. Taken together, these analyses evaluate whether a one-size-fits-all EAP index adequately captures systematic behavior and whether additional data are needed for high-risk offers.

We find that the EAP explains little of the variance for high-KDPI offers on average, and there is no strong link between real-world and experimental choices. However, the averages hide significant heterogeneity in how clinicians respond to the same medical information. For roughly half of clinicians, experimental and real-world decisions align closely, suggesting that simplified models, such as EAP, based solely on medical factors, can capture their decision rules. The lack of explanatory power of the simplified models for the remainder does not seem to come at the expense of short-term patient outcomes, as remarkably, these "off-model" clinicians accept a larger share of high-KDPI kidneys yet achieve one-year graft-survival outcomes comparable to their peers, suggesting that the current EAP benchmark may underrate practices that safely expand access.

These insights complement recent work on market design and assignment mechanisms (Agarwal et al., 2020, 2021), suggesting that variation in *interpretation* of medical factors, not just incentives, shapes organ allocation outcomes. Previous work has also shown physician-level heterogeneity (Molitor, 2018; Cutler et al., 2019; Badinski et al., 2023; Finkelstein et al., 2016; Green et al., 2025); however, we are the first to show significant differences in how clinicians respond to identical medical factors when making acceptance decisions, suggesting that clinician choices do not neatly align with a single, universal mapping from clinical inputs to final decisions (Mullainathan and Obermeyer, 2022). Instead, clinicians appear to differ systematically in their decision rules, potentially reflecting deeper heterogeneity in altruistic preferences (Li et al., 2022), beliefs about risk (Brosig-Koch et al., 2024), training backgrounds (Molitor, 2018), or institutional incentives (Chan and Roth, 2024).

The clinician-specific patterns identified in this study are generally useful for economic modeling and policy. They contribute to the broader economic literature documenting substantial individual variation

[2] Stewart et al. (2020) previously used SimUNet^{SM} to understand the effect of biopsies on acceptance. McCulloh et al. (2023) presents core clinical variables that enter the EAP model. For example, donor Kidney Donor Profile Index (KDPI), donor age and cause of death, cold-ischemia time, human-leukocyte-antigen (HLA) mismatch, blood-type compatibility, and recipient wait-time priority.

in economic preferences (Einav et al., 2012; Dohmen et al., 2011; Charness et al., 2013) and complicate one-size-fits-all allocation algorithms. The key question, therefore, is whether the observed dispersion reflects stable traits (e.g., risk tolerance) or unmodeled features of the decision environment (e.g., how medical information is interpreted). Clarifying that distinction is essential for welfare analysis, because policy levers differ depending on the underlying source of heterogeneity (Kapor et al., 2016).

2. Methodology and Procedures

We used two sources of data for the same group of clinicians: (1) a retrospective dataset capturing their actual kidney offer acceptance decisions in the field, linked with on-call logs from 18 transplant centers to identify clinician-specific decisions, and (2) an experimental dataset collected via SimUNet^{SM}, a platform designed to conduct experiments with hypothetical organ offers that replicate the information, structure, and randomness of DonorNet—the interface where clinicians make real-world decisions to accept or reject a deceased donor kidney. This dual-dataset approach allows us to directly compare clinicians' real-world behavior with their responses in a controlled experimental environment.

2.1. Retrospective Data:

This study used data from the Organ Procurement and Transplantation Network (OPTN). The OPTN data system includes data on all donor, wait-listed candidates, and transplant recipients in the US, submitted by the members of the Organ Procurement and Transplantation Network. The Health Resources and Services Administration (HRSA), U.S. Department of Health and Human Services provides oversight to the activities of the OPTN contractor.

We obtained offer-level data from the OPTN for a set of U.S. transplant centers spanning January 1, 2016, to January 1, 2021. These data include all deceased donor kidney offers made to each center, along with donor and candidate characteristics, the offer's time and date, and the acceptance decision (if any). To assign decisions to individual clinicians, we merged OPTN offer records with on-call schedules from participating centers, creating a clinician-level panel of real-world decision-making. Our sample includes multiple centers selected to reflect variation in geographic location, size, and aggressiveness in accepting kidneys (as measured by observed-to-expected acceptance ratios; see Green et al. (2025)). This results in thousands of decisions per clinician, allowing us to estimate baseline acceptance behavior under actual clinical conditions.

Within the OPTN dataset, we focus on the accept/reject decision for the first offer matched to a given candidate-donor pair (since multiple match runs are used to allocate kidneys from some donors). We exclude observations from donors where no acceptance was recorded at any point in the OPTN dataset, bypassed candidates, as well as offers presented after the final accepted offer. This ensures a consistent measure of acceptance behavior across both settings. These exclusions were in place because factors could have been uncovered throughout the offer process and not recorded, which would imply the offer was not usable and should not have been considered as a viable offer. Thus, our focus on offers that were eventually accepted and transplanted puts the focus on offers that went through a rigorous screening process and were deemed by multiple medical professionals - the organ procurement team and the transplant team - to be viable.

2.2. Experimental Data (SimUNet^{SM}):

From October 2021 to September 2023, we recruited clinicians from 18 participating transplant centers to participate in SimUNet^{SM} experiments.[3] SimUNet^{SM}, developed by UNOS Labs, replicates the informational environment of DonorNet and thus mirrors the real-world decision-making platform. Each participant received the same ten hypothetical offers over five consecutive days, with two offers delivered per day at randomized times. The hypothetical offers were drawn from actual deceased donors, ensuring realism, and varied systematically in quality. To prevent order effects, the sequence of offers was randomized across participants. Clinicians were asked whether or not they believe they would, in a real clinical setting, accept or reject each offer on behalf of a specific hypothetical patient profile, and the entire SimUNet^{SM} exercise required approximately 30-45 minutes of total engagement.[4]

By combining risk-adjusted retrospective data with carefully controlled experimental data, and by leveraging randomization in the timing and sequence of experimental offers, we are able to isolate the extent to which simplified representations of the decision environment (via SimUNet^{SM}) align with actual clinical practice.

2.3. Variable Descriptions

Our main outcome variable is a binary accept/reject decision. To measure the perceived riskiness of an offer, we employ three commonly used metrics. The first is focused on medical factors tied to the donor. The second considers medical factors tied to the potential recipient. The third is a predictive tool that combines these factors, as well as other offer characteristics.

The Kidney Donor Profile Index (KDPI), ranging from 0.0 to 1 and derived from the Kidney Donor Risk Index (Rao et al., 2009), provides an estimate of donor kidney quality, with lower values indicating longer expected graft survival.[5] KDPI is derived based on a normative national reference that is updated annually and is assigned to all recovered kidneys. A KDPI of 0.5, for instance, indicates that this DDK is expected to last as long as 50% of all kidneys recovered in the last year. A KDPI of 0.95 indicates 95% of kidneys in the last year are expected to last longer than the current offer. Given that a kidney with a higher KDPI is riskier, all else equal, we expect more cautious clinicians to reject higher KDPI offers more frequently.

The expected post-transplant survival (EPTS) score is a measure of the candidate's health.[6] The higher the candidate's EPTS score, the lower the prospective recipient's expected post-transplant longevity. EPTS thus signals a candidate's health and need for an organ and is used in policy to match higher-quality offers (KDPI $\leq$ 20%) with candidates who are expected to have a better chance to realize the offer's potential (EPTS $\leq$ 20%). It is not clear a priori what the predicted direction of

[3] This study was part of a larger study on clinical decision-making where the SimUNet^{SM} choices were always made first. We could not incentivize the choices made in SimUNet^{SM}, but clinicians were given a payment for their participation in addition to a small payment received from a portion of the study not related to SimUNet^{SM}.

[4] Clinicians received a fixed fee for participating in the study, but no marginal incentives were given for their hypothetical choices.

[5] KDPI is sometimes listed as a percent, such as 25% or 50%. The data records this in the range from 0.0 to 1, which is how we will describe it in the text. For a more detailed description, see https://optn.transplant.hrsa.gov/professionals/by-topic/guidance/kidney-donor-profile-index-kdpi-guide-for-clinicians/

[6] More information on this score can be found here: https://unos.org/news/in-focus/what-is-epts/

this effect will be. On the one hand, clinicians may deem those with a higher score as having more need and thus more willing to transplant them. On the other hand, a higher score may imply more frailty, which could lead to graft failure or death.

Finally, EAP is an empirically derived metric that incorporates 42 donor, candidate, and match characteristics to predict acceptance likelihood based on national decision-making patterns.[7] EAP approximates the broader informational structure clinicians encounter in SimUNet^{SM} and DonorNet by reducing dimensionality from 42 elements to a single likelihood of acceptance scale. Because EAP incorporates a broad set of patient, donor, and contextual factors, it provides a benchmark for whether clinicians follow commonly observed national patterns of risk acceptance.

3. Hypotheses

A central objective of this study is to gain a deeper understanding of how clinicians use commonly available information in their choices. We will accomplish this by first examining the choices from the retrospective data and second by investigating the link between choices in the hypothetical and retrospective data.

The widely used EAP model serves as a baseline for understanding the value of available medical information. We will focus on EAP's ability to capture the variance in choices via psuedo-R^2 from a simple logit regression. A low R^2 is typically not a concern, as meaningful causal effects can still be identified in the regression. We are not as interested in the causal effects of a single variable, but rather in how well EAP captures the variables clinicians use when making choices. A low R^2 indicates that EAP is missing important explanatory variables, or is not combining the current variables in a way that captures variation. Although there is no universally agreed-upon threshold for R^2, some suggest values as low as .10 and as high as .30 are the minimum for modeling human behavior (Gupta et al., 2024). Much higher values are expected for predictive models. In our first hypothesis, we chose the middle value of 0.20; however, we note that this value has no special statistical properties. Nevertheless, it is more conservative than one might expect for models intended to provide predictions.

Hypothesis 1: *The psuedo-R^2 from a simple logit regression of EAP on acceptance will be above 0.20.*

The first hypothesis could fail because the EAP model is not parameterized correctly, or because clinicians do not generally use the observable factors to the extent believed. To pin down the influence of observable factors on clinicians' willingness to accept, we assess whether a controlled experimental environment can approximate clinicians' decision-making patterns. Our experimental setting (SimUNet^{SM}) abstracts away certain complexities—such as scheduling operations or assembling surgical teams—and focuses instead on the core characteristics commonly used in predictive frameworks like the EAP model. A key question is whether the resulting hypothetical decisions correlate with those observed in actual clinical practice.

Hypothesis 2: *Clinicians who accept more offers in SimUNet^{SM} should also be more likely to ac-*

[7] More information can be found at the following website: https://www.srtr.org/tools/offer-acceptance

cept higher-risk offers in the real world.

Note that this hypothesis does not imply identical acceptance rates across settings. Without the logistical frictions and ethical complexities of the real world, clinicians may appear more willing to accept riskier offers in the hypothetical scenario. Our test is thus whether those who appear more accepting in the experiment also tend to be relatively more accepting in the field, indicating that essential decision-making factors are similarly weighted in both environments.

If we fail to find support for the first two hypotheses, then a reasonable case can be made that clinicians are not as reliant on the observable factors as believed when making choices for the higher-risk DDKs. However, not all clinicians may rely on these standard clinical factors to the same extent. Evidence suggests substantial heterogeneity in clinician behavior (Green et al., 2025). Some clinicians may adhere closely to the attributes captured by EAP, while others incorporate additional, unmeasured considerations.

Hypothesis 3: *Clinicians whose real-world acceptance behavior is more closely predicted by EAP—i.e., those who rely more systematically on the standard medical and donor characteristics—will exhibit stronger correlations between their experimental and actual decisions.*

Put differently, the more a clinician's real-world choices align with the EAP model's informational structure, the better the experimental environment should align with their field behavior. If the hypothetical setting omits essential elements in the decision-making process, then a link between the hypothetical and real-world choices will not exist. The second hypothesis tests the relevance of those omitted factors in the decision-making process for the average clinician, while the third allows for a specific form of heterogeneity under the assumption that the link will only exist for a subset of clinicians.

Offer	KDPI	EPTS	No. of Choices	Percent of offers accepted	Used in Analysis
1	0.14	0.76	66	100%	Omitted
2	0.22	0.84	57	100%	Omitted
3	0.71	0.58	64	100%	Omitted
4	0.78	0.02	61	73.21%	Included
5	0.81	0.24	66	83.33%	Included
6	0.81	0.24	63	73.02%	Included
7	0.89	0.13	66	78.79%	Included
8	0.89	0.23	66	30.30%	Included
9	0.95	0.20	63	12.70%	Included
10	0.98	0.67	66	59.09%	Included

Table 1: Summary of SimUNet^{SM} offers

4. Results

4.1. Summary of the Data

We begin by describing the experimental (hypothetical) data. A total of 73 clinicians participated in the SimUNet^{SM} study, each receiving a series of hypothetical offers drawn from real donor cases. Table 1 summarizes acceptance rates across the ten hypothetical offers. We intentionally chose high KDPI offers to focus on how decisions were made for riskier DDKs. The first three offers, all with relatively low KDPI values, were universally accepted, providing no variation for analysis. This acceptance pattern indicates that the clinicians took the task seriously, which can be seen as an attention check. Therefore, we focus the subsequent analysis on offers 4–10, each of which had a KDPI > 77%. Limiting the sample in this way results in 70 clinicians making at least one decision and 63 clinicians completing at least five out of seven choices. This table also provides the EPTS for the offers, demonstrating variation in both the KDPI and EPTS. Figure 1 shows the frequency of the average acceptance rate per clinician for the 63 clinicians, which ranges from 0 if they did not accept any offers to 1 if they accepted all offers. The distribution skews toward higher acceptance counts, reflecting a greater willingness to accept risk in the frictionless hypothetical environment as anticipated.

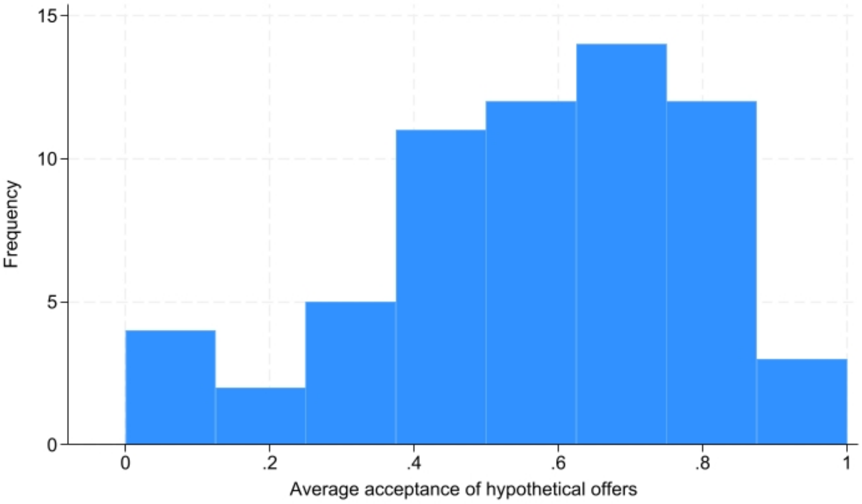

Figure 1: Distribution of the mean number of SimUNet^{SM} offers accepted out of 10, restricted to clinicians who made at least 5 choices.

Turning to the retrospective data, 69 clinicians at participating centers received over 670,000 candidate-specific offers, with roughly 5,900 accepted. Not all clinicians represented in the retrospective data participated in SimUNet^{SM}, and vice versa.[8] Because a key area of interest is in understanding the role of EAP in explaining choices, Figure 2 displays the pseudo-R^2 of logit regressions for different values of KDPI where the dependent variable is the accept/reject decision and the explanatory variable is the calculated EAP for that offer. From the Figure, it is seen that as KDPI increases, the amount of variance explained by EAP decreases, though these averages hide substantial heterogeneity between clinicians. When we conduct clinician-level regressions conditioned on high-risk offers (KDPI > 0.77), the pseudo-R^2 values range from as low as 0.000 to as high as 0.674. These results suggest that EAP may not capture the relevant variables used in the decision-making process on average, and the way

[8] Table 7 in the Appendix provides an overview of the 18 participating centers, including the number of clinicians and the volume of deceased donor kidney offers they collectively evaluated from 2016 to 2021. Of the 69, we are only able to use 59 for this analysis, as we do not have enough data for clinicians who are residents or newly appointed at their transplant center.

clinicians use that information is heterogeneous. We return to this heterogeneity in a subsequent section.

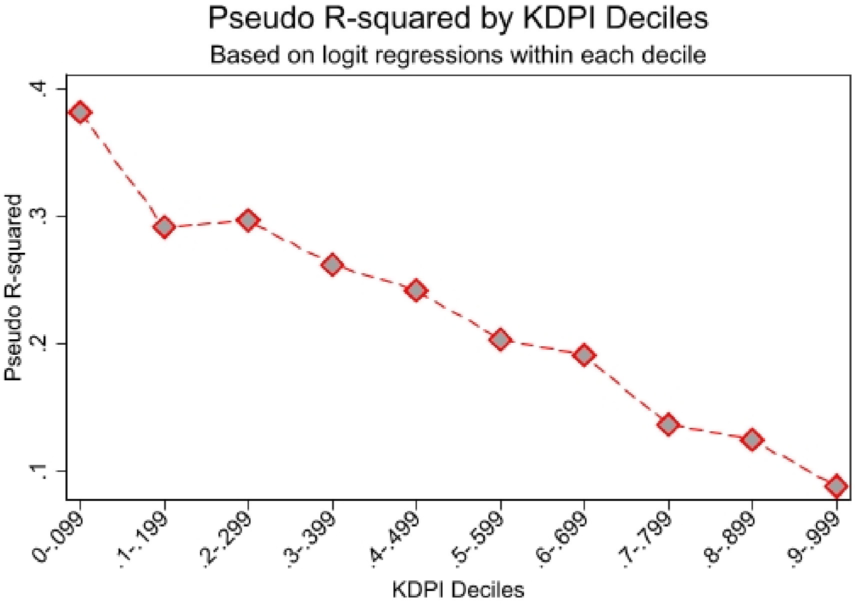

Figure 2: The Psuedo-R^2 was calculated for each decile from logit regressions where the acceptance decision was the dependent variable and EAP was the sole explanatory variable.

Testing our first hypothesis, the R^2 from a simple logit regression that uses EAP as the only explanatory variable for all choices is 0.266, but falls to 0.111 for choices where KDPI > 0.77. For the riskiest offers, which comprise those offers that are more likely to go unused (Wilk et al., 2017), we do not find support for the first hypothesis.

For the analysis linking experimental and real-world decisions, we focus on the 44 clinicians present in both datasets, who together made over 400,000 real-world decisions. To get a general sense of how medical factors influence choices, Table 2 presents a logistic regression analysis of hypothetical and retrospective acceptance decisions. The model includes KDPI, representing donor kidney quality, and EPTS, which reflects the urgency of the recipient's need for transplantation. To ensure comparability with the hypothetical data, we restrict the analysis to offers with KDPI > 0.77. The regression uses clustered standard errors at the clinician level.

The results reveal key similarities and differences between the retrospective and hypothetical settings. Similar to the hypothetical data, the coefficient on EPTS is positive and statistically significant. Also, the coefficient on KDPI is negative in both; however, it is statistically insignificant in the retrospective data ($p = 0.69$), while in the hypothetical setting, higher KDPI values were strongly associated with reduced acceptance rates. This lack of correlation between KDPI and acceptance behavior in the retrospective data is unexpected and suggests that for high-risk offers, clinicians may not consistently rely as much on this specific medical factor of donor quality as a decisive factor.

Table 2: Regression Results for Offer Acceptance Decisions

	Offer Acceptance	
	Hypothetical	Retrospective
KDPI	-23.606^{***}	-0.335
	(2.78)	(0.98)
EPTS	6.356^{***}	0.445^{***}
	(0.92)	(4.52)
Constant	19.527^{***}	-5.365^{***}
	(2.29)	(16.79)
Observations	451	291,433

Notes: Logistic regression estimates of clinician acceptance decisions based on hypothetical and retrospective (real-world) data. KDPI (Kidney Donor Profile Index) reflects donor organ quality; EPTS (Estimated Post-Transplant Survival) captures recipient benefit. Analyses are restricted to kidneys with KDPI > 0.77. $\ln \sigma_u^2$ is the log of the variance of the random effect. Standard errors are clustered at the clinician level. t-statistics in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4.2. Testing Hypothesis 2: Alignment Between Hypothetical and Real-World Decisions

To formally test Hypothesis 2—that clinicians who are more accepting in the hypothetical environment are also relatively more accepting in the real world—we link their SimUNet^{SM} acceptance patterns to their field decisions. Specifically, we regress their actual acceptance decisions on the ratio of offers they accepted in SimUNet^{SM}. Because all participants in SimUNet^{SM} faced the same set of hypothetical offers, this ratio captures their relative willingness to accept riskier organs. If clinicians' core decision-making criteria are similar across contexts, we expect a positive relationship between experimental and actual behavior.

Table 3 presents the results. The first specification includes only the ratio of hypothetical acceptances, while the second adds a control for EAP to account for variation in offer characteristics. Both specifications indicate a positive but statistically insignificant relationship between the propensity to accept offers in SimUNet^{SM} and actual acceptance decisions. The wide confidence intervals indicates that some clinicians may rely heavily on the core medical factors reflected in the hypothetical environment, while others incorporate additional, unobserved considerations in their real-world decisions.

The results suggest that not all clinicians' acceptance behavior aligns well with the hypothetical environment, presumably because some rely on unobserved or additional factors in practice. To examine whether clinicians who do rely on these medical factors show more similarity between hypothetical and real-world decisions, we turn to Hypothesis 3.

4.3. Testing Hypothesis 3: Heterogeneity in Model Predictability

Hypothesis 3 states that clinicians whose real-world acceptance decisions are well-explained by the EAP model (i.e., those more reliant on medical attributes) will exhibit stronger consistency between hypothetical and actual behaviors. If, on the other hand, a clinician prioritizes factors beyond offer, donor, and candidate characteristics, then the link between these two settings should be weaker.

Table 3: Logistic Regression of Retrospective Acceptance on Hypothetical Decisions and EAP

	Retrospective Acceptance	
	(I)	(II)
SimUNet^{TM} Acceptance Rate	0.719	0.982
	(0.79)	(1.17)
EAP		28.990***
		(12.83)
Constant	-5.699^{***}	-6.086^{***}
	(-9.06)	(-9.98)
Observations	151,460	100,357

Notes: Logistic regression estimates of real-world (retrospective) kidney offer acceptance decisions. Column (I) includes only the clinician's acceptance rate in the hypothetical (SimUNet) setting. Column (II) additionally includes the Expected Acceptance Probability (EAP), a simulation-derived measure of clinician behavior. Standard errors are clustered at the clinician level. z-statistics in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

To assess how each clinician utilizes the available information, we measure the extent to which EAP explains the variance in their real-world acceptance decisions by recording clinician-specific R^2 values. As previously mentioned, there is a great deal of heterogeneity in this measure.

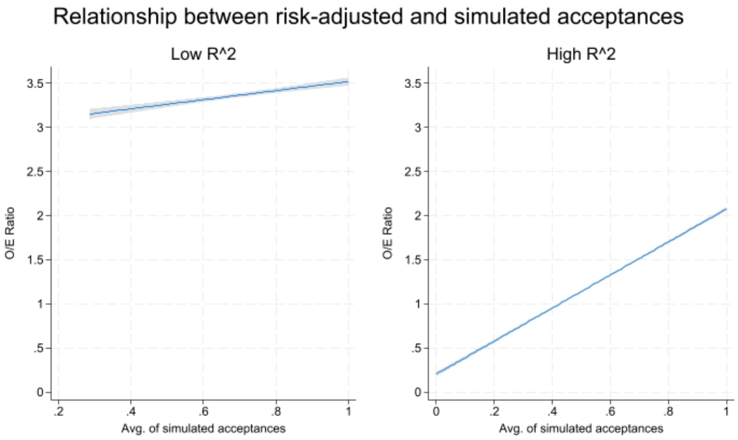

Figure 3: Unadjusted simple correlation of the simulated and real-world data.

To test whether the heterogeneity in EAP's predictive power relates to the observed correlation between hypothetical and real-world decisions, we divide clinicians into two groups based on a pseudo-R^2 cutoff of 0.19.[9] Specifically, those with a higher pseudo-R^2 (20 clinicians) and those with a lower pseudo-R^2 (24 clinicians). Similar results hold if we use three tiers (high, mid, low).[10]

Figure 3 provides an initial look at the relationship between the simulated (hypothetical) and risk-

[9] This cutoff is close to the 0.20 used in the second hypothesis, but was not drawn from the intuition from the literature. Instead, this cutoff was guided by the analysis, which determined the point at which there was no longer a statistically significant relationship between acceptance behavior in the two settings.

[10] Table 6 in the Appendix divides the sample into thirds. The positive and significant effect on hypothetical acceptance remains for the high-R^2 group and becomes negative (though not statistically significant) for the low-R^2 group.

adjusted real-world outcomes for these two R^2 groups. Among clinicians whose real-world decisions are better explained by EAP (high R^2), we observe a stronger correlation between hypothetical and actual acceptance choices.

We now formally test the third hypothesis.

Table 4: Robustness of Retrospective Acceptance Models by Fit Quality

	Higher R^2 Group		Lower R^2 Group	
	W.C. Bootstrap	RE Jackknife	W.C. Bootstrap	RE Jackknife
Hypothetical Acceptance	2.783**	3.589*	−0.415	−1.496
	(0.998)	(1.297)	(1.569)	(2.275)
EAP	33.55***	35.23***	33.57***	31.10***
	(2.904)	(4.072)	(2.293)	(2.743)
Constant	−7.743***	−8.707***	−5.033***	−3.966
	(0.603)	(0.760)	(1.318)	(1.843)
z-statistic (diff. in hyp. accept)	1.949*		−0.272	
# of clinicians	20	20	24	24
Observations	51,961	51,961	34,034	34,034

Notes: Table reports robustness of retrospective offer acceptance regressions across two clinician groups: those whose the EAP had higher vs. lower predictive fit (R^2). Estimates are shown using wild-cluster bootstrapped standard errors (W.C. Bootstrap) and random effects jackknife (RE Jackknife). The main predictors are the rate of acceptance in the hypothetical setting and EAP (Expected Acceptance Probability). A z-statistic tests the difference in coefficients on hypothetical acceptance across groups.
Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 4 presents regression results for these two groups (higher vs. lower R^2). The dependent variable in all regressions is the binary acceptance decision in the retrospective (real-world) data. The key independent variable is Hypothetical Acceptance, measured as the average number of SimUNet^{SM} offers a clinician accepted. Columns (1) and (3) report estimates using wild-cluster bootstrapped standard errors (shown in parentheses), while columns (2) and (4) use a jackknifed random-effects estimator for robustness. We will return to the rationale for these two specifications shortly. The focus now is on the row labeled *z-statistic (diff. in hyp. accept)*, which is a basic test of whether the coefficient on Hypothetical Acceptance differs significantly between the higher and lower R^2 groups.

For the group with higher R^2 values, we find a positive and statistically significant association between hypothetical and actual choices. By contrast, for the lower R^2 group, there is no significant relationship.[11] These results indicate that the simplified experimental setting provides a closer approximation to real-world decision-making among clinicians whose behavior aligns more closely with the EAP model. In other words, in support of our third hypothesis, the strength of the link between hypothetical and actual acceptance decisions depends critically on how closely a clinician's real-world behavior aligns with the factors emphasized by EAP.

4.3.1. Robustness Checks

Our robustness checks address two potential concerns. First, standard clustering methods may not satisfy the asymptotic assumptions due to limited clusters. Using wild-cluster bootstrapping, we

[11] Table 5 in the Appendix shows that excluding EAP as a control variable yields similar conclusions.

find that our main conclusions from the prior subsection remain. Second, a single clinician's behavior could disproportionately influence the results. Jackknifed random effects estimates show that no single clinician drives the observed relationships.

5. Discussion

The results thus far have been somewhat unexpected. In this section, we briefly explore the relationship between these results and other key metrics in kidney allocation: risks taken and graft survival. One distinguishing feature of clinicians who deviate from the EAP's predictions - those with the lowest R^2 - is that they tend to accept more offers - average number of acceptances is 153 vs. 107 -, and accept riskier offers - O/E ratio is 2.08 vs. 1.20. That is, those in the low R^2 group take roughly 73% more risk than those in the high R^2 group. In a simple logit regression on acceptance, the coefficient estimate on pseudo-R^2 is negative ($p < 0.001$), and remains negative and statistically significant when a control for EAP is added ($p < 0.01$), implying that acceptance is lower for those more likely to adhere to the predictions from EAP.[12] Likewise, in a simple OLS, there is a statistically significant negative relationship between pseudo-R^2 and the O/E ratio - a metric that measures how much risk a clinician takes relative to the expected risk.[13] Importantly, these additional risks do not appear to produce worse one-year graft survival rates, suggesting that a deeper understanding of how these clinicians assess organ quality could inform policies aimed at expanding transplant access without compromising outcomes.[14]

Moreover, our finding that clinicians with similar one-year post-transplant outcomes use medical data differently raises a fundamental question: Should clinicians rely on standard medical models for high-risk offers? Pruett et al. (2021) finds little difference in survival for transplants using KDPI below and above 0.85, indicating high KDPI offers may not have substantially worse outcomes.[15] Additionally, Bae et al. (2019), whose machine learning approach was designed to capture complex nonlinearities and interactions between EPTS and KDPI, shows minimal risk-ratio increases as EPTS rises for very high KDPI offers.[16] These findings suggest that current data for studying high KDPI offers may be insufficient, implying a fundamental issue with risk-adjusted metrics used to encourage more high-risk offer acceptance, and current models that rely on this data for predictions.

6. Conclusion

This study examines whether simplified models of decision-making can predict the complex real-world behavior of clinicians choosing to accept deceased donor kidneys. Leveraging a unique dataset that tracks the same clinicians over time and across both a controlled experimental platform (SimUNet^{SM}) and their actual acceptance decisions (DonorNet), we focus on high-risk, lower-quality offers—precisely

[12] However, when a fixed effect for listing center is included, the statistical significance goes away, suggesting that either clinicians who select into different practices differ on their comfort with accepting risk, or center-level characteristics are tied to risk-taking. With the current dataset, we do not have a way to distinguish between the two, suggesting more can be done to gather this kind of relevant information.

[13] All regressions include clustering at the clinician level.

[14] In a logit model regressing one-year graft failure rate on pseudo-R^2 with errors clustered at the clinician level (with or without controls for EAP) the coefficient estimate on pseudo-R^2 is never statistically significant ($p > 0.76$).

[15] See Figure 3

[16] See Figure 2

the margin where improvements could meaningfully reduce the discard rate and potentially increase the availability of kidneys for transplantation. By using individual-level data, our findings offer practical insights into how policy interventions might affect clinicians' acceptance of these high-impact offers.

Our analysis shows that while a simplified setting captures some essential elements of clinician decision-making, it does not uniformly predict all clinicians' choices in practice. In fact, the most widely used model, the EAP framework, does a poor job of capturing the relevant components for many of the clinicians in our study. It is for these same clinicians that we fail to find a correlation between hypothetical and real-world choices. Further, the experimental environment approximates real-world decisions more closely for those clinicians whose actual acceptance patterns align well with the standard informational factors captured in the EAP model.

Our findings emphasize two key takeaways. First, even though an empirical model like EAP captures a nationally averaged notion of *risk* over all kidney offers, it fails to predict perceived risk for high-risk offers on average. Further, our results suggest that clinicians may hold different views about which factors best capture an organ's overall desirability. Heterogeneity in how much of the variation in acceptance behavior is explained by this model, as reflected in our measures of R^2, implies that not all clinicians agree on what constitutes "risky" or "desirable." Policies that hinge on equating risk with common medical metrics assume a uniform risk across clinicians, potentially weakening their effectiveness. Consistent with this concern, we find that the clinicians whose acceptance decisions are poorly explained by the EAP accept a larger share of high-KDPI kidneys yet achieve one-year graft-survival rates comparable to their peers, indicating that the benchmark can undervalue decision rules that safely expand access. If policymakers seek to encourage greater use of higher-KDPI kidneys, learning from these more risk-tolerant clinicians—and adapting policies to accommodate their decision-making criteria—may be especially valuable.

Second, these results contribute to ongoing discussions in experimental economics regarding the external validity of laboratory experiments, particularly the conditions under which experimentally derived predictions generalize to complex, high-stakes settings. Prior studies highlight that laboratory findings may not consistently translate to field contexts due to differences in stakes, scrutiny, and contextual influences (Levitt and List, 2007; Al-Ubaydli et al., 2017). Yet, other work has shown that experimental insights often generalize qualitatively, especially when carefully designed experiments capture essential behavioral mechanisms (Kessler and Vesterlund, 2015; Galizzi and Navarro-Martinez, 2019; Athey et al., 2025; Cox et al., 2016). In our setting, we show that the experimental model aligned with external behavior *only* when the modeling assumption of the experiment – that clinicians' use of medical information in a predictable manner – was satisfied. That the assumptions failed for some, but not for others, indicates that the degree to which we expect externally valid results rests on an assumption of a homogeneous decision-making process and that homogeneity is as assumed in the model the experiment is built around. Simply relying on average behavior as the yardstick to measure the two settings may miss the bigger issue that there is not a single underlying latent decision process, and dismissing the experiment/model on these grounds implies discarding important information on how decisions are made. Our study was well-suited to pick up on this kind of heterogeneity, given our access to EAP data. The challenge in other studies is to find a reliable external measure to assess better the degree to which the model works. By identifying when and for whom experimental models, or predictive models like the EAP, successfully predict real-world clinician behavior, we address key calls in the experimental economics literature to better understand the boundaries and generalizability of lab-based insights in a health context (Falk and Heckman, 2009; Al-Ubaydli et al., 2017).

Future research should investigate what additional considerations drive the behavior of clinicians whose actions are not well-predicted by standard measures. Understanding whether these differences stem from institutional incentives, nuanced patient-level information, or ethical principles not captured in existing models can guide the refinement of both experimental designs and policy approaches. More generally, our results suggest that a nuanced, clinician-level perspective is crucial for developing policies that enhance efficiency and equity in organ allocation.

References

Agarwal, N., Ashlagi, I., Rees, M. A., Somaini, P., and Waldinger, D. (2021). Equilibrium allocations under alternative waitlist designs: Evidence from deceased donor kidneys. *Econometrica*, 89(1):37–76.

Agarwal, N., Hodgson, C., and Somaini, P. (2020). Choices and outcomes in assignment mechanisms: The allocation of deceased donor kidneys. Technical report, National Bureau of Economic Research.

Al-Ubaydli, O., List, J. A., and Suskind, D. L. (2017). What can we learn from experiments? understanding the threats to the scalability of experimental results. *American Economic Review*, 107(5):282–286.

Athey, S., Chetty, R., and Imbens, G. (2025). The experimental selection correction estimator: Using experiments to remove biases in observational estimates. Technical report, National Bureau of Economic Research.

Badinski, I., Finkelstein, A., Gentzkow, M., and Hull, P. (2023). Geographic variation in healthcare utilization: The role of physicians. Technical report, National Bureau of Economic Research.

Bae, S., Massie, A. B., Thomas, A. G., Bahn, G., Luo, X., Jackson, K. R., Ottmann, S. E., Brennan, D. C., Desai, N. M., Coresh, J., et al. (2019). Who can tolerate a marginal kidney? predicting survival after deceased donor kidney transplant by donor–recipient combination. *American Journal of Transplantation*, 19(2):425–433.

Brosig-Koch, J., Hennig-Schmidt, H., Kairies-Schwarz, N., Kokot, J., and Wiesen, D. (2024). A new look at physicians' responses to financial incentives: Quality of care, practice characteristics, and motivations. *Journal of Health Economics*, 94:102862.

Chan, A. and Roth, A. E. (2024). Regulation of organ transplantation and procurement: A market-design lab experiment. *Journal of Political Economy*, 132(11):3827–3866.

Charness, G., Gneezy, U., and Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization*, 87:43–51.

Cox, J. C., Green, E. P., and Hennig-Schmidt, H. (2016). Experimental and behavioral economics of healthcare.

Cutler, D., Skinner, J. S., Stern, A. D., and Wennberg, D. (2019). Physician beliefs and patient preferences: a new look at regional variation in health care spending. *American Economic Journal: Economic Policy*, 11(1):192–221.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550.

Einav, L., Finkelstein, A., Pascu, I., and Cullen, M. R. (2012). How general are risk preferences? choices under uncertainty in different domains. *American Economic Review*, 102(6):2606–2638.

Falk, A. and Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, 326(5952):535–538.

Finkelstein, A., Gentzkow, M., and Williams, H. (2016). Sources of geographic variation in health care: Evidence from patient migration. *The quarterly journal of economics*, 131(4):1681–1726.

Galizzi, M. M. and Navarro-Martinez, D. (2019). On the external validity of social preference games: A systematic lab-field study. *Management Science*, 65(3):976–1002.

Green, E., Dutcher, E. G., Schold, J. D., and Stewart, D. (2025). The dynamics of deceased donor kidney transplant decision-making: Insights from studying individual clinicians' offer decisions. *American Journal of Transplantation*.

Gupta, A., Stead, T. S., and Ganti, L. (2024). Determining a meaningful r-squared value in clinical medicine. *Academic Medicine & Surgery*.

Kapor, A., Neilson, C., and Zimmerman, S. (2016). Heterogeneous beliefs and school choice. *American Economic Review*.

Kessler, J. B. and Vesterlund, L. (2015). The external validity of laboratory experiments: Qualitative rather than quantitative effects. In Fréchette, G. R. and Schotter, A., editors, *Handbook of Experimental Economic Methodology*, pages 391–406. Oxford University Press.

Kizer, K. W., English, R., and Hackmann, M. (2022). Committee on a fairer and more equitable cost-effective and transparent system of donor organ procurement allocation and distribution, national academies of sciences engineering and medicine (us). board on health sciences policy., national academies of sciences engineering and medicine (us). board on health care services., et al. realizing the promise of equity in the organ transplantation system. *National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division*.

Levitt, S. D. and List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, 21(2):153–174.

Li, J., Casalino, L. P., Fisman, R., Kariv, S., and Markovits, D. (2022). Experimental evidence of physician social preferences. *Proceedings of the National Academy of Sciences*, 119(28):e2112726119.

McCulloh, I., Stewart, D., Kiernan, K., Yazicioglu, F., Patsolic, H., Zinner, C., Mohan, S., and Cartwright, L. (2023). An experiment on the impact of predictive analytics on kidney offers acceptance decisions. *American Journal of Transplantation*, 23(7):957–965.

Molitor, D. (2018). The evolution of physician practice styles: evidence from cardiologist migration. *American Economic Journal: Economic Policy*, 10(1):326–356.

Mullainathan, S. and Obermeyer, Z. (2022). Diagnosing physician error: A machine learning approach to low-value health care. *The Quarterly Journal of Economics*, 137(2):679–727.

Pruett, T. L., Vece, G. R., Carrico, R. J., and Klassen, D. K. (2021). Us deceased kidney transplantation: Estimated gfr, donor age and kdpi association with graft survival. *EClinicalMedicine*, 37.

Rao, P. S., Schaubel, D. E., Guidinger, M. K., Andreoni, K. A., Wolfe, R. A., Merion, R. M., Port,

F. K., and Sung, R. S. (2009). A comprehensive risk quantification score for deceased donor kidneys: the kidney donor risk index. *Transplantation*, 88(2):231–236.

Stewart, D., Shepard, B., Rosendale, J., McGehee, H., Hall, I., Gupta, G., Reddy, K., Kasiske, B., Andreoni, K., and Klassen, D. (2020). Can behavioral research improve transplant decision-making? a mock offer study on the role of kidney procurement biopsies. *Kidney360*, 1(1):36–47.

U.S. Government (2024). Organ donation statistics - detailed description. Accessed: 2025-02-13.

Wilk, A. R., Beck, J., and Kucheryavaya, A. Y. (2017). The kidney allocation system (kas): The first two years. Technical report, United Network for Organ Sharing (UNOS), Richmond, VA. Prepared for the OPTN Kidney Transplantation Committee.

7. Appendix

	High R^2		Low R^2	
	1	2	3	4
Hyp. accept	2.800*	2.559*	-0.807	-1.264
	(1.115)	(1.209)	(0.866)	(0.783)
Constant	-7.566^{***}	-7.496^{***}	-4.306^{***}	-3.807^{***}
	(0.613)	(0.605)	(0.650)	(0.560)
# of clinicians	20	20	24	24
Observations	76773	76773	74687	74687

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: Regression Results for Higher and Lower R^2 Groups without EAP as a control.

	High R^2	Mid R^2	Low R^2
Hyp. accept	1.870$**$	0.0178	-1.070
	(0.597)	(1.017)	(1.473)
EAP	35.57$* * *$	33.29$* * *$	16.45$* * *$
	(3.822)	(2.437)	(2.788)
Constant	$-7.744* * *$	$-5.441* * *$	$-3.802* * *$
	(0.279)	(0.771)	(0.942)
Observations	38420	47617	14320

Standard errors in parentheses
* $p < 0.01$...

Table 6: Regression Results for Higher and Lower R^2 Groups split into thirds

7.1. Summary of Retrospective Data

7.1.1. Centers

Table 7 provides more information about each center.

Center	# of Clinicians	# of Offers
1	3	16,393
2	3	39,460
3	3	29,347
4	4	19,234
5	4	100,756
6	6	32,673
7	3	88,553
8	4	38,139
9	3	5,566
10	4	11,605
11	3	13,381
12	8	202,939
13	4	32,122
14	4	2,300
15	4	4,915
16	4	8,751
17	4	16,693
18	1	8,910
Average	3.8	9.745.3
Total	69	671,736

Table 7: Summary of retrospective data.

7.1.2. Clinicians

Figure 4 illustrates the distribution of the observed-to-expected (O/E) ratios for clinicians in the retrospective dataset, providing a measure of their relative aggressiveness in accepting donor kidneys. The median O/E ratio of 1.23 indicates that the participating clinicians, on average, accepted more offers than expected based on national patterns. This suggests a slightly more aggressive approach to risk in this sample relative to typical clinician behavior across the United States.[17] This measure provides important context for understanding how the clinicians in this study approach high-risk offers.

[17] The O/E ratio reflects the ratio of observed acceptances to expected acceptances, where a value of 1 indicates alignment with national acceptance patterns over the previous year.

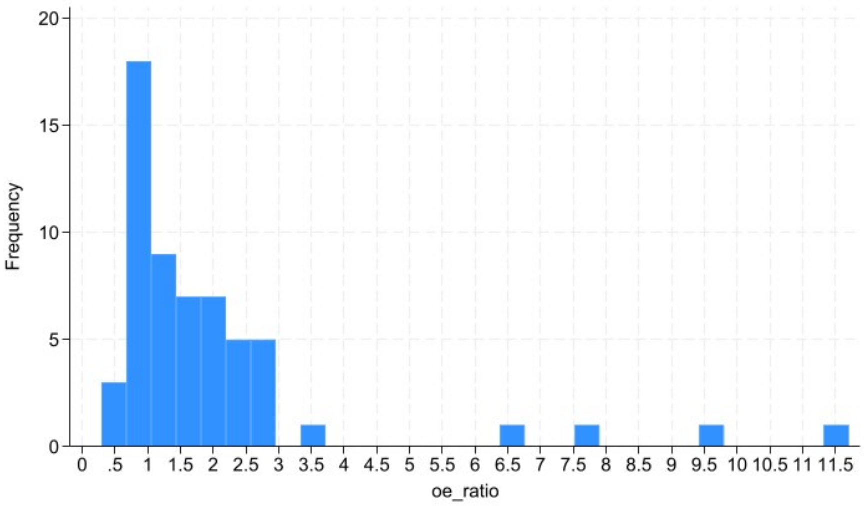

Figure 4: The distribution of the O/E ratio.