A Data-Driven Approach towards Improving Deceased-Donor Kidney Utilization

Aditya Shankar Garg AG4808@COLUMBIA.EDU
Department of Industrial Engineering and Operations Research
Columbia University
New York, NY, USA

Itai Feigenbaum ITAI@ITAIFEIGENBAUM.COM
Lehman College and the Graduate Center
City University of New York
New York, NY, USA

Miko E. Yu MY2750@CUMC.COLUMBIA.EDU
Division of Nephrology, Department of Medicine
Vagelos College of Physicians and Surgeons
Columbia University
New York, NY, USA

Sumit Mohan SM2206@CUMC.COLUMBIA.EDU
Division of Nephrology, Department of Medicine
Vagelos College of Physicians and Surgeons
Columbia University
New York, NY, USA

Jay Sethuraman JAY@IEOR.COLUMBIA.EDU
Department of Industrial Engineering and Operations Research
Columbia University
New York, NY, USA

Abstract

Deceased donor kidneys in the United States remain substantially underutilized: approximately 30% of recovered kidneys are discarded, and many other kidneys are transplanted in suboptimal condition due to delays. A major contributing factor is the allocation procedure, which offers kidneys sequentially to patients in priority order ("match run"). Offers are often declined, and repeated declines delay transplant, prolong cold ischemia time, and increase the risk of discarding the kidney. We propose a machine-learning framework to identify kidneys at risk of underutilization. Such a framework can be used to expedite the offering process and shorten the time lost due to strings of declines. We use only medical and historical information available upon arrival of the donor to the system. On the dataset of U.S. kidney offers, we predict discards well (AUROC 0.993, AUPRC 0.977). Furthermore, we are able to accurately identify 86% of the kidneys at the risk of underutilization. Our framework consists of three models, which are each an ensemble of decision trees. The first model, OfferPred, directly evaluates the chances of an offer acceptance for a particular kidney-patient pair; however, applying this model directly to evaluate an entire match run leads to error accumulation. Therefore, we train models DiscardPred and LocationPred, which indirectly use OfferPred to predict the chance of discard and identify some of the

initial string of declines. Remarkably, DiscardPred and LocationPred do not directly use granular information about patients and transplant centers in the offer sequence, but replace these with aggregates from OfferPred. Our models show that underutilization risk is not solely a function of donor characteristics but is strongly correlated with the structure of the match run and recent historical center behavior.

1. Introduction

Deceased-donor kidneys are allocated through a sequential procedure in which organs are offered to eligible candidates in a predefined order until one accepts or the organ can no longer be used. Long chains of declines often precede transplantation, leading to significant additional waiting time, which lowers the quality of the kidney and reduces the efficacy of the transplant (Cohen et al., 2018; Mohan and Schold, 2022). In fact, the kidney's quality may deteriorate to the point of not being viable: nearly 30% of recovered kidneys are discarded, despite many of them being of use to patients who do not receive offers (Husain et al., 2018; Mohan et al., 2023). Our work is motivated by the urgent need to improve the utilization of deceased-donor kidneys. To this end, we start with the following key question: When a kidney from a deceased donor becomes available, what is the risk of its being discarded under the allocation procedure currently in use? Next, assuming the kidney is likely to be transplanted, is it likely to be declined by many patients in the priority list before being accepted? Answering these questions can support targeted interventions for expediting the allocation of "hard-to-place" kidneys (Narvaez et al., 2018; Massie et al., 2010)—those that are likely to be declined by many patients—to patients further down in the priority list who may be willing to accept the kidney, while minimizing the risk of overlooking a higher priority patient (Guan et al., 2026). In addition to increasing the utilization of donor kidneys, this approach has the potential to increase transplant efficacy, especially for hard-to-place kidneys, by reducing the cold ischemia time.

We define some terminology that will be useful in the rest of the paper. A *match run* is the ordered sequence of potential offers generated for a donor kidney once it becomes available. Each run evolves as a trajectory of declines and up to two acceptances (one for each kidney). A run is *placed* if at least one offer is accepted and *discarded* otherwise. For placed runs, we define *first acceptance* as the earliest rank in the sequence at which an offer is accepted. These are run-level endpoints and do not distinguish unilateral placement from bilateral placement or nonuse.

Existing approaches for predicting kidney underutilization operate at different levels of resolution. Offer-level models estimate the probability that a specific offer will be accepted (Martinez et al., 2026). While these models capture fine-grained donor–candidate interactions, using them to sequentially characterize the evolution of an entire match run can lead to compounding error[1]. At the other extreme, donor-level models (Barah and Mehrotra, 2021; Guan et al., 2026) and center-ranking or process-aware approaches (Berry et al., 2024, 2025) capture broader system-level patterns, but do not explicitly model the ordered structure of the match run or how acceptance likelihood evolves along it.

1. Suppose an offer-level model predicts each offer outcome independently with error probability p. If recovering the evolution of a run of length n requires all n offer outcomes to be predicted correctly, then the probability of correctly recovering the full sequence is $(1-p)^n$, which decays exponentially in n.

Our approach takes a middle route by treating the match run as the primary object of prediction. We begin by assigning an offer-level score to each offer in the sequence using donor, candidate, and center-level features. Rather than chaining these predictions to simulate the entire run, which would amplify small per-offer errors across the sequence, we aggregate them into run-level summaries to estimate whether the run will end without any acceptance. Conditional on eventual placement, we estimate where in the match run acceptance is likely to occur. This decomposition separates predicting whether a run resolves from predicting where it resolves, aligning the model with the sequential structure of the allocation process without relying on independent offer-level decisions. By explicitly modeling the offer trajectory, we capture signals related to sequencing and the distribution of acceptance likelihood along the run. These signals are not available to models that operate solely on static donor characteristics or single-offer predictions. We are able to accurately identify 86% of kidneys at high-risk of underutilization on the dataset of match runs executed in 2024. Expediting the offering process for these kidneys could improve the system-wide efficiency. The main contributions of this paper are as follows,

(i) We introduce the match run as a predictive object for kidney allocation and formalize two decision-relevant tasks: discard risk and first acceptance localization.

(ii) We develop a framework that combines offer-level scoring, run-level aggregation, and within-run localization.

(iii) We demonstrate that run-level structure and historical center behavior provide substantive predictive signals in addition to donor characteristics.

(iv) We provide predictions that can be used to design expedited placement strategies for select (or hard-to-place) kidneys.

More broadly, our results suggest that hard-to-place kidneys are not solely defined by donor quality but emerge from the interaction between donor attributes, allocation order, and center behavior.

Generalizable Insights about Machine Learning in the Context of Healthcare

This paper suggests broader lessons for machine learning in healthcare workflows that unfold through repeated, ordered actions rather than a single isolated decision. First, it can be useful to separate prediction of *whether the workflow resolves* from estimation of *where resolution occurs within the workflow*, because these targets support different operational decisions and may rely on different signals. Second, while an individual step-level model is unlikely to succeed in predicting the process due to error accumulation across the sequence, it can still be valuable as intermediate representations for workflow-level prediction, since the distribution of predicted likelihood across a trajectory may encode process information beyond static information alone. Third, because many healthcare systems are partially longitudinal, temporally ordered development and evaluation, with historical features computed only from prior data, offer a practical way to reduce leakage and better approximate prospective use.

2. Related Work

Kidney allocation as a sequential process. Effective allocation of deceased-donor kidneys cannot be achieved by evaluating organ quality in isolation. The community has long recognized the importance of formulating this as a sequential assignment process in which geographical distance (Yu et al., 2025; King et al., 2023), time of offer (Mohan et al., 2016; Carpenter et al., 2019), and program behavior shape outcomes (Agarwal et al., 2025). Recent policy changes expanded eligibility and increased offer volume, raising the operational burden on transplant programs and OPOs (Organ Procurement and Transplantation Network, 2021; Adler et al., 2021; Cron et al., 2023). In parallel, OPTN introduced tools such as offer filters and predictive analytics to help programs manage this growing stream of offers (Organ Procurement and Transplantation Network, 2022, 2023). Together, this literature motivates viewing kidney placement as an evolving allocation process.

Declines, center behavior, and hard-to-place kidneys. Clinical studies show that decline cascades are shaped by more than donor risk alone. Offer outcomes depend on donor and candidate characteristics, geography, sequencing, and center behavior, and similar kidneys may follow different allocation paths depending on where and when they are offered (Cohen et al., 2018; Husain et al., 2019; Guan et al., 2025). Donor-side studies of discard and nonuse reach a similar conclusion: although donor factors matter, utilization is also influenced by center practices, geographic frictions, and system-level coordination (Mohan et al., 2018; Brennan et al., 2019; King et al., 2020; Bradbrook et al., 2025; Green et al., 2025). Additional process variation arises from OPO notification practices and out-of-sequence placement behavior (Cron et al., 2024; King et al., 2022; Adler et al., 2025). These findings suggest that hard-to-place status is partly a property of the allocation trajectory, not solely a property of the donor (King et al., 2021).

Predictive models in transplantation and the remaining gap. Prior predictive work in transplantation has focused on adjacent but distinct targets, including donor-recipient matching, donor-level discard or nonuse prediction, and offer-level acceptance support (Yoon et al., 2017; Barah and Mehrotra, 2021; Ashiku et al., 2021; Li et al., 2025; McCulloh et al., 2023). More recent process-aware approaches have moved beyond static donor prediction toward identifying kidneys likely to require broader outreach or ranking likely accepting centers for difficult placements (Massie et al., 2010; Berry et al., 2024; Guan et al., 2026). Related operational work studies simultaneous-offer burden and batching policies under expiring offers (Berry et al., 2025). However, these approaches do not model the ordered offer sequence itself as the primary object of prediction. Consequently, they do not directly address two decision-relevant questions that arise at allocation time: whether a run is likely to end without any acceptance, and where first acceptance is likely to occur if the run is ultimately placed.

Position of this paper. We address this gap by treating the ordered match run as the unit of analysis. Rather than predicting donor-level nonuse or only evaluating isolated offer decisions, we model the trajectory of the offer sequence itself and decompose the problem into offer-level signals, run-level discard prediction, and conditional localization of first acceptance. This framing is intended to better match the operational decision problem faced during kidney placement.

3. Data, Cohort, and Allocation Context

3.1. Endpoint construction

Each kidney match run is treated as an ordered offer sequence. Because a donor may yield two kidneys, a single run can contain up to two realized acceptances. For the run-level task, we classify a run as *placed* if at least one accepted offer results in transplant and as *discarded* otherwise. For placed runs, the localization target is the sequence rank of the first accepted offer. We focus on the first acceptance, as in general the difference in demand for the two kidneys is negligible: if we identify one kidney as one that should be expedited, the same generally applies to the other kidney.

3.2. Cohort Selection

Historical features for the transplant center, candidate and OPO are constructed using prior match runs beginning in 2015. The most recent iteration of the kidney allocation system, also known as KAS250 was introduced in March 2021(Organ Procurement and Transplantation Network, 2021). As a result, our analysis dataset consists of match runs from May 2021 through December 2024. We partition the data temporally, using May 2021–December 2022 for training, calendar year 2023 for validation for hyper-parameter tuning and model-selection, and calendar year 2024 for held-out testing.

We adopt a temporal split rather than random partitioning to prevent leakage of future outcomes for a donor. In particular, random splits can introduce temporal leakage about a candidate's or center's future behavior, or downstream outcomes of a donor. Our split ensures that all training runs precede validation and test runs. Our split might still introduce identity leakage, however that is not a problem. Because allocation is a longitudinal process, the same candidate or center may appear across splits; this does not cause an issue because information regarding the entity's past is available when making decisions.

The dataset contains 140,440 match runs, of which 128,483 meet one of the two study definitions (93,444 placed runs and 35,039 discard runs). After filtering out pre-KAS250, the final dataset consists of 57,313 match runs (41,384 placed runs and 15,929 discard runs), corresponding to approximately 104.2 million offers. For training OfferPred, we include only the offers up to the second acceptance in each match run. This removes trailing offers that were never materially made, avoiding learning redundant patterns.

Table 1: Temporal data split used for model development.

Split	Placed	Discarded	Total
Train (2021–2022)	17,911	6,438	24,349
Validation (2023)	11,607	4,553	16,160
Test (2024)	11,866	4,938	16,804

3.3. Feature construction

We construct features from information available at donor arrival, including donor attributes, candidate characteristics, compatibility, geographic context, and historical behavior of candidates, centers, and OPOs. Features are organized into five groups: donor static features

Table 2: Prediction tasks in the framework.

Model	Task	Information and output	Evaluation Criteria
OfferPred	Score offers on likelihood of being accepted.	Uses donor, candidate, compatibility, geography, and recent-history covariates. Outputs offer score s_{ri}.	AUROC, Average Precision
DiscardPred	Identify if a kidney is placed or discarded	Uses ordered-run summaries and OfferPred score aggregate summaries. Outputs discard score q_r.	AUROC, Average Precision
LocationPred	For a placed kidney, identify initial declines	Uses run context and score mass distribution. Outputs localization distribution $\{p_{ri} : i \in [n_r]\}$	Within-k accuracy, Mean Absolute Error.

(for ex- age, height, weight, disease history, KDPI etc.), candidate static features (for ex - age, height, weight, previous transplant information), compatibility (A/B/DR antigen mismatch between donor and candidate), geography (distance from OPO to patient's hospital), listing-center historical practices, and OPO placement practices and calendar context. All historical features are computed using data strictly prior to donor arrival, with candidate/center-level summaries derived from rolling windows over past offers to prevent temporal leakage. This representation defines the information available to OfferPred at allocation time; DiscardPred and LocationPred operate on transformations and aggregations of these inputs.

4. Methods

We model kidney placement at the level of the match run. Our goal is not to simulate every downstream offer decision one at a time. Instead we ask two decision relevant questions that arise upon donor arrival: will this match run end in a discard, and if the kidney is ultimately place how far down the run is the first acceptance likely to occur? To answer these questions we model three objects: an offer score s_{ri} capturing the probability of an offer being accepted, a run-level discard score q_r representing the probability that a run ends in a discard and for the placed runs a distribution of estimated rank at which first acceptance occurs $\{p_{ri} : i \in [n_r]\}$ (Table 4).

4.1. Tasks and notation

A match run r contains ordered sequence of offers indexed by $i = 1, \ldots, n_r$ where n_r is the length of the run (refers to the length of the ordered list of potential offers generated for a donor). For each offer we have covariates x_{ri}. The run label is $z_r = 1$ for a run that ends in a discard and $z_r = 0$ for placed run; for placed runs, t_r denotes the offer index of the first acceptance. Our framework consists of three models namely, OfferPred is trained on the observed offer-outcome (accept versus decline) and outputs a nonnegative score s_{ri}. We interpret s_{ri} as a offer-level measure of acceptance support, not as a direct prediction of the match run's evolution via sequential application of OfferPred. To summarize how this

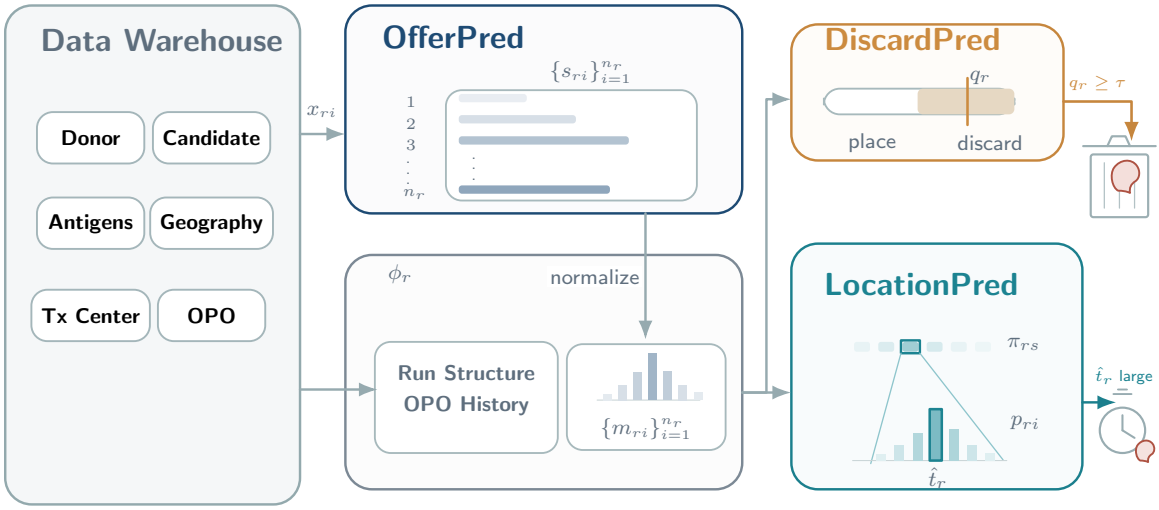

Figure 1: A visual description of our prediction framework. At inference time, the run passes through OfferPred and then through DiscardPred and LocationPred. If the discard risk is low, we localize the rank of the first acceptance. Hence we can identify hard-to-place kidneys.

support is distributed within the run, we define the normalized within-run score mass,

$$m_{ri} = \frac{s_{ri}}{\sum_{j=1}^{n_r} s_{rj}}$$

The vector $(m_{r1}, \ldots, m_{rn_r})$ describes where the acceptance support is concentrated along the match run. DiscardPred maps run-level summaries $\phi_r = \phi(x_{r1:n_r}, s_{r1:n_r}, m_{r1:n_r})$ to a discard score $q_r = f_2(\phi_r)$. LocationPred outputs a localization distribution p_{ri} over the first acceptance. Our main usage of LocationPred is to identify hard-to-place kidneys[2]. Because the ordered offer list is available when the run opens, DiscardPred and LocationPred use information available at that point rather than information revealed later in the sequence. While training OfferPred we do not utilize any offers beyond the second acceptance within the training match runs. This is done to avoid learning accept/decline behavior based on offers that were never materialized.

4.2. Technical Description

OfferPred. is a CatBoost classifier (Prokhorenkova et al., 2018) trained on a sampled training set of offers. It learns the offer-level score s_{ri} from donor, candidate, compatibility, geography, and recent-history covariates for offer i in run r. We interpret s_{ri} as a measure of *local acceptance support*: larger values indicate that the corresponding offer looks more favorable for acceptance given the information available at allocation time. The normalized quantity m_{ri} then describes how this support is distributed across the ordered run. Thus, OfferPred provides a score distribution over the run, while m_{ri} indicates where that score

2. Placed kidneys that suffer from at least a 100 declines before the first acceptance.

mass is concentrated. We do not use these offer scores directly to predict discard or first acceptance, because those are run-level tasks. In addition, naively chaining offer-level predictions across a long sequence would be sensitive to compounding error and would ignore the fact that the later tasks depend on the overall geometry of acceptance support along the run rather than on exact reconstruction of every offer outcome.

DiscardPred. is a CatBoost classifier trained on run-level features and OfferPred score summaries. It predicts the run label z_r from the run-level feature vector ϕ_r, with larger values of q_r indicating that run r ends in discard i.e. - $q_r = f_2(\phi_r)$. The key idea is moving from the offer level to the run level rather than querying OfferPred over the match run. Discard is a property of the match run as a whole, not of any single offer. We therefore summarize the OfferPred score distribution into run-level descriptors rather than applying offer scores independently across the sequence. These descriptors capture the overall level of acceptance support, whether that support is concentrated early or late in the run, and whether it is diffuse or sharply localized across the ordered list. The feature vector ϕ_r combines these score summaries with donor context, compatibility and geography, distance composition of the run, and OPO operational history. This allows DiscardPred to use the structure of the run without explicitly simulating its full offer-by-offer evolution.

LocationPred. We use a boosted-tree hazard learner on the placed-run segments. It predicts where first acceptance occurs within the ordered match run. Directly predicting the exact accepted offer is difficult because match runs vary substantially in length: the same absolute offer index can correspond to a relatively early rank in a long run and a relative late rank in a short run. To address this we convert each offer index to a normalized rank

$$u_{ri} = \frac{i}{n_r},$$

so that positions are comparable across runs of different lengths. We then partition the unit interval using fixed cutpoints $C = (0, 0.01, 0.02, 0.05, 0.10, 0.20, 0.30, 0.40, 0.60, 0.80, 1.00)$, and define the segment of offer i as $s(i) = \min\{s \in \{1, \ldots, 10\} : u_{ri} \leq C_s\}$. For run r, let the set of offers in segment s be $I_{rs} = \{i : s(i) = s\}, \qquad s = 1, \ldots, 10$, and let the set of nonempty segments be, $\mathcal{S}_r = \{s \in \{1, \ldots, 10\} : I_{rs} \neq \emptyset\}$ Because the normalized positions i/n_r are discrete, some fixed relative-rank segments may contain no offers for a given run. We therefore define hazards and renormalization only over $\mathcal{S}_r$. For each nonempty segment $s \in \mathcal{S}_r$, we model the discrete hazard that first acceptance occurs in segment s, conditional on not having occurred in any earlier nonempty segment:

$$h_{rs} = \Pr\left(t_r \in I_{rs} \,\middle|\, t_r \notin \bigcup_{\substack{j \in \mathcal{S}_r \\ j < s}} I_{rj} \right), \qquad s \in \mathcal{S}_r.$$

These hazards define a coarse probability model over the rank of first acceptance along the run. The induced segment mass is

$$\tilde{\pi}_{rs} = h_{rs} \prod_{\substack{j \in \mathcal{S}_r \\ j < s}} (1 - h_{rj}), \qquad s \in \mathcal{S}_r.$$

Because some original segments may be empty, we renormalize this mass over the nonempty segments in run r:

$$\pi_{rs} = \begin{cases} \dfrac{\tilde{\pi}_{rs}}{\sum_{\ell \in \mathcal{S}_r} \tilde{\pi}_{r\ell}}, & \text{if } \sum_{\ell \in \mathcal{S}_r} \tilde{\pi}_{r\ell} > 0, \\ \dfrac{1}{|\mathcal{S}_r|}, & \text{otherwise,} \end{cases} \qquad s \in \mathcal{S}_r.$$

Segment probabilities localize first acceptance only coarsely. To recover an offer-level distribution, we redistribute each segment's probability mass across its offers using the normalized OfferPred mass m_{ri}. Intuitively, offers with larger OfferPred mass are treated as more plausible first acceptance locations within the segment. Specifically, define the within-segment weight

$$w_{ri} = \begin{cases} \dfrac{m_{ri}}{\sum_{j \in I_{r,s(i)}} m_{rj}}, & \text{if } \sum_{j \in I_{r,s(i)}} m_{rj} > 0, \\ \dfrac{1}{|I_{r,s(i)}|}, & \text{otherwise,} \end{cases}$$

and set

$$p_{ri} = \pi_{r,s(i)}\, w_{ri}.$$

Thus, p_{ri} is the final offer-level probability that rank i is the first accepted offer. We pick the mode of this distribution as the prediction for the first acceptance.

4.3. Training and Evaluation

All model development uses the temporal split defined in Section 3.2, with historical features constructed only from prior runs. Missing categorical values were encoded as an explicit missing level, numeric features were imputed inside the model encoder, and undefined engineered summaries were set to zero after feature creation. To assess what DiscardPred gains from richer ordered-run summaries, we also evaluated prespecified restricted-information variants of the DiscardPred feature set. To keep offer-level evaluation tractable, OfferPred validation and test metrics were computed on deterministic hashed subsets of match runs.

The OfferPred score sequence used by DiscardPred and LocationPred was generated from OfferPred positive-class scores across the ordered offer list in each run. DiscardPred and LocationPred training therefore use OfferPred scores from the fitted training-period model; validation and test runs are scored out-of-sample. When we report thresholded classification counts, we use the reference rule: $\hat{z}_r = \mathbb{I}(q_r \geq \tau)$, with $\tau = 0.5$. OfferPred is assessed on the sampled held-out offer set using AUROC, average precision, and top-1% recall. DiscardPred is assessed on the held-out runs using AUROC, AUPRC, Brier score, and routed counts. LocationPred is assessed on placed runs using the modal predicted offer $\hat{t}_r = \arg\max_i p_{ri}$ with exact-hit and within-k accuracy relative to the observed first acceptance t_r. All model families, cutoffs, and operating points were fixed on the 2023 validation split before final evaluation on calendar-year 2024 data.

Data Statement

This study used data from the Scientific Registry of Transplant Recipients (SRTR). The SRTR data system includes data on all donors, wait-listed candidates, and transplant recipients in the US, submitted by the members of the Organ Procurement and Transplantation Network (OPTN). The Health Resources and Services Administration (HRSA), U.S. Department of Health and Human Services, provides oversight to the activities of the OPTN and SRTR contractors. The data reported here have been supplied by the Hennepin Healthcare Research Institute (HHRI) as the contractor for the Scientific Registry of Transplant Recipients (SRTR). 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 SRTR or the U.S. Government [3].

5. Results

We organize the results around the main clinical question: can the framework identify kidneys at risk of underutilization? We define underutilization as either discard or a kidney that is eventually placed only after a long sequence of declines. We therefore begin with the system-level result and then decompose it into the two component behaviors that matter clinically: predicting discard and localizing delayed first acceptance among placed kidneys. Lastly we discuss OfferPred's performance.

5.1. The framework identifies kidneys at risk of underutilization

On the 2024 test set, the framework identified underutilization risk across both nonplacement and substantially delayed placement. DiscardPred identified 4,882 of the 4,938 runs that ended without an accepted offer. Within the LocationPred evaluation cohort, 1,872 placed runs met the hard-to-place criterion, of which LocationPred identified 1,019. Across these two complementary outcomes, the framework therefore identified 5,901 of 6,810 at-risk runs (86.7%). Furthermore the false-alert burden is modest, i.e. - 380 of 9,723 not-underutilized kidneys were flagged as underutilized (3.9%). The value of the framework is not only that it predicts discard, but that it identifies kidneys likely to require intervention, including kidneys that are transplanted only after a prolonged sequence of declines.

5.2. Run-level prediction separates placed runs from discard runs

5.2.1. Performance improves with match-run structure

DiscardPred improves sharply as richer summaries of the ordered offer list are added. In the 2024 test set, AP rose from 0.611 with donor/OPO context alone to 0.838 after adding top-50 ranked-offer summaries, and the full model reached 0.977 (Table 3). The main gain therefore comes from capturing structure in the match run itself, not from donor descriptors alone. Among the 16,804 test runs, 4,938 (29.4%) were discard runs.

3. Due to data-use restrictions, we cannot publicly release the dataset. Our code can be found at, https://github.com/adi2809/A-Data-Driven-Approach-towards-Improving-Deceased-Donor-Kidney-Utilization/

Table 3: DiscardPred performance on the 2024 test set, with discard as the positive class. Offers add progressively richer summaries of the ordered offer list.

Information set	**AUROC**	**AP**	**Brier**
Prevalence only	0.500	0.294	0.208
Donor/OPO context	0.779	0.611	0.206
Donor + compatibility/geography	0.816	0.649	0.176
Top-10 ranked-offer summaries	0.917	0.759	0.125
Top-50 ranked-offer summaries	0.953	0.838	0.090
Full DiscardPred model	**0.993**	**0.977**	**0.032**

5.2.2. High-risk runs concentrate in a small fraction of the population

Ranking runs by the DiscardPred score q_r produced a steep concentration of discard outcomes. The highest-score 10% of test runs captured 33.5% of all discard runs while including only 0.24% of placed runs. Expanding the stratum to the highest-score 30% captured 95.3% of discard runs, and the highest-score 33% captured 98.7%. Most discard runs sit in a relatively small high-risk stratum.

5.2.3. Operating point reflects a sensitivity–precision tradeoff

Thresholded routing answers a different question: how the framework behaves once a branch rule is fixed. We report $q_r \geq 0.5$ as a reference operating point. At this cutoff, 4,882 of 4,938 observed discard runs (98.9%) were assigned to the discard branch, while 689 of 11,866 placed runs (5.8%) were also assigned there. Raising the cutoff to 0.90 reduced placed-run assignment to the discard branch from 5.8% to 1.8%, but discard recall fell from 98.9% to 90.3%. This is the core tradeoff for prospective use: broader sensitivity for finding hard-to-place kidneys versus a smaller, higher-precision review stratum.

5.2.4. Comparison to KDPI-based risk stratification

Kidney Donor Profile Index (KDPI) is a standard measure of donor quality, with higher KDPI indicating lower quality. A donor-risk ranking based on KDPI alone was markedly weaker. Treating KDPI as the score gave AUROC 0.777 and AUPRC 0.613 on the same test set. Fixed KDPI$\geq$ 75 and KDPI$\geq$ 85 rules captured 63.5% and 47.6% of discard runs, respectively, with positive predictive values of 55.5% and 64.3%. By contrast, reviewing the highest-score 30% of runs under DiscardPred captured 95.3% of discard runs, whereas the same review budget applied to KDPI captured 58.8% (Figure 2, Panel B).

Figure 3 shows what the DiscardPred score means in run-level terms. As the score increases, runs shift from early acceptance toward delayed placement and discard completion. The lowest two quintiles were dominated by placed runs, with median first acceptance rank 4 in both strata. By the fourth quintile, 49.3% of runs were discard, and by the highest quintile 97.6% were discard; the small subset of placed runs that remained in the highest quintile placed very late (median rank 858; 90th percentile 4,724). This matters for study design because high DiscardPred scores capture both likely discards and rare late-acceptance placements.

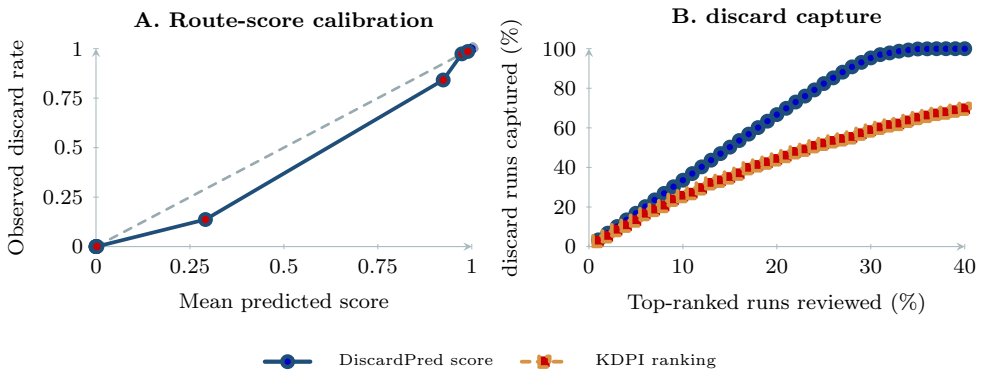

Figure 2: DiscardPred route diagnostics on the 2024 test set. Panel A compares observed discard frequency with the mean predicted score across equal-count bins. Panel B shows how quickly discard runs accumulate as larger top-ranked review sets are formed.

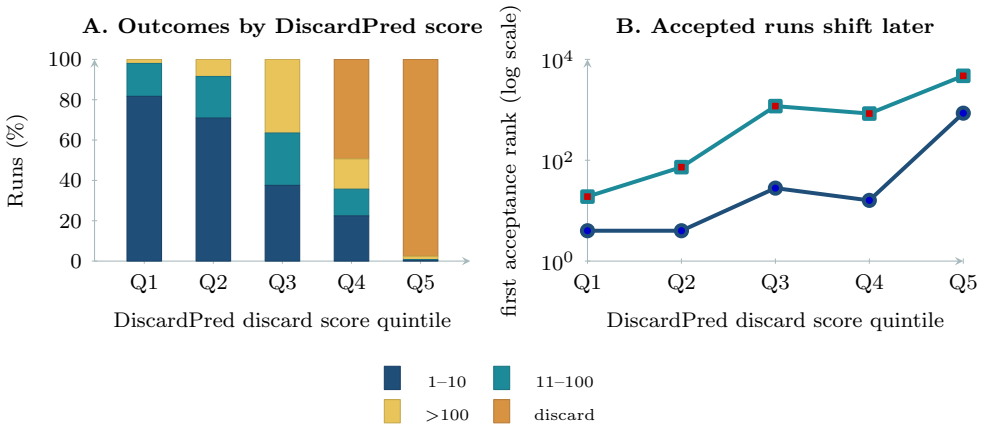

Figure 3: Run structure across DiscardPred discard score quintiles. Panel A shows the mix of discard and placed runs. Panel B shows that, among runs that place, first acceptance shifts later as the DiscardPred score rises. Panel B uses a log y-axis.

Subgroup performance remained strong across donor quality, donation pathway, local-center availability, and rank of first placement strata (Table 4). AUROC stayed above 0.98 and AUPRC stayed above 0.90 in every subgroup. AUPRC was highest in the highest-KDPI and donation after circulatory death (DCD) strata and lowest in runs where acceptance happened in 101–500, suggesting that the most difficult regime is intermediate-to-long trajectories rather than any single donor-quality slice.

Routing errors have a different consequence from ordinary classification errors because they change which downstream task is defined. At the reference threshold, 689 placed runs would not receive localization because they were assigned to the discard branch. Conversely,

Table 4: DiscardPred subgroup robustness on the 2024 test set. AUROC is reported as a point estimate; AP is reported with 95% bootstrap confidence intervals.

Subgroup	Discard (%)	AUROC	AP (95% CI)
KDPI <35	9.7	0.995	0.939 (0.915, 0.963)
KDPI 35–84	24.9	0.992	0.968 (0.960, 0.974)
KDPI ≥85	64.3	0.984	0.987 (0.983, 0.992)
Non-DCD	26.6	0.992	0.971 (0.965, 0.977)
DCD	33.0	0.994	0.983 (0.978, 0.987)
Any local offer	29.3	0.993	0.978 (0.973, 0.983)
No local offer	29.5	0.992	0.976 (0.969, 0.982)
Length 1–10	16.7	0.995	0.972 (0.960, 0.983)
Length 11–100	11.4	0.994	0.947 (0.923, 0.962)
Length 101–500	18.6	0.983	0.908 (0.885, 0.932)
Length 501+	51.2	0.989	0.985 (0.981, 0.989)

the 56 discard runs assigned to the placed-run branch have no true first acceptance target. We therefore report routed counts separately from LocationPred localization, which is evaluated on placed runs.

5.3. Segment hazards localize first acceptance within placed runs

5.3.1. Localization enables identification of hard-to-place kidneys

A key use of LocationPred is to identify kidneys that are hard to place (Narvaez et al., 2018), defined as runs where first acceptance occurs deep in the sequence. Using predicted positions to classify runs as hard-to-place or not. We achieve a high specificity (98.6%) and strong precision (88.8%), indicating that when the model flags a kidney as hard-to-place, it is usually correct. Recall is more moderate (54.4%), reflecting the tendency of the model to under-predict extreme tail cases A.10. Importantly, this result is obtained without directly training on the hard-to-place classification task. Instead, it emerges from predicting the full acceptance rank and then thresholding. We compare this approach to a direct binary classifier trained on the same inputs to predict whether a kidney is hard-to-place. This improves recall (76.1%) but at a substantial cost in precision (76.8%) and specificity (95.5%) A.10. In contrast, LocationPred yields fewer false positives (128 vs. 431), producing a much cleaner high-risk set. This highlights a key advantage of the proposed approach: modeling the full run structure provides a more precise identification of difficult cases, even if it does not maximize recall. In practice, this leads to a smaller and more reliable set of candidates for intervention.

5.3.2. Localization accuracy captures rank within the run

LocationPred improved first acceptance localization beyond both a shared empirical profile and simply selecting the highest score OfferPred offer. On the 11,866 placed-run test cases, exact accuracy rose from 16.1% under the shared profile and 44.0% under OfferPred argmax to 52.2% with segment hazards; within-5 accuracy rose to 69.8% and mean absolute error fell

to 103.4(Table 5; Figure 5). The segment hazard model therefore adds information beyond both a fixed placement profile and a single-offer maximum.

Performance was strongly moderated by run length. Within-5 accuracy was 99.5% for runs of length 1–10, 82.3% for runs of length 11–100, 58.0% for runs of length 101–500, and 34.1% for runs longer than 500 offers. The shared-profile baseline also degrades with run length, but the segment hazard model preserves an 11.6–22.5% advantage across all four strata. This pattern is expected: as runs grow long, far-tail segments contain many more possible offers, so offer-level localization becomes harder even when the model identifies the correct broad region of the run.

Table 5: LocationPred first acceptance localization on placed test runs ($n = 11{,}866$). Values are percentages except MAE, reported in offer ranks.

Model	**Exact**	**Within 1**	**Within 5**	**Within 10**	**Within 50**	**MAE**
Segment Hazard	52.2	59.0	69.8	74.9	84.3	103.4
Max OfferPred Score	44.0	52.5	65.2	70.5	80.3	125.7
Shared Profile	16.1	27.3	50.1	61.9	77.5	187.2

For interpretation, we grouped placed runs into early placement (offers 1–10), delayed placement (offers 11–100), and far-tail placement (offers >100). The segment hazard model preserved this coarse structure better than offer-level MAE alone suggests: early placements were usually retained as early (95.1%), delayed placements were most often retained as delayed (65.6%), and far-tail placements were most often classified as far-tail (57.0%). The dominant residual pattern is asymmetric, with far-tail cases more often pulled earlier than early cases pushed into the tail.

5.4. Offer scores recover acceptance structure

OfferPred recovered offer-level acceptance structure despite extreme imbalance (only 1 in every 1679 offers is an accepted offer). On the sampled 2024 offer set, it reached AUROC 0.992, specificity 99.6%, sensitivity 64%, balanced accuracy 82%, average precision 0.3258, and top-1% recall 92.5% (Table 6). The sampled 2024 evaluation set contains 371 accepted

Table 6: OfferPred held-out ranking comparison on the sampled 2024 offer benchmark. Average-precision lift is relative to the accepted-offer prevalence (0.09%).

Scoring rule	**AUROC**	**AP**	**AP lift**	**Top 1% recall**
Random/base rate	0.500	0.0009	1.0×	1.0%
Offer-order heuristic	0.943	0.0850	94×	65.5%
OfferPred offer score	**0.992**	**0.3258**	**361×**	**92.5%**

offers among 411,027 evaluated offers (0.09%) and is constructed as a deterministic held-out subset of 2024 test runs. The main comparison is with offer order, which ranks earlier offers above later offers and therefore captures the strong head-of-run tendency in kidney

acceptance. Even against this baseline, OfferPred increased average precision from 0.0850 to 0.3258 and top-1% recall from 65.5% to 92.5%, showing that the learned score surface carries more than sequence rank alone. That sharper surface is used to distinguish concentrated acceptance opportunity from weak or tail-shifted support.

6. Discussion

Underutilization emerges from the match run. Kidneys at risk of underutilization are not fully characterized by donor attributes alone, but by how those attributes interact with the allocation process. A clinically viable kidney may still be underutilized if acceptance support is weak across the run. These outcomes depend on the structure of the match run: candidate ordering, center-level acceptance behavior, geographic constraints, and the distribution of acceptance likelihood along the sequence. Modeling the match run as the primary object of prediction therefore exploits system-level signals that are not available to donor-level or isolated offer-level models.

Score distributions encode allocation dynamics. Although OfferPred is trained at the level of individual offers, its primary role is not to predict isolated acceptances but to induce a structured representation of the run. The distribution of scores along the sequence reflects how acceptance likelihood is concentrated or dispersed across candidates. Runs with early, concentrated support tend to resolve quickly, whereas runs with diffuse or tail-shifted support are more likely to experience long decline cascades or end in discard. This representation provides a stable signal for downstream prediction, even when individual offer-level predictions are noisy.

Implications for expedited placement. The framework identifies a subset of kidneys at high risk of underutilization that could benefit from modified allocation strategies. In particular, when acceptance support is weak or shifted toward the tail of the run, expediting offers to candidates who are more likely to accept may reduce cold ischemia time and decrease discard. Importantly, these predictions are derived from information available at donor arrival, making them applicable for decision support.

Limitations and translation. Localization is inherently more difficult in long runs, where many positions are plausible candidates for first acceptance; while the model captures coarse placement regimes, fine-grained rank prediction degrades in the far tail. In addition, the framework identifies kidneys that are difficult to place but does not prescribe specific interventions or quantify their downstream impact. Further evaluation, including calibration, center-level heterogeneity, and the effect of deployment on utilization and equity, is required before clinical adoption.

References

Joel T. Adler, S. Ali Husain, Kristen L. King, and Sumit Mohan. Greater complexity and monitoring of the new Kidney Allocation System: Implications and unintended consequences of concentric circle kidney allocation on network complexity. *American Journal of Transplantation*, 21(6):2007–2013, 2021. doi: 10.1111/ajt.16441.

Joel T. Adler, David C. Cron, Arnold E. Kuk, Miko Yu, Sumit Mohan, S. Ali Husain, and Layla Parast. Association between out-of-sequence allocation and deceased donor kidney nonuse across organ procurement organizations. *American Journal of Transplantation*, 25 (8):1707–1714, 2025. doi: 10.1016/j.ajt.2025.02.005.

Nikhil Agarwal, Charles Hodgson, and Paulo Somaini. Choices and outcomes in assignment mechanisms: The allocation of deceased donor kidneys. *Econometrica*, 93(2):395–438, 2025. doi: 10.3982/ECTA20203.

Lirim Ashiku, Md. Al-Amin, Sanjay Madria, and Cihan Dagli. Machine learning models and big data tools for evaluating kidney acceptance. *Procedia Computer Science*, 185:177–184, 2021. doi: 10.1016/j.procs.2021.05.019.

Masoud Barah and Sanjay Mehrotra. Predicting kidney discard using machine learning. *Transplantation*, 105(9):2054–2071, 2021. doi: 10.1097/TP.0000000000003620.

Sean Berry, Berk Görgülü, Sait Tunç, Mücahit Çevik, and Matthew J. Ellis. Optimizing hard-to-place kidney allocation: A machine learning approach to center ranking. *arXiv preprint arXiv:2410.09116*, 2024. doi: 10.48550/arXiv.2410.09116.

Sean Berry, Berk Görgülü, Sait Tunç, and Mücahit Çevik. Predicting offer burden to optimize batch sizes in simultaneously expiring kidney offers. *Frontiers in Artificial Intelligence*, 8: 1662960, 2025. doi: 10.3389/frai.2025.1662960.

Keighly Bradbrook, David Klassen, Allan B. Massie, and Darren E. Stewart. Does a changing donor pool explain the recent rise in the United States kidney nonuse rate? *American Journal of Transplantation*, 25(8):1685–1695, 2025. doi: 10.1016/j.ajt.2025.02.004.

Corey Brennan, Syed Ali Husain, Kristen L. King, Demetra Tsapepas, Lloyd E. Ratner, Zhezhen Jin, Jesse D. Schold, and Sumit Mohan. A donor utilization index to assess the utilization and discard of deceased donor kidneys perceived as high risk. *Clinical Journal of the American Society of Nephrology*, 14(11):1634–1641, 2019. doi: 10.2215/CJN.02770319.

Dustin J. Carpenter, Mariana C. Chiles, Elizabeth C. Verna, Karim J. Halazun, Jean C. Emond, Lloyd E. Ratner, and Sumit Mohan. Deceased brain dead donor liver transplantation and utilization in the United States: Nighttime and weekend effects. *Transplantation*, 103(7):1392–1404, 2019. doi: 10.1097/TP.0000000000002533.

J. B. Cohen, J. Shults, D. S. Goldberg, P. L. Abt, D. L. Sawinski, and P. P. Reese. Kidney allograft offers: Predictors of turndown and the impact of late organ acceptance on allograft survival. *American Journal of Transplantation*, 18(2):391–401, 2018. doi: 10.1111/ajt.14449.

David C. Cron, Syed A. Husain, Kristen L. King, Sumit Mohan, and Joel T. Adler. Increased volume of organ offers and decreased efficiency of kidney placement under circle-based kidney allocation. *American Journal of Transplantation*, 23(8):1209–1220, 2023. doi: 10.1016/j.ajt.2023.05.005.

David C. Cron, Arnold E. Kuk, Layla Parast, S. Ali Husain, Vanessa M. Welten, Miko Yu, Sumit Mohan, and Joel T. Adler. Variation across organ procurement organizations in deceased-donor kidney offer notification practices. *Clinical Transplantation*, 38(11):e70024, 2024. doi: 10.1111/ctr.70024.

Ellen Green, E. Glenn Dutcher, Jesse D. Schold, and Darren Stewart. The dynamics of deceased donor kidney transplant decision making: insights from studying individual clinicians' offer decisions. *American Journal of Transplantation*, 25(7):1471–1480, 2025. doi: 10.1016/j.ajt.2025.01.040.

Grace Guan, Sanjit Neelam, Joachim Studnia, Xingxing S. Cheng, Marc L. Melcher, Michael A. Rees, Alvin E. Roth, Paulo Somaini, and Itai Ashlagi. Insights from refusal patterns for deceased donor kidney offers. *Transplantation*, 109(11):1774–1782, 2025. doi: 10.1097/TP.0000000000005434.

Grace Guan, Joachim Studnia, Sanjit Neelam, Xingxing S. Cheng, Marc L. Melcher, Nikhil Agarwal, Paulo Somaini, and Itai Ashlagi. Machine learning predictions for assessing hard-to-place deceased donor kidneys. *Kidney Medicine*, 8(4):101276, 2026. doi: 10.1016/j.xkme.2026.101276.

S. Ali Husain, Kristen L. King, Stephen Pastan, Rachel E. Patzer, David J. Cohen, Jai Radhakrishnan, and Sumit Mohan. Association between declined offers of deceased donor kidney allograft and outcomes in kidney transplant candidates. *JAMA Network Open*, 2 (8):e1910312, 2019. doi: 10.1001/jamanetworkopen.2019.10312.

Syed Ali Husain, Mariana C. Chiles, Samnang Lee, Stephen O. Pastan, Rachel E. Patzer, Bekir Tanriover, Lloyd E. Ratner, and Sumit Mohan. Characteristics and performance of unilateral kidney transplants from deceased donors. *Clinical Journal of the American Society of Nephrology*, 13(1):118–127, 2018. doi: 10.2215/CJN.06550617.

Kristen L. King, S. Ali Husain, Jesse D. Schold, Rachel E. Patzer, Peter P. Reese, Zhezhen Jin, Lloyd E. Ratner, David J. Cohen, Stephen O. Pastan, and Sumit Mohan. Major variation across local transplant centers in probability of kidney transplant for wait-listed patients. *Journal of the American Society of Nephrology*, 31(12):2900–2911, 2020. doi: 10.1681/ASN.2020030335.

Kristen L. King, Sulemon G. Chaudhry, Lloyd E. Ratner, David J. Cohen, S. Ali Husain, and Sumit Mohan. Declined offers for deceased donor kidneys are not an independent reflection of organ quality. *Kidney360*, 2(11):1807–1818, 2021. doi: 10.34067/KID.0004052021.

Kristen L. King, S. Ali Husain, Adler Perotte, Joel T. Adler, Jesse D. Schold, and Sumit Mohan. Deceased donor kidneys allocated out of sequence by organ procurement organizations. *American Journal of Transplantation*, 22(5):1372–1381, 2022. doi: 10.1111/ajt.16951.

Kristen L. King, S. Ali Husain, Miko Yu, Joel T. Adler, Jesse Schold, and Sumit Mohan. Characterization of transplant center decisions to allocate kidneys to candidates with lower waiting list priority. *JAMA Network Open*, 6(6):e2316936, 2023. doi: 10.1001/jamanetworkopen.2023.16936.

Ruoting Li, Sait Tunç, Osman Y. Özaltın, and Matthew J. Ellis. Improving deceased donor kidney utilization: predicting risk of nonuse with interpretable models. *Frontiers in Artificial Intelligence*, 8:1638574, 2025. doi: 10.3389/frai.2025.1638574.

Carlos Martinez, Md Nasir, Meghana Kshirsagar, Cass McCharen, Rae Shean, Juan Lavista Ferres, Rahul Dodhia, and William B. Weeks. Predictive models for kidney offer acceptance: Challenges and strategies. *Journal of Transplantation*, 2026(1):8243450, 2026. doi: 10.1155/joot/8243450.

Allan B. Massie, Naimish M. Desai, Robert A. Montgomery, Alfred L. Singer, and Dorry L. Segev. Improving distribution efficiency of hard-to-place deceased donor kidneys: Predicting probability of discard or delay. *American Journal of Transplantation*, 10(7):1613–1620, 2010. doi: 10.1111/j.1600-6143.2010.03163.x.

Ian McCulloh, Darren Stewart, Kevin Kiernan, Ferben Yazicioglu, Heather Patsolic, Christopher Zinner, Sumit Mohan, and Laura Cartwright. An experiment on the impact of predictive analytics on kidney offers acceptance decisions. *American Journal of Transplantation*, 23(7):957–965, 2023. doi: 10.1016/j.ajt.2023.03.010.

Sumit Mohan and Jesse D. Schold. Accelerating deceased donor kidney utilization requires more than accelerating placement. *American Journal of Transplantation*, 22(1):7–8, 2022. doi: 10.1111/ajt.16866.

Sumit Mohan, Karl Foley, Mariana C. Chiles, Geoffrey K. Dube, Rachel E. Patzer, Stephen O. Pastan, R. John Crew, David J. Cohen, and Lloyd E. Ratner. The weekend effect alters the procurement and discard rates of deceased donor kidneys in the United States. *Kidney International*, 90(1):157–163, 2016. doi: 10.1016/j.kint.2016.03.007.

Sumit Mohan, Mariana C. Chiles, Rachel E. Patzer, Stephen O. Pastan, S. Ali Husain, Dustin J. Carpenter, Geoffrey K. Dube, R. John Crew, Lloyd E. Ratner, and David J. Cohen. Factors leading to the discard of deceased donor kidneys in the United States. *Kidney International*, 94(1):187–198, 2018. doi: 10.1016/j.kint.2018.02.016.

Sumit Mohan, Miko Yu, Kristen L. King, and S. Ali Husain. Increasing discards as an unintended consequence of recent changes in United States kidney allocation policy. *Kidney International Reports*, 8(5):1109–1111, 2023. doi: 10.1016/j.ekir.2023.02.1081.

J. Reinier F. Narvaez, Jing Nie, Katia Noyes, Mary Leeman, and Liise K. Kayler. Hard-to-place kidney offers: Donor- and system-level predictors of discard. *American Journal of Transplantation*, 18(11):2708–2718, 2018. doi: 10.1111/ajt.14712.

Organ Procurement and Transplantation Network. March 15 policy implementation: Removal of DSA from kidney and pancreas allocation. OPTN policy notice, 2021.

Organ Procurement and Transplantation Network. OPTN offer filters, an innovative tool that increases efficiency, is available to all U.S. kidney programs. OPTN program announcement, 2022.

Organ Procurement and Transplantation Network. OPTN predictive analytics launched to all kidney transplant programs. OPTN program announcement, 2023.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jinsung Yoon, Ahmed M. Alaa, Martin Cadeiras, and Mihaela van der Schaar. Personalized donor-recipient matching for organ transplantation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 1647–1654, 2017. doi: 10.1609/aaai.v31i1.10711.

Miko Yu, Syed Ali Husain, Joel T. Adler, Lindsey M. Maclay, Kristen L. King, Prateek V. Sahni, David C. Cron, Jesse D. Schold, and Sumit Mohan. Decreasing efficiency in deceased donor kidney offer notifications under the new distance-based kidney allocation system. *American Journal of Transplantation*, 25(8):1696–1706, 2025. doi: 10.1016/j.ajt.2025.03.010.

Appendix A. Supplementary Appendix Analyses

A.1. Model feature inventory and construction

We selected a separate information set for each prediction unit and then derived the run- and segment-level summaries described below. Table 7 gives the resulting dimensions. Match, candidate, center, and OPO identifiers were used only to construct histories; they were not model inputs. The acceptance response, first-acceptance rank, run outcome, and donor-disposition fields were used only to define labels or evaluation cohorts.

Table 7: Model-specific feature sets. Counts refer to the fitted models reported in the main text

Model	Prediction unit	Feature blocks	Input dimension
OfferPred	Offer row	Candidate and donor attributes; donor–candidate compatibility; geography; calendar context; candidate offer history; listing-center offer and placement history	88 predictors
DiscardPred	Offer run	Donor/OPO and run context; composition of the match run; raw OfferPred score-sequence summaries	26 context plus 10 score predictors
LocationPred	Run–segment pair	Run context; normalized OfferPred mass distribution; peak locations and row characteristics; local and segment-level score geometry; current-segment position and risk-set summaries	145 run/sequence plus 11 segment predictors

OfferPred: offer-row acceptance evidence. OfferPred represents each row by the information attached to that donor–candidate offer and by historical summaries available before the current match run. Candidate variables include age, sex, race, ABO type, dialysis status, diabetes, prior kidney transplant, height and weight, and current calculated panel-reactive antibody (cPRA). Donor and organ variables include age, sex, race, ABO type, height and weight, cause of death, creatinine and blood urea nitrogen, diabetes, hypertension, cancer history, donation after circulatory death (DCD), Kidney Donor Risk Index, Kidney Donor Profile Index (KDPI), and hepatitis C and hepatitis B core-antibody indicators.

Compatibility and access variables include HLA-A, HLA-B, and HLA-DR mismatch counts, total mismatch count, OPO-to-listing-center distance, high-distance indicator, and the transplant center density in 250nm circle. Calendar variables encode day of week, week of month, month of year, and hour of day at match generation. These variables allow the row score to reflect both clinical plausibility and where the offer appears in the allocation environment.

Two historical blocks describe prior behavior. Candidate history includes counts of prior transplant events and declined offers over 30-, 90-, 150-, and 365-day windows; the KDPI bin of the most recent earlier offer; the mean and variability of KDPI, donor creatinine, donor age, and mismatch count among recently declined offers; the fraction of those declines involving DCD or hepatitis C-positive donors; and time since the candidate's previous offer. Listing-center history includes prior offer volume and positive-response rates, response rates for clinically or operationally similar offers (same DCD, high-KDPI, hepatitis C, long-distance, or mismatch stratum), and the center's prior mean acceptance position and late-placement

rate. OfferPred uses short-term 30-day histories together with selected 365-day summaries to represent both recent practice and a more stable baseline.

DiscardPred: run context and raw score-sequence summaries. DiscardPred first uses one row per match run. Its 26 context predictors comprise run length and log run length; the fraction of offers linked to a listing center; donor age, KDPI, DCD, and high-KDPI indicators; the number of transplant centers within 250 nautical miles; and prior OPO summaries over the preceding 365 days. The OPO block includes historical DCD mix, placement and bilateral-nonuse rates, placement rate within the current KDPI stratum, mean declines before first acceptance, and mean run length. The remaining variables describe the composition of the match run: counts of offers within 10, 100, and 250 nautical miles; counts in four HLA-mismatch bands; counts with cPRA at least 80 and at least 98; and counts involving DCD and hepatitis C-positive donors.

Let s_{ri} be the OfferPred score for row i in run r. The DiscardPred model adds the following aggregates:

$$\max_i s_{ri}, \quad \frac{1}{n_r}\sum_{i=1}^{n_r} s_{ri}, \quad \sum_{i\le k} s_{ri} \text{ for } k \in \{10, 25, 50\}, \quad \sum_{i>50} s_{ri},$$

together with $\max_{i\le k} s_{ri}$ for $k \in \{10, 25, 50\}$ and the contrast

$$\frac{1}{|\{i : i \le 25\}|}\sum_{i\le 25} s_{ri} - \frac{1}{|\{i : i > 50\}|}\sum_{i>50} s_{ri},$$

where an empty tail is assigned mean zero. The sums retain information about both the intensity and the amount of predicted acceptance support, while the restricted maxima and head-minus-tail contrast encode whether that support appears early or only later in the match run. These are distinct from the normalized mass features used by LocationPred.

LocationPred: score geometry and segment hazard features. For localization, OfferPred scores are normalized within each run to the mass m_{ri} defined in the main text. The LocationPred run representation contains 145 context and score-geometry predictors. Rather than enumerate many mechanically related columns, we group them into their prespecified constructions:

- *Run and OPO context:* run length, donor age, KDPI, DCD, transplant-center density, and the same prior OPO placement and run-length summaries used above.
- *Global mass shape:* minimum, maximum, mean, standard deviation, selected quantiles, interquartile range, skewness, kurtosis, entropy, Herfindahl concentration, and effective support of $\{m_{ri}\}_{i=1}^{n_r}$.
- *Peak location and strength:* absolute and normalized locations of the three largest peaks, their masses and pairwise gaps, the score-weighted mean and spread of offer rank, and the share of mass held by the three largest peaks.
- *Characteristics of high-support offers:* for each of the five largest OfferPred peaks, normalized rank, raw score, normalized mass, cPRA, mismatch count, distance, time since the candidate's previous offer, and two listing-center history summaries.

- *Local and regional geometry:* mass and near-tie counts in windows around the leading peaks, left-versus-right mass contrasts, and mass assigned to each of the 10 normalized-rank intervals with cutpoints $(0, .01, .02, .05, .10, .20, .30, .40, .60, .80, 1)$; the number of offers in each interval is retained separately.

The hazard learner appends 11 descriptors for the segment currently at risk: normalized segment index; segment start, end, midpoint, and width; OfferPred mass and number of offers inside the segment; cumulative mass and row count before the segment; and remaining mass and row count after it. Segment hazards are converted to first-event probabilities and renormalized over the nonempty segments. Within the selected segment, probability is allocated across offers in proportion to m_{ri}, with a uniform fallback only when the segment has no positive OfferPred mass.

Temporal construction, missingness, and interpretation. The complete match run and its row covariates are available when the match run opens; consequently, run length, composition counts, and OfferPred sequence summaries do not require waiting for responses to accumulate. All candidate, listing-center, and OPO histories exclude the current run and use only events before its submission time. OfferPred scores for validation and test runs are generated out of sample, and no held-out acceptance labels enter the DiscardPred or LocationPred features. Missing categorical values are represented by an explicit missing level, numeric missingness is handled inside the model encoder, and undefined engineered summaries are set to zero after construction. Finally, race, sex, center-history, and OPO-history variables describe patterns in the observed allocation system; their associations should not be interpreted as causal effects or as normative measures of candidate, center, or OPO quality.

A.2. DiscardPred threshold sensitivity

The main text reports $\tau = 0.5$ as a reference routing threshold. Table 8 shows the corresponding operating characteristics across a wider range of thresholds on the 2024 test set. The key pattern is stable: lowering the threshold offers nearly complete discard recall at the cost of routing more placed runs into the discard branch, whereas raising the threshold produces a smaller and more precise review stratum but misses more hard-to-place runs.[4]

Table 8: Additional DiscardPred operating points on the 2024 test set, with discard as the positive class. Review share is the fraction of runs routed to the discard branch.

Threshold	Review (%)	Recall (%)	Precision (%)	Placed routed (%)
0.10	34.9	99.8	84.1	7.8
0.25	34.1	99.5	85.6	7.0
0.50	33.2	98.9	87.6	5.8
0.75	31.4	97.5	91.3	3.9
0.90	27.8	90.3	95.4	1.8

4. Code : https://github.com/adi2809/A-Data-Driven-Approach-towards-Improving-Deceased-Donor-Kidney-Utilization/

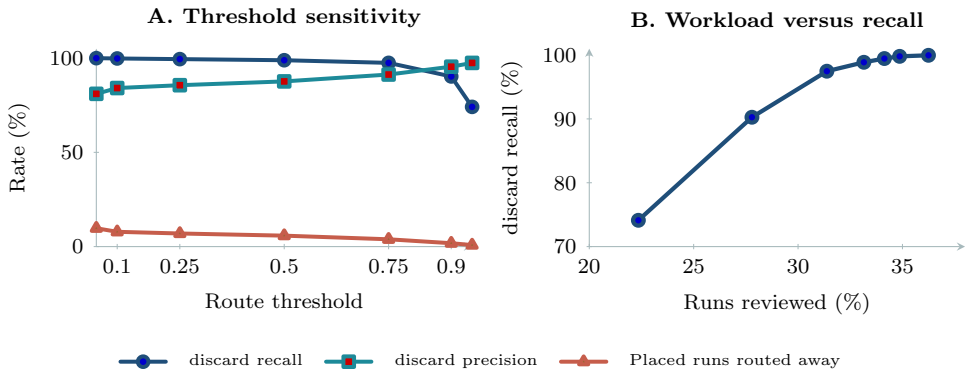

Figure 4: Supplementary DiscardPred routing-threshold diagnostics on the 2024 test set. Panel A shows how recall, precision, and placed-run false routing change with the threshold. Panel B reframes the same thresholds as workload versus discard recall.

A.3. Localization Error Structure

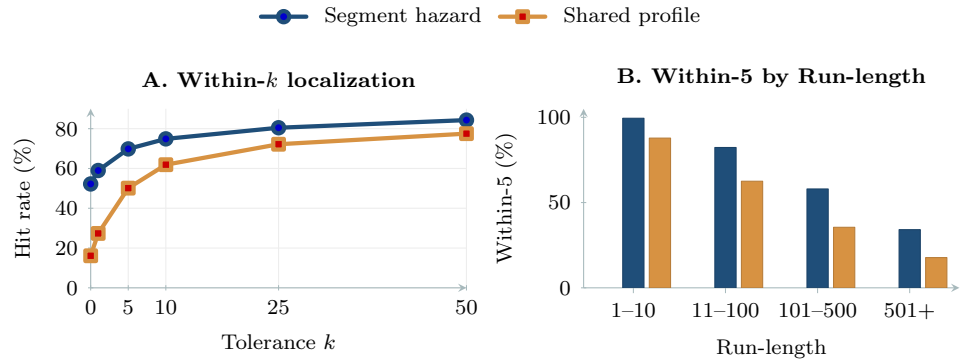

Figure 5: LocationPred localization diagnostics on placed test runs. Panel A shows within-k hit rate. Panel B shows within-5 accuracy by run-length.

A.4. Robustness to lower prior-patient overlap

Because the temporal split allows the same centers and some of the same candidates to recur across years, we evaluated DiscardPred discrimination in test subsets with lower overlap between 2024 runs and previously observed patients. Table 9 shows that route performance remains strong even in runs with no prior-patient overlap and in runs where overlap is at most 10%, suggesting that the main result is not driven by memorizing previously seen candidate sets.

Table 9: DiscardPred binary discard discrimination in lower-overlap 2024 test slices. Prior-patient overlap is defined as the share of distinct patients in a 2024 run that were seen in training or validation runs.

Slice	Runs	discard (%)	AUROC	AP
Full test	16,804	29.4	0.993	0.977
Zero overlap	468	25.6	0.996	0.984
Overlap 10% or less	496	25.0	0.996	0.984
Bottom overlap decile	1,695	17.2	0.996	0.975

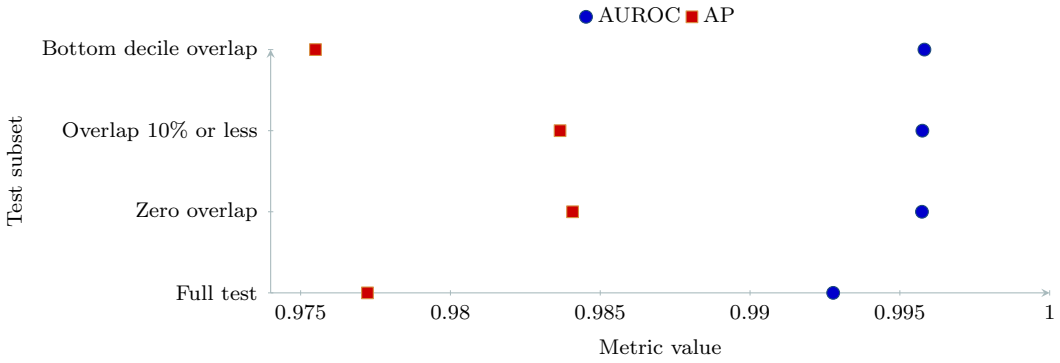

Figure 6: Supplementary DiscardPred robustness analysis under lower prior-patient overlap. Each point shows binary discard discrimination in a lower-overlap subset of the 2024 test runs.

A.5. Robustness, error bars, and statistical significance

The evaluation target is the fixed 2024 temporal test set: models are trained before 2024 and evaluated on 2024 match runs. We use a temporal rather than random split because the intended use case is forward-in-time prediction and random partitioning can leak future information. However, while temporal separation addresses leakage, it does not quantify sampling variability. We therefore provide uncertainty analyses on the test set. For error bars, the trained models are held fixed, 2024 match runs are resampled $N = 1000$ times, and each metric is recomputed. These intervals measure test-cohort uncertainty. They are not random-split error bars; random splits would answer a different question and mix earlier and later allocation behavior. The estimates are stable. DiscardPred has AUROC $0.993(0.992 - 0.994)$, AP $0.977(0.973 - 0.981)$, recall $98.9\%(98.6 - 99.1)$, and precision $87.6\%(86.8 - 88.6)$. LocationPred exact/within-10 accuracy is $52.2\%(51.4 - 53.1)/74.9\%(74.2 - 75.7)$. A paired bootstrap against KDPI on the same 2024 runs gives DiscardPred-minus-KDPI deltas of $+0.216$ AUROC $(0.209 - 0.224)$ and $+0.364$ AP $(0.350 - 0.378)$. For the composite underutilization endpoint, sensitivity/precision/specificity/accuracy are 88.5/94.3/96.1/92.9. Across 2024 quarters, DiscardPred AP is 0.968-0.984 and recall is 98.5-99.0. Moving the LocationPred cutoff from rank > 50 to > 150 keeps composite sensitivity $87.7 - 89.3$,

precision 93.7 – 95.3, and specificity 95.9 – 96.4. Quarter-wise evaluation within the test period also showed stable performance: AP 0.968–0.984, precision 0.868–0.886, and recall 0.989–0.991

A.6. Extended descriptive context for hard-to-place kidneys

The main text emphasizes the rise in discard runs and the shift in median first-acceptance rank. Figure 7 extends that description in two ways. First, the right tail of first acceptance moved much more sharply than the median: the 90th percentile rank increased from 256 in 2015 to 870 in 2024. Second, the worsening run tail coincided with worsening utilization outcomes, with kidney discard rising from 22.5% to 30.3% and donor nonutilization rising from 18.3% to 24.6%. Taken together, these patterns suggest that the hard-to-place burden grew largely through a thicker tail of very delayed placements and discard runs rather than through a uniform shift of the entire placement distribution.

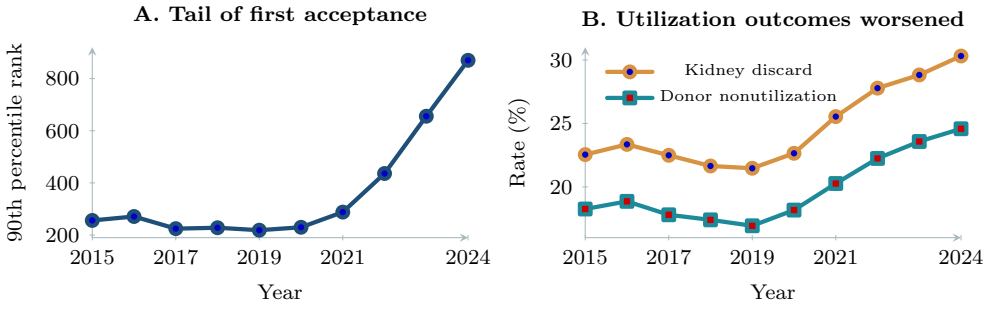

Figure 7: Supplementary descriptive context for hard-to-place kidneys in the linked cohort, 2015–2024. Panel A tracks the 90th percentile of first-acceptance rank. Panel B shows concurrent changes in kidney discard and donor nonutilization.

A.7. Hard-to-place patterns across donor and allocation strata

Figure 8 reframes hard-to-place kidneys as a mixture of four run outcomes: early placement (1–10), delayed placement (11–100), far-tail placement (>100), and discard completion. The burden concentrates strongly in high-KDPI kidneys: in 2024, kidneys with KDPI$\geq$ 85 ended in discard 64.3% of the time and placed in the first 10 offers only 9.1% of the time, compared with 9.7% and 72.8% for KDPI$<$ 35. DCD kidneys also shifted toward hard-to-place outcomes, with a 33.0% discard share versus 26.6% for non-DCD kidneys. By contrast, the presence of any local offer changed the mix of placed runs more than the overall discard rate: runs with no local offer had nearly the same discard share as runs with a local offer (29.5% versus 29.3%) but, conditional on placement, were more concentrated in the first 10 offers. This pattern is descriptive, but it suggests that local access alone does not determine difficulty of placement; the sequence of centers and candidate mix within the run remain important.

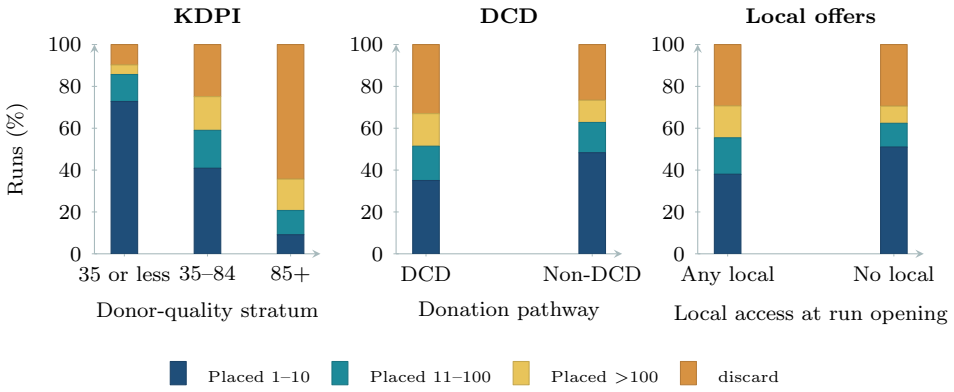

Figure 8: Supplementary outcome mix across 2024 donor and allocation strata. Each stacked bar partitions runs into early placement (1–10), delayed placement (11–100), far-tail placement (>100), and discard completion.

A.8. Variation by listing-center and OPO practice proxies

The model uses historical center and OPO behavior as descriptive context rather than as causal or normative signals. Figure 9 summarizes two such proxies on the 2024 test set: the mean recent positive-response rate among the first 10 listed centers in the run, and the historical average run length of the match OPO over the prior 365 days. Both summaries are available when the run opens and are grouped here into terciles for descriptive comparison.

The center-history gradient is substantial. Runs in the lowest tercile of early-run center responsiveness ended in discard 39.0% of the time, compared with 19.6% in the highest tercile, and the median first-acceptance rank among placed runs shifted from 11 to 5. The OPO gradient is weaker but still visible: runs from OPOs with the longest historical run lengths had a 32.2% discard share and median placed-run first acceptance at rank 8, versus 27.2% and rank 6 for runs from OPOs in the shortest-run tercile. These patterns are descriptive and should be interpreted as system-behavior summaries rather than quality scores for centers or OPOs.

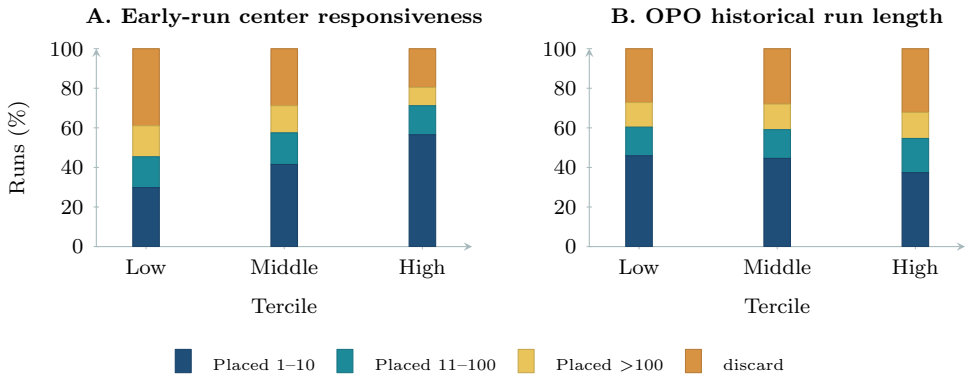

Figure 9: Supplementary outcome mix across two descriptive practice proxies in the 2024 test set. Panel A groups runs by the mean recent positive-response rate among the first 10 listed centers. Panel B groups runs by the match OPO's historical average run length over the prior 365 days.

Table 10: Selected summaries for the descriptive center- and OPO-practice strata shown in Figure 9. Median acceptance rank is computed only among placed runs.

Proxy	Tercile	Runs	discard (%)	Median acceptance rank
Early-run center responsiveness	Low	5,521	39.0	11
Early-run center responsiveness	Middle	5,521	28.8	7
Early-run center responsiveness	High	5,521	19.6	5
OPO historical run length	Low	5,521	27.2	6
OPO historical run length	Middle	5,521	28.0	6
OPO historical run length	High	5,521	32.2	8

A.9. Features associated with delayed placement

Because this study is observational, we treat feature comparisons across outcome groups as descriptive associations rather than causal effects. Figure 10 focuses on four interpretable summaries that distinguish early placement, delayed placement, far-tail placement, and discard completion. Relative to early placement, delayed and far-tail placements were associated with higher KDPI, much longer runs, lower maximum Stage 1 support anywhere in the run, and more Stage 1 score mass concentrated in the tail of the run.

Far-tail placements and discard runs looked similar on donor quality and run length but differed sharply on residual acceptance support. The mean highest offer score remained high in far-tail placements (0.780) and in runs accepted in 11–100 (0.850), while it was much lower in discard runs (0.053). Far-tail placements also carried substantially more tail score mass than discard runs (18.3 versus 2.2 on average), suggesting that very delayed placement often reflects diffuse or tail-shifted support rather than a complete absence of plausible accepting offers.

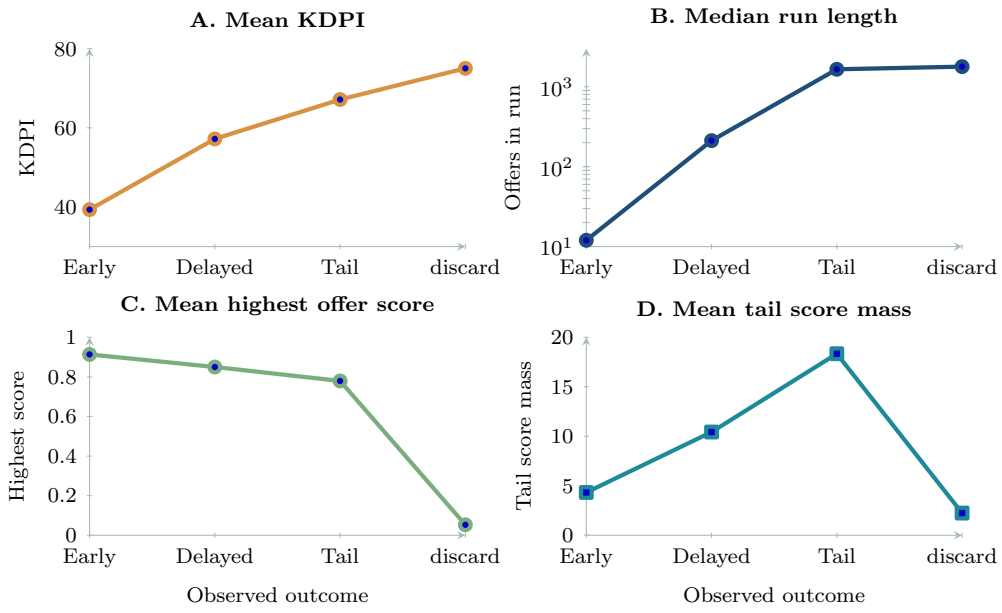

Figure 10: Supplementary descriptive comparison of four interpretable markers across early placement, delayed placement, far-tail placement, and discard runs in the 2024 test set. The purpose is descriptive: to show how donor quality, run burden, and acceptance support differ across outcome types on their native scales.

A.10. Confusion Matrices for Identification of Hard-to-Place Kidneys

The following confusion matrix reports component-level LocationPred performance. At the operating point used for this analysis, runs assigned to the discard branch by DiscardPred did not receive a LocationPred prediction. A run is classified as hard to place when first acceptance occurs after at least 100 declines.

Table 11: Component-level LocationPred confusion matrix on the 11,346 placed runs that received a LocationPred prediction. Hard-to-place runs are those with first acceptance after at least 100 declines.

	Predicted Not Hard-to-Place	**Predicted Hard-to-Place**
True Not Hard-to-Place	9,346 (TN)	128 (FP)
True Hard-to-Place	853 (FN)	1,019 (TP)

Table 12: Identification of hard-to-place runs using a direct binary classifier with the same inputs as LocationPred ($n = 11{,}346$).

	Predicted Not Hard-to-Place	**Predicted Hard-to-Place**
True Not Hard-to-Place	9,043 (TN)	431(FP)
True Hard-to-Place	447 (FN)	1,425 (TP)