Towards AI-Based Kidney Transplant Matching with Multi-Modal Inputs and Long-Term Outcome Predictions

Bowen Fan[*1], Manuel Schürch[*2], Yuan Tian[3], Anna Mallone[3], Lukas Frischknecht[3], Michael Koller[4], Christian Van Delden[5], Alexander Leichtle[6,7,10], Dela Golshayan[8], Jean Villard[5], Thomas Schachtner[3], Daniel Sidler[9], Stefan Schaub[4], Jakob Nilsson[†3], Michael Krauthammer[†1], Swiss Transplant Cohort Study

[1]University of Zurich, Zurich, Switzerland.
[2]Harvard University, Cambridge, MA, United States.
[3]University Hospital Zurich, Zurich, Switzerland.
[4]University Hospital Basel, Basel, Switzerland.
[5]Geneva University Hospitals, Geneva, Switzerland.
[6]Cantonal Hospital Baden, Baden, Switzerland.
[7]University of Bern, Bern, Switzerland.
[8]Lausanne University Hospital & University of Lausanne, Lausanne, Switzerland.
[9]Bern University Hospital, Bern, Switzerland.
[10]University Hospital Augsburg, Augsburg, Germany.

Abstract

Kidney allocation for organ transplantation must be decided within hours using only pre-transplant information. We built an allocation-time decision-support framework that forecasts five endpoints — death, graft loss, AMR, TCMR, and eGFR—across 13 annual horizons from baseline donor–recipient features. Using Swiss Transplant Cohort Study data, we created horizon-specific labels that respect terminal versus pre-terminal events, compared gradient-boosted trees (XGBoost) with a tabular foundation model (TabPFN) across increasing immunologic resolution, and summarized forecasts into a single, higher-is-better compatibility score to rank feasible recipients for an incoming donor. Internal cross-validation showed discrimination generally improved with horizon; late-horizon AUROC for death approached 0.90 with high-resolution immunology, with consistent gains for graft loss, AMR, and TCMR, while eGFR regression yielded stable R^2. Donor-specific counterfactual rankings displayed right-shifted distributions of absolute risk reduction versus the factual pairing, suggesting practical headroom for quantitative matching. The framework is policy-complementary, auditable, and equipped with case-level explanations. Coupled with post-transplant risk prediction model, it outlines a pragmatic transplant pathway, from offer triage to long-term monitoring.

1 Introduction

Kidney transplantation improves survival and quality of life for patients with end-stage renal disease, but the allocation of a scarce organ to one of several eligible candidates must be decided within hours and under uncertainty about long-term outcomes [1, 2]. At allocation time, clinicians must

[*]These authors jointly contributed to this work.
[†]These authors jointly supervised this work.
[‡]Corresponding author: bowen.fan@uzh.ch

integrate immunologic compatibility (e.g., HLA matching, presence of non acceptable HLA antigens based on patient's immune history), donor quality, recipient comorbidities, center logistics, and policy constraints. Existing tools used around allocation, such as KDRI/KDPI for donor quality and EPTS for recipient factors, summarize narrow variable sets and typically target single short-term endpoints, offering limited visibility into long-horizon trade-offs relevant to patients (long-term survival, graft durability, rejection risk, kidney function) and to programs (equitable, efficient placement) [3].

Validated post-transplant prognostic systems (e.g., iBox [4]) and dynamic machine-learning models [5, 6] that update risk using follow-up measurements have shown that outcome prediction improves once clinical data are available. However, these approaches operate after transplantation and cannot inform the allocation decision itself, where only pre-transplant information is available. In our companion post-transplant study [7] using the Swiss Transplant Cohort Study (STCS) [8], incorporating longitudinal laboratory data and clinical events markedly improved next-year risk estimation for death and graft loss compared to baseline-only models, underscoring the value of time-updated signals after surgery, and highlighting the distinct challenge at allocation time, when such signals do not yet exist.

Allocation-time prediction nonetheless must account for risks that unfold over many years. In STCS, cumulative incidence increases substantially over long horizons (e.g., mortality from ∼2.6% at year 1, to ∼40% by year 10), and non-terminal rejection events accumulate well before terminal outcomes, implying clinically meaningful, multi-outcome trade-offs that are invisible to short-horizon or single-endpoint scores. This motivates a pre-transplant framework that (i) explicitly models multiple endpoints over extended horizons and (ii) enables transparent comparisons across hypothetical donor–recipient pairings for an incoming offer.

Study objective and approach

We develop an allocation-time decision-support framework that uses *only* pre-transplant data to generate long-term forecasts for five endpoints: all-cause death, graft loss, antibody-mediated rejection (AMR), T-cell–mediated rejection (TCMR), and kidney function (eGFR), across 13 annual horizons. Predictions are constructed with tabular ML models (gradient-boosted trees and a tabular foundation model) and trained with a horizon-specific censoring scheme that respects competing risks (terminal vs. pre-terminal events) while remaining compatible with fixed-horizon classification/regression. We adopt the gradient-boosted trees as the default engine given near-parity with the foundation model and its advantages for calibration, interpretability, and deployment.

From forecasts to allocation decisions

To make forecasts actionable at the point of offer, we translate predictions of multi-outcome with multiple horizons for each candidate pairing into a single compatibility score that rewards higher projected eGFR and penalizes higher risks for death, graft loss, AMR, and TCMR using prespecified weights. This score enables side-by-side ranking of feasible recipients for a given donor and supports counterfactual matching analyses that compare the factual historical pairing with the top-ranked counterfactual candidate to estimate potential, multi-endpoint benefit. Explanations and calibration checks are embedded to support clinical audit and shared decision-making.

Contributions

This work makes three contributions beyond existing allocation-time practice and post-transplant risk tools:

- **Pre-transplant, long-horizon, multi-outcome prediction.** Using only information available at or before transplantation, we forecast five endpoints over 13 years, with labels constructed via horizon-specific censoring to respect terminal and pre-terminal event structure.
- **Transparent donor–recipient compatibility scoring.** We aggregate multi-year, multi-endpoint forecasts into a higher-is-better scalar score with prespecified weights, enabling interpretable, donor-specific ranking of feasible recipients at offer time.
- **Counterfactual matching for potential benefit estimation.** For each donor, we compare the factual match with the top-ranked counterfactual candidate to quantify expected absolute risk reductions (death, graft loss, AMR, TCMR) and eGFR gains, providing a concrete, auditable estimate of prospective benefit at the patient and cohort levels.

Together, this work assists allocation-time decision-making and aims to augment existing allocation policies rather than replace them. Moreover, it is designed to pair with the recently proposed post-transplant prognostication pipeline [7] that updates risk using follow-up measurements, outlining a pragmatic, end-to-end precision pathway from kidney offer through long-term care.

2 Methods

2.1 Cohort and Study Design

We analyzed data from the Swiss Transplant Cohort Study (STCS) [8], a nationwide, prospective, multi-center cohort initiated in May 2008 that enrolls recipients at the time of transplantation (defined as allograft reperfusion) across all six Swiss transplant centers (Figure 1 **(a)**). All solid organ transplant recipients (including islet transplantation) are eligible, and no exclusion criteria were applied. By the end of 2019, the STCS had enrolled 4,728 kidney transplant cases.

The STCS follows a standardized schedule: baseline at transplantation, 6 and 12 months, and annually thereafter, coupled with event-driven reporting of clinical outcomes (e.g., rejection, graft failure, death). This infrastructure integrates demographic, clinical, and immunologic information, links donor–recipient matching records (including ABO/HLA typing and donor characteristics) and maintains high-granularity longitudinal follow-up, thereby minimizing missingness in core pre-transplant covariates relevant to allocation-time modeling. Routine follow-ups capture immunosuppressive therapies, biopsy results, infectious events, and graft function; estimated glomerular filtration rate (eGFR) is calculated using the creatinine equation in [9]. An overview of kidney donor–recipient characteristics is provided in Table 1, with extended descriptors in the Supplementary Information (Supp. Table 2).

2.2 Data Modalities and Outcomes

The dataset is multi-modal, integrating immunologic, clinical, and demographic variables collected at or before transplantation. Immunologic inputs include ABO blood group; HLA typing for donors and recipients across class I (A, B, C) and class II (DPB1, DQB1, DRB1/3/4/5) loci, with complete typing available for most transplants but with resolution that varies across records and is not uniformly at allele-level; pre-transplant donor-specific antibodies (DSA) with mean fluorescence intensity (MFI) summaries, and compatibility summaries curated by the STCS Immunology working group. Clinical information encompasses comorbidities and medical history, peri-operative logistics (e.g., cold ischaemia time), immunosuppressive induction intent, and routine laboratory measurements (Figure 1 **(c)**). These variables are recorded within a harmonized, nationwide data infrastructure designed for long-term outcome surveillance and research use.

We model five major post-transplant outcomes over a 13-year horizon: all-cause death, graft loss, antibody-mediated rejection (AMR), T-cell–mediated rejection (TCMR), and kidney function (eGFR), as illustrated in Figure 1 **(d)**–**(e)**. Outcomes are represented as recipient-level longitudinal trajectories. For the rejection endpoints, events are captured and curated according to standardized clinical and pathology definitions used within the STCS framework; death and graft loss are recorded as terminal outcomes. We account for temporal dependencies and competing risks (e.g., graft loss precluding subsequent rejection) when constructing labels and evaluating predictions, so that interdependencies among events are preserved over time.

2.3 Predictive Modeling

We benchmarked two state-of-the-art AI models for clinical prediction, gradient-boosted decision trees (XGBoost) [10] and a pretrained tabular foundation model (TabPFN) [11]. All predictions for horizons 1–13 years post-transplant use only pre-transplant data collected at or before transplantation; no post-transplant information is used. For each horizon, binary endpoints (death, graft loss, AMR, TCMR) are framed as classification, and eGFR is modeled as regression.

XGBoost is well-suited to heterogeneous, mixed-type clinical features, capturing non-linearities and higher-order interactions while remaining robust to monotone trends, sparsity, and moderate missingness; it has repeatedly shown strong performance in electronic health record risk prediction and perioperative prognostics [12, 13]. TabPFN provides a meta-learned prior over tabular tasks via a transformer, yielding competitive, low-tuning performance on small-to-medium datasets, improved

stability under class imbalance and center heterogeneity, and strong out-of-the-box calibration when combined with simple post-hoc methods [14].

Continuous variables are standardized and categorical variables are one-hot encoded. We did not perform any explicit missingness imputation as both models can naturally handle missing values. Class imbalance is addressed via built-in class weighting; probabilistic outputs are calibrated via isotonic regression. Performance is summarized by AUROC/AUPRC for classification and R^2/RMSE for regression, with 95% confidence intervals estimated from 5-fold cross-validated assessments. Model interpretability is provided by post hoc SHAP attributions [15] per horizon and aggregated across horizons, supporting clinical review.

Hierarchical Personalization Models

We use a stepwise hierarchy to personalize pre-transplant prediction. **M1** (*population level*) relies on broadly available donor–recipient clinical and demographic variables (e.g., donor/recipient age and sex, donor type, cold ischaemia time, baseline vitals and laboratory values). It offers rapid and well-calibrated population risk but limited individual specificity. **M2** (*immunology-informed*) augments M1 with immunological compatibility and sensitization features, including ABO blood group, HLA mismatch summaries, recipient HLA compatibility summaries, pre-transplant DSA with MFI, and prior sensitizing events such as transplants, transfusions or pregnancies. These features capture alloimmune risk factors, especially relevant for AMR, that are not encoded in M1. **M3** (*HLA-typing–aware compatibility*) further integrates the raw donor HLA typing data as recorded across class I (A, B, C) and class II (DPB1, DQB1, DRB1/3/4/5) loci (at the available resolution), complementing the recipient HLA compatibility summaries already present in M2 and enabling the model to learn locus- and antigen-level, and where available higher-resolution, interactions relevant to long-term outcomes. The hierarchy degrades gradually: when detailed immunologic data are absent, predictions fall back to M1; with partial immunology, M2 improves stratification; and M3 can exploit it to provide the most granular degree of personalization. A full feature list for each tier appears in Supp. Table 1 (Supplementary Information).

Handling Competing Risks via Horizon-Specific Censoring

In this study, death and graft loss are terminal events that preclude subsequent observations, whereas antibody-mediated rejection (AMR) and T-cell–mediated rejection (TCMR) are non-terminal events that can occur only prior to a terminal event; kidney function (eGFR) is only observed among recipients alive with a functioning graft. To ensure that our fixed-horizon models (trained independently for each outcome and for each year $t = 1, \ldots, 13$ using pre-transplant covariates only) reflect this structure, we construct horizon-specific analysis sets and labels by censoring competing events and unknown statuses as follows: (i) **Death @** t: label positive if death occurs on or before t; otherwise label negative if confirmed alive at t. (ii) **Graft loss @** t: label positive if graft loss occurs on or before t; otherwise label negative if the graft is confirmed functioning at t or if death with a functioning graft occurs before t; censor if graft status by t is unknown. (iii) **AMR/TCMR @** t **(pre-terminal incidence)**: label positive only if the episode occurs on or before t *and* strictly precedes any terminal event; label negative if no AMR/TCMR and no terminal event by t; censor recipients who experience a terminal event before t without prior AMR/TCMR or whose status is otherwise unknown. (iv) **eGFR @** t: train the regressor on recipients who are alive with a functioning graft at t and have an eGFR measurement near t (pre-specified window); others are censored for this endpoint at t. This simple, horizon-specific censoring yields labels that respect competition and truncation while remaining compatible with our independent, fixed-horizon classification/regression setup.

2.4 Donor–Recipient Compatibility Scoring and Counterfactual Matching

Following Figure 3, we convert the five longitudinal predictions into a single, transparent compatibility score. For each feasible donor–recipient pair, the M3 model produces 65 pre-transplant outputs (5 outcomes $\times$ 13 years). Each outcome is summarized by the mean of its horizon-specific predictions (constructed with censoring that respects competing risks) and combined with fixed clinical weights: death 30%, graft loss 30%, AMR 15%, TCMR 15%, eGFR 10%, rewarding higher eGFR and penalizing higher risks. This score enables side-by-side ranking of candidates for an incoming donor.

Specifically, for each candidate donor–recipient pairing we predict, for each year $t = 1, \ldots, 13$, the outcomes $o \in \{\text{Death}, \text{Graft Loss}, \text{AMR}, \text{TCMR}, \text{eGFR}\}$. Let $\hat{y}_{o,t}$ denote the model's year-t prediction

for outcome o (probabilities for adverse outcomes; eGFR in mL/min/1.73m^2). We summarize each outcome by its mean over the years with available predictions:

$$\bar{y}_o = \frac{1}{|T_o|} \sum_{t \in T_o} \hat{y}_{o,t}, \quad T_o \subseteq \{1, \ldots, 13\}. \tag{1}$$

We then combine the five averages with the nonnegative weights mentioned above (summing to one). The compatibility score is:

$$\begin{aligned} S(d,r) \;=\; 100\Big(& w_{\text{death}}(1-\bar{y}_{\text{death}}) + w_{\text{graft\_loss}}(1-\bar{y}_{\text{graft\_loss}}) \\ & + w_{\text{AMR}}(1-\bar{y}_{\text{AMR}}) + w_{\text{TCMR}}(1-\bar{y}_{\text{TCMR}}) + w_{\text{eGFR}}\frac{\bar{y}_{\text{eGFR}}}{100}\Big). \end{aligned} \tag{2}$$

The complement $(1-\bar{y}_o)$ makes lower probabilities of adverse events contribute positively to the score, while eGFR is scaled by 100 to place it on a comparable $[0,1]$ range before weighting. With these choices, $S \in [0,100]$ and larger values indicate more favorable pairings. For each donor, we compute $S(d,r)$ for all candidate recipients and rank recipients by $S(d,r)$.

We then quantify the improvement of the top-ranked counterfactual match relative to the factual pairing. Let r_f denote the factual recipient for donor d, and let $r^\star = \arg\max_r S(d,r)$ be the optimal counterfactual recipient by the compatibility score. Using the averaged predictions $\bar{y}_o(\cdot)$ defined above, per-outcome gains are:

$$\text{ARR}_o(d) \;=\; \bar{y}_o(d, r_f) \;-\; \bar{y}_o(d, r^\star), \qquad o \in \{\text{death, graft\_loss, AMR, TCMR}\}, \tag{3}$$

$$\Delta_{\text{eGFR}}(d) \;=\; \bar{y}_{\text{eGFR}}(d, r^\star) \;-\; \bar{y}_{\text{eGFR}}(d, r_f). \tag{4}$$

Positive ARR_o indicates a reduction in the average predicted risk of adverse event o, and positive Δ_{eGFR} indicates an increase in average predicted eGFR (mL/min/1.73m^2).

To benchmark the learned compatibility score against a simple immunologic strategy, we additionally construct an HLA-based heuristic. For each feasible donor–recipient pair, we compute a total HLA mismatch count across class I and class II loci (A, B, C, DR, DQ, DP) using donor and recipient HLA typings. For a given donor, the heuristic selects the candidate recipient with the lowest total HLA mismatch among all eligible recipients, formalizing a "best immunologic match" rule that approximates current practice by prioritizing minimization of alloimmune risk without using longitudinal outcome predictions. For both the optimal counterfactual match $r^\star$ and the HLA-heuristic match, we compute the donor-level quantities $\{\text{ARR}_o(d), \Delta_{\text{eGFR}}(d)\}$ and compare their distributions to assess the additional adverse risk reduction achieved by the model-based compatibility score over this immunologic baseline.

Because clinical data reveal only the factual donor–recipient pairing, all quantities $\{\text{ARR}_o(d), \Delta_{\text{eGFR}}(d)\}$ are model-based counterfactual estimates of unobserved outcomes. We therefore interpret them as an auditable summary of how the model would redistribute predicted risk across feasible matches, rather than as causal effects of changing allocation policy.

As a sensitivity analysis to assess whether living donors disproportionately drive the counterfactual recipient ranking, we repeated the entire counterfactual matching and gain estimation pipeline after restricting the donor pool to deceased donors only. The resulting gain distributions are reported in Supplementary Fig. 7.

3 Results

3.1 Event Overview

Figure 1 **(f)** summarizes horizon-specific *observed* outcome frequencies for four adverse endpoints and the longitudinal kidney function trajectory across $t = 1, \ldots, 13$ years post-transplant. For each binary endpoint at horizon t, we report the proportion of positive labels among recipients with non-missing endpoint status at t in the corresponding horizon-specific analysis set (i.e., after censoring competing events and unknown statuses as described in Section 2.3). Under this definition, the observed mortality proportion increased from 2.6% at year 1 to 13.0% at year 5, and 39.7% at year 10; the observed graft loss proportion increased from 4.1% to 11.9%, and 32.9% at the same horizons. Observed pre-terminal rejection proportions increased more gradually for AMR (7.5%, 13.6%, 27.9% at years 1, 5, and 10 respectively) and more steeply for TCMR (32.5%, 40.4%, 61.6%). Mean eGFR is shown among recipients alive with a functioning graft who have an eGFR measurement within the

pre-specified window at each horizon, remaining in the mid–50s to low–60s mL/min/1.73 m^2, peaking at year 2 (59.9 ± 24.2) and declining modestly to 55.8 by year 13 (shaded band denotes ± 0.5 SD). These descriptive trajectories characterize the endpoint availability and label distribution across horizons used in our fixed-horizon prediction tasks and reflect the competing-risk and truncation handling detailed in Section 2.3.

3.2 Prediction Performance

The left column in Figure 2 presents fixed-horizon performance across t=1, . . . , 13 years for three feature specifications. We report AUROC for binary endpoints (death, graft loss, AMR, TCMR; panels 2 **(a)**–**(d)**) and RMSE for eGFR (panel 2 **(e)**); complementary AUPRC (binary) and R^2 (eGFR) appear in Supp. Figure 1 (Supplementary Information), and similar trends can be observed. Curves show mean and 95% CIs from 5-fold cross-validation, with labels/analysis sets constructed via the horizon-specific censoring scheme (Section 2.3).

Across binary outcomes, discrimination generally improves with horizon length, consistent with accumulating signal under horizon-specific censoring. Death (Figure 2 **(a)**) exhibits the strongest late-horizon gains, approaching AUROC ≈ 0.90 by year 13 with **M3**; the ordering **M3** $\geq$ **M2** $>$ **M1** holds for most horizons. Graft loss (Figure 2 **(b)**) shows a similar upward trajectory into years 10–13 with a consistent **M3** advantage. For pre-terminal rejections, early horizons are modest but rise over time: AMR reaches the high 0.8s by late follow-up (Figure 2 **(c)**), and TCMR improves into the 0.8 range by year 12 with a minor drop at year 13 (Figure 2 **(d)**). In both, **M3** typically matches or exceeds **M2**, and both outperform **M1**. For eGFR, RMSE increases gradually from early to mid follow-up and peaks around years 8–9 before stabilizing with modest variation across horizons (Figure 2 **(e)**); differences between feature sets are small, with **M3** generally achieving slightly lower (better) error than **M1** and **M2**. We observed incorporating higher-fidelity immunologic features and raw HLA/DSA representations (**M3**) yields the best overall performance across outcomes and horizons, with the largest gains evident in later years for the binary endpoints. We therefore use **M3** as the default feature specification in subsequent analyses.

Model choice (XGBoost vs. TabPFN)

The results in Figure 2 are generated with XGBoost. We additionally performed a head-to-head comparison with TabPFN, presented in Supp. Tables 3–7, using the same cross-validation folds and the **M3** feature set. For each outcome and horizon (years 1–13), we pooled out-of-fold predictions and evaluated AUROC/AUPRC for binary endpoints and R^2/RMSE for eGFR. We tested performance differences using paired bootstrap resampling over patients and applied Benjamini–Hochberg FDR correction across horizons within each outcome and metric (α=0.05). In this updated comparison, discrimination for the binary endpoints was broadly similar between models, with few horizons surviving FDR correction (e.g., limited AUROC differences for death and a single late-horizon difference for graft loss, while AMR showed no FDR-significant gaps). In contrast, for eGFR regression, XGBoost consistently outperformed TabPFN across most horizons, with higher R^2 and lower RMSE that remained significant after FDR correction. Beyond accuracy, we favor gradient-boosted trees for clinical deployment because they provide mature, efficient, regulator-aligned explainability (exact TreeSHAP [16]) that has already been demonstrated in real-time clinician-facing systems [12, 17], supporting clinician trust, auditability, and calibration. By contrast, while tabular foundation models such as TabPFN report strong accuracy on small/medium tabular datasets, their interpretability remains an active research area and their deployment ecosystem is comparatively immature [18]. Given these considerations and the observed eGFR advantage, we adopt XGBoost as the default modeling approach in downstream analyses.

Leave-one-center-out validation

Using the selected **M3** feature tier, we conducted a leave-one-center-out evaluation to assess cross-center generalizability. For each endpoint and prediction horizon, we retrained the XGBoost model on patients from five centers and evaluated performance on the held-out sixth center. As shown in Supp. Fig. 2–6, model performance was largely consistent across test centers, with no center exhibiting systematic degradation across horizons. While early horizons showed slightly higher variability (particularly for rejection-related endpoints), performance remained comparable overall, and discrimination for death and graft loss was stable and generally improved at later horizons. For eGFR,

prediction error (RMSE) also remained within a relatively narrow range across centers and years, indicating robust and equitable performance under center shift.

3.3 Model Interpretation

To interpret predictors across horizons we computed TreeSHAP values on the XGBoost models and summarized them alongside the performance curves (Figure 2, middle and right columns of panels **(a)**–**(e)**): the middle column ranks the *top-10* features by their 13-year average absolute SHAP value and the right column shows *temporal dynamics* as the SHAP value of each top 10 feature at each horizon. For death (Figure 2 **(a)**), recipient age is the dominant driver with its largest contributions in early–mid follow-up, while recipient glucose increases sharply after year 9 and becomes the leading signal in the outermost horizons (years 11–13); CIT and donor age remain moderate, and recipient DRB3/4/5 (allele 2) peaks in mid follow-up. For graft loss (Figure 2 **(b)**), CIT and donor age are most influential early (with donor type strongest in years 1–3), whereas recipient glucose is small initially but rises steeply to dominate from about year 10 onward; donor/recipient DRB3/4/5 (allele 2) peak in mid follow-up and DSA MFI is comparatively modest. AMR (Figure 2 **(c)**)) is largely immunologic: the sum of current DSA MFI is the strongest contributor with a pronounced early peak (around year 3), donor DRB3/4/5 (allele 2) is prominent in early–mid years, recipient DRB3/4/5 (allele 2) peaks later (around years 9–10), and recipient glucose becomes dominant at very late horizons (notably year 12). TCMR (Figure 2 **(d)**)) is driven by metabolic and class-II signals: recipient glucose surges after year 8 to dominate late horizons, BMI is consistently important (largest early), and DRB3/4/5 peaks mid follow-up (especially around year 7), with additional early contributions from diastolic BP and DQB1 alleles. For eGFR (Figure 2 **(e)**), donor age dominates throughout, recipient age is second but declines steadily with horizon length, and CIT and BMI contribute at intermediate levels.

Across outcomes, the explanations recover clinically expected, horizon-dependent drivers: ischemia exposure and donor factors (CIT, donor age, donor type) dominate early graft-loss risk and long-term eGFR, immunologic burden (DSA MFI and class-II HLA) concentrates in rejection risk (especially AMR), and metabolic burden (recipient glucose, BMI) becomes increasingly influential at late horizons for death, TCMR, and graft loss.

3.4 Counterfactual Donor–Recipient Matching

Using the compatibility score $S(d, r)$ defined in Section 2.4, we generated, for each donor d, a donor-specific ranking over all feasible recipients r. We report donor-level gains for two baselines (Figure 3, right): (i) the factual pairing (d, r_f), and (ii) an HLA-based heuristic match that selects the eligible recipient with the lowest total HLA mismatch across class I and class II loci. For each donor, we contrasted the *optimal* counterfactual $(d, r^\star)$, where $r^\star = \arg\max_r S(d, r)$, against each baseline using absolute risk reduction (ARR; baseline − optimal) for death, graft loss, AMR, and TCMR, and mean eGFR gain (optimal − baseline).

Relative to the factual pairing (Figure 3, top-right), gain distributions are consistently right-shifted for all four adverse endpoints, indicating that the top-ranked counterfactual match often reduces predicted long-term risks across multiple outcomes. The eGFR gains are uniformly positive and tightly distributed, suggesting a consistent functional improvement under the optimal match alongside reductions in adverse-event risk.

When benchmarking against the HLA heuristic (Figure 3, bottom-right), the distributions remain centered on the benefit side for all endpoints, but are attenuated relative to the factual comparison and show greater spread, indicating that an immunologic "best-match" strategy captures part of the achievable improvement for some donors, yet the model-based compatibility score can still identify additional multi-endpoint benefit beyond HLA mismatching alone.

Restricting counterfactual matching to deceased donors yielded gain distributions that were qualitatively similar to those obtained using all donors, with only a modest reduction in magnitude. This suggests the observed benefits are not primarily driven by living-donor kidneys (Supplementary Fig. 7).

4 Discussion

4.1 Principal findings

We present and evaluate an allocation-time kidney transplant decision-support framework that uses only pre-transplant donor–recipient information, consistent with the few-hour offer window to produce long-horizon forecasts for five clinically salient endpoints (death, graft loss, AMR, TCMR, eGFR) across 13 annual horizons. These 5×13 predictions are converted into a higher-is-better compatibility score that ranks feasible candidates for a given donor, making cross-endpoint trade-offs explicit at the moment of the offer (Figure 3). Methodologically, clinical endpoints were built with a horizon-specific censoring scheme that respects terminal versus pre-terminal event structure so that death/graft loss preclude later risk attribution, and eGFR is modeled only among survivors with functioning grafts at the horizon. This choice allowed us to use fixed-horizon classification/regression while aligning with transplant clinical trajectories [19, 20].

In internal, cross-validated evaluations (Figure 2), discrimination generally strengthened with horizon length for the binary endpoints. Late-horizon AUROC approached ~ 0.90 for death, with consistent upward trends for graft loss, AMR, and TCMR. Across feature specifications, the high-resolution immunology model (M3) typically matched or exceeded the leaner baselines (M1/M2), with the most pronounced and consistent gains observed for the long-term terminal outcomes of death and graft loss. For AMR, TCMR, and eGFR, performance differences between M2 and M3 were more modest, and at some late horizons (e.g., year 13) M2 slightly outperformed M3. This pattern suggests that the strongest gains from adding high-resolution immunologic features occur for long-term terminal risk stratification, whereas pre-terminal events such as AMR/TCMR and graft function may be more strongly influenced by short-term or post-transplant factors not fully captured in the pre-transplant feature set. In a head-to-head comparison, XGBoost performed on par with a recently proposed tabular foundation model (TabPFN) after multiple-comparison control, and we selected XGBoost as the default engine of this study for its interpretability and operational simplicity in clinical deployment.

Beyond per-outcome metrics, donor-specific rankings demonstrated practically relevant net-benefit patterns not only relative to historical (factual) pairings but also relative to a simple immunologic baseline: the optimal counterfactual match yielded right-shifted ARR distributions across death, graft loss, AMR, and TCMR and consistently positive eGFR gains versus the factual match, and it retained positive (though smaller and more variable) incremental gains when compared against an HLA-mismatch heuristic (Figure 3). This pattern suggests that while HLA matching can recover some of the headroom in predicted outcomes, integrating broader pre-transplant signals (e.g., donor factors, ischemia exposure, metabolic/clinical risk, and sensitization markers) enables additional quantitative improvement across long-horizon, multi-outcome trade-offs.

4.2 Clinical interpretation and implications

The proposed tool is designed to augment clinical judgment and existing policy rather than replace them. In practice, once an offer becomes available, the transplant team can (i) view 13-year risk trajectories for each donor-recipient pairing, (ii) compare a single, higher-is-better compatibility score across feasible candidates, and (iii) inspect case-level explanations that attribute predictions to concrete pre-transplant factors (e.g., immunologic burden, donor age, CIT, comorbidity). Because all outputs derive from baseline data only, they fit within the time constraints of offer review and can be repeated for subsequent offers to the same candidate.

Clinically, making multi-outcome trade-offs explicit can support several high-stakes decisions: (1) *accept/deferral*: identifying pairings where the composite long-term benefit exceeds center-defined thresholds; (2) *patient counseling and consent*: communicating expected trajectories, not just short-term risks; (3) *center-level consistency*: applying the same weighted objectives for all candidates to reduce ad hoc variation under time pressure; and (4) *equitable placement*: enabling auditability of how immunologic and clinical factors jointly shaped a recommendation. These functions at allocation time complement dynamic post-transplant management, where follow-up models (evaluated separately) improved next-year risk stratification and can subsequently guide surveillance and intervention—together forming a continuous pipeline from organ offer through long-term care.

Because the model is trained on observational historical allocation and outcome data, it may inherit existing inequities: if certain patient subgroups historically faced different access, comorbidity burdens, or follow-up intensity, a purely predictive allocation-time tool could inadvertently

reinforce those patterns. We therefore frame the system as decision support within existing policy constraints rather than an automated allocator, and we recommend routine auditing of discrimination/calibration and ranking impacts across clinically relevant subgroups (e.g., sex, age strata, ethnicity/nationality proxies where permitted) in addition to the center-level generalization checks we report.

Attribution patterns were clinically plausible and time-dependent: ischemia exposure and donor factors (CIT, donor age, donor type) were most informative for early graft loss and for long-term eGFR, immunologic burden (DSA MFI and class-II HLA markers) concentrated in rejection risk (particularly AMR), and metabolic burden (baseline glucose and BMI) became increasingly important at late horizons for death, TCMR, and graft loss. We did not have an explicit pre-transplant diabetes diagnosis (or HbA1c/medication indicators) in our baseline feature set; thus, baseline glucose serves as the closest available surrogate of dysglycemia. In our data, glucose co-exists with other metabolic risk components (BMI, blood pressure, HDL/LDL), so high baseline glucose likely summarizes a broader metabolic syndrome phenotype that becomes increasingly determinant of late graft and patient outcomes. Importantly, SHAP values explain model behavior rather than causal effects; attributions may reflect correlated risk profiles, center-level practice, and selection mechanisms. We therefore use explanations to support transparency, auditing, and counseling, not to infer that modifying a single factor will necessarily change outcomes.

4.3 Complementary to post-transplant monitoring model

Our allocation-time framework complements our recently proposed post-transplant prognostication framework [7] that update outcome risk prediction using time-varying labs, events, and comorbidities. Validated tools such as iBox only provide point-in-time post-transplant risk estimates [4], while dynamic ML follow-up models that ingest annually refreshed data further improve next-year risk discrimination and can surface patients in need of further surveillance or therapy adjustment. In combination, these components outline a precision kidney transplant pathway: (i) **Triage at offer** using multi-outcome, long-horizon forecasts from pre-transplant data to prioritize recipients for a specific donor; (ii) **Early course calibration** via a brief first-year update to anchor initial risk and surveillance thresholds before entering long-term monitoring; and (iii) **Long-term monitoring** using iterative, interpretable post-transplant predictions to target biopsies, adjust immunosuppression, and schedule follow-up. The envisioned clinical impact is twofold: improving patient-centered outcomes by aligning the right kidney with the right recipient upfront, and sustaining benefit through timely, data-informed interventions after surgery.

4.4 Comparison with current allocation practice and scores

Widely used scores around allocation (e.g., KDRI/KDPI for donor quality [3], EPTS for recipient factors [21]) summarize narrow variable sets and typically target single, relatively short-horizon endpoints. We did not report KDPI/KDRI because the standard OPTN/SRTR calculation requires donor height/weight, donor history of hypertension/diabetes, cause of death, and donor serum creatinine (in addition to age and DCD status), which are not available in our baseline donor variable set. Current clinical practice also depends on expert synthesis without quantitative comparison of hypothetical pairings for the incoming donor. Our approach differs by: (i) forecasting multiple outcomes that matter to patients and programs over 13 years; (ii) ranking donor-specific feasible pairings, enabling explicit trade-offs for the offer at hand; and (iii) embedding calibration and explanation to support clinical audit and trust. The framework is policy-complementary: it does not override allocation rules, but can be inserted as a transparent assistance within existing workflows (e.g., Eurotransplant kidney allocation system (ETKAS) [22]).

4.5 Limitations

First, although this is a multi-center study, it remains observational, and residual confounding from center-level practice patterns is possible. Even with careful handling of missingness and harmonization, unmeasured immunologic complexity (e.g., non-HLA antibodies, nuanced epitope landscapes) may attenuate performance, especially in highly sensitized candidates [23, 24]. Moreover, some clinically relevant features, such as detailed histopathology, are not available at offer time by design; our framework is restricted to pre-transplant information that can realistically be accessed within the allocation window. Second, our fixed-horizon labeling scheme approximates competing risks rather

than fitting a fully joint multi-state model. As a consequence, certain transitions (e.g., death with a functioning graft) may be under- or overestimated at specific horizons [25, 26]. This trade-off was made to retain compatibility with standard fixed-horizon classification/regression while respecting the main terminal versus pre-terminal structure, but it does not exhaust the space of possible event-process models. Third, donor-specific rankings are counterfactual projections derived from observational data. They quantify expected benefit under modeling assumptions and do not guarantee feasibility once logistics, allocation rules, and patient preferences are taken into account. By construction, the magnitude of any counterfactual gain cannot be empirically verified for unrealized matches (the fundamental problem of counterfactual learning) [27]; we therefore view the compatibility score as a decision-support signal rather than as a causal effect estimate. Finally, external and temporal generalizability across health systems and allocation regimes remain to be established. To this end, we have begun to explore large transplant registries (e.g., UNOS [28], ANZDATA [29], and UKTR [30]) as opportunities for further validation of the proposed framework. Beyond retrospective validation, we have not yet conducted a prospective impact evaluation. As with iBox and other clinical scores, prospective studies will ultimately be required to demonstrate that the full pathway (pre-transplant triage plus post-transplant monitoring) improves patient-centered outcomes and organ utilization within ethical and policy constraints.

Data availability

The data analyzed in this study is subject to the following licenses/restrictions: because of the STCS data protection contract, data are not directly available but can be requested from the STCS Scientific Committee (FUP 221). Requests to access these datasets should be directed to https://www.stcs.ch/about/data-center.

Code availability

All scripts used for modelling and data analysis are available on GitHub: https://github.com/uzh-dqbm-cmi/pre-transplant-allocation.

Acknowledgments

This study has been conducted in the framework of the Swiss Transplant Cohort Study, supported by the Swiss National Science Foundation, Unimedsuisse, the Swiss University Hospitals (G15) and transplant centers.

Author contribution

B.F., M.S., J.N., and M.K. conceived the study. J.N. and M.K. supervised the study. L.F., M.K., M.Ko., C.V.D., A.L., D.G., J.V., T.S., D.S., S.S. and J.N. contributed to data acquisition. B.F. and M.S. performed the data cleaning, pre-processing, and label annotation. B.F. implemented the machine learning models and ran the experiments. M.S. and M.K. advised on modeling, statistical interpretation, and evaluation. Y.T., A.M., and J.N. advised on medical content and clinical relevance. B.F., M.S. wrote the manuscript with feedback and guidance from all other co-authors. All authors contributed to interpreting the findings and refining the manuscript.

Competing Interests

The authors declare no competing interests.

Author Information

Consortium authors (Swiss Transplant Cohort Study)

The members of the Swiss Transplant Cohort Study are: Patrizia Amico, Adrian Bachofner, Vanessa Banz, Sonja Beckmann, Guido Beldi, Christoph Berger, Ekaterine Berishvili, Annalisa Berzigotti, Françoise-Isabelle Binet, Pierre-Yves Bochud, Petra Borner, Sanda Branca, Anne Cairoli, Emmanuelle Catana, Yves Chalandon, Philippe Compagnon, Sabina De Geest, Sophie De Seigneux, Michael Dickenmann, Joëlle Lynn Dreifuss, Thomas Fehr, Sylvie Ferrari-Lacraz, Andreas Flammer, Jaromil Frossard, Déla Golshayan, Nicolas Goossens, Fadi Haidar, Jürg Halter, Christoph Hess, Sven Hillinger, Hans Hirsch, Patricia Hirt, Linard Hoessly, Uyen Huynh-Do, Franz Immer, Nina Khanna, Michael Koller, Angela Koutsokera, Andreas Kremer, Thorsten Krueger, Christian Kuhn, Arnaud

L'Huillier, Bettina Laesser, Frédéric Lamoth, Roger Lehmann, Alexander Leichtle, Oriol Manuel, Hans-Peter Marti, Michele Martinelli, Valérie McLin, Katell Mellac, Aurélia Merçay, Karin Mettler, Sara Christina Meyer, Nicolas Müller, Jelena Müller, Ulrike Müller-Arndt, Mirjam Nägeli, Dionysios Neofytos, Jakob Nilsson, Manuel Pascual, Rosmarie Pazeller, David Reineke, Juliane Rick, Fabian Rössler, Silvia Rothlin, Thomas Schachtner, Stefan Schaub, Dominik Schneidawind, Macé Schuurmans, Simon Schwab, Thierry Sengstag, Daniel Sidler, Federico Simonetta, Jürg Steiger, Guido Stirnimann, Ueli Stürzinger, Christian Van Delden, Jean-Pierre Venetz, Jean Villard, Julien Vionnet, Laura Walti, Caroline Wehmeier, Patrick Yerly.

Competing Interests

The authors declare no competing interests.

References

[1] Wolfe, R.A., Ashby, V.B., Milford, E.L., Ojo, A.O., Ettenger, R.E., Agodoa, L.Y., Held, P.J., Port, F.K.: Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. New England journal of medicine **341**(23), 1725–1730 (1999)

[2] Chaudhry, D., Chaudhry, A., Peracha, J., Sharif, A.: Survival for waitlisted kidney failure patients receiving transplantation versus remaining on waiting list: systematic review and meta-analysis. Bmj **376** (2022)

[3] Rao, P.S., Schaubel, D.E., Guidinger, M.K., Andreoni, K.A., Wolfe, R.A., Merion, R.M., Port, F.K., Sung, R.S.: A comprehensive risk quantification score for deceased donor kidneys: the kidney donor risk index. Transplantation **88**(2), 231–236 (2009)

[4] Loupy, A., Aubert, O., Orandi, B.J., Naesens, M., Bouatou, Y., Raynaud, M., Divard, G., Jackson, A.M., Viglietti, D., Giral, M., et al.: Prediction system for risk of allograft loss in patients receiving kidney transplants: international derivation and validation study. Bmj **366** (2019)

[5] Ravindhran, B., Chandak, P., Schafer, N., Kundalia, K., Hwang, W., Antoniadis, S., Haroon, U., Zakri, R.H.: Machine learning models in predicting graft survival in kidney transplantation: meta-analysis. BJS open **7**(2), 011 (2023)

[6] Van Loon, E., Zhang, W., Coemans, M., De Vos, M., Emonds, M.-P., Scheffner, I., Gwinner, W., Kuypers, D., Senev, A., Tinel, C., *et al.*: Forecasting of patient-specific kidney transplant function with a sequence-to-sequence deep learning model. JAMA network open **4**(12), 2141617–2141617 (2021)

[7] Fan, B., Schürch, M., Tian, Y., Mallone, A., Frischknecht, L., Koller, M., Van Delden, C., Leichtle, A., Golshayan, D., Villard, J., *et al.*: Enhancing post-kidney transplant prognostication: an interpretable machine learning approach for longitudinal outcome prediction. npj Digital Medicine **8**(1), 684 (2025)

[8] Stampf, S., Mueller, N.J., Delden, C., Pascual, M., Manuel, O., Banz, V., Binet, I., De Geest, S., Bochud, P.-Y., Leichtle, A., *et al.*: Cohort profile: The swiss transplant cohort study (stcs): A nationwide longitudinal cohort study of all solid organ recipients in switzerland. BMJ open **11**(12), 051176 (2021)

[9] Levey, A.S., Stevens, L.A., Schmid, C.H., Zhang, Y., Castro III, A.F., Feldman, H.I., Kusek, J.W., Eggers, P., Van Lente, F., Greene, T., *et al.*: A new equation to estimate glomerular filtration rate. Annals of internal medicine **150**(9), 604–612 (2009)

[10] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)

[11] Hollmann, N., Müller, S., Eggensperger, K., Hutter, F.: Tabpfn: A transformer that solves small tabular classification problems in a second. arXiv preprint arXiv:2207.01848 (2022)

[12] Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K.-W., Newman, S.-F., Kim, J., *et al.*: Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature biomedical engineering **2**(10), 749–760 (2018)

[13] Li, B., Warren, B.E., Eisenberg, N., Beaton, D., Lee, D.S., Aljabri, B., Verma, R., Wijeysundera, D.N., Rotstein, O.D., Mestral, C., *et al.*: Machine learning to predict outcomes of endovascular intervention for patients with pad. JAMA Network Open **7**(3), 242350–242350 (2024)

[14] Tran, V.Q., Byeon, H.: Predicting dementia in parkinson's disease on a small tabular dataset using hybrid lightgbm–tabpfn and shap. Digital Health **10**, 20552076241272585 (2024)

[15] Lundberg, S.: A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874 (2017)

[16] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I.: From local explanations to global understanding with explainable ai for trees. Nature machine intelligence **2**(1), 56–67 (2020)

[17] Lyu, X., Fan, B., Hüser, M., Hartout, P., Gumbsch, T., Faltys, M., Merz, T.M., Rätsch, G., Borgwardt, K.: An empirical study on kdigo-defined acute kidney injury prediction in the intensive care unit. Bioinformatics **40**(Supplement_1), 247–256 (2024)

[18] Rundel, D., Kobialka, J., Crailsheim, C., Feurer, M., Nagler, T., Rügamer, D.: Interpretable machine learning for tabpfn. In: World Conference on Explainable Artificial Intelligence, pp. 465–476 (2024). Springer

[19] Kurland, B.F., Johnson, L.L., Egleston, B.L., Diehr, P.H.: Longitudinal data with follow-up truncated by death: match the analysis method to research aims. Statistical science: a review journal of the Institute of Mathematical Statistics **24**(2), 211 (2009)

[20] Coemans, M., Tran, T.H., Döhler, B., Massie, A.B., Verbeke, G., Segev, D.L., Gentry, S.E., Naesens, M.: A competing risks model to estimate the risk of graft failure and patient death after kidney transplantation using continuous donor-recipient age combinations. American Journal of Transplantation **25**(2), 355–367 (2025)

[21] Time, P.S.: A guide to calculating and interpreting the estimated post-transplant survival (epts) score used in the kidney allocation system (kas). Kidney **2** (2012)

[22] Mayer, G., Persijn, G.G.: Eurotransplant kidney allocation system (etkas): rationale and implementation. Nephrology Dialysis Transplantation **21**(1), 2–3 (2006)

[23] Lefaucheur, C., Loupy, A., Hill, G.S., Andrade, J., Nochy, D., Antoine, C., Gautreau, C., Charron, D., Glotz, D., Suberbielle-Boissel, C.: Preexisting donor-specific hla antibodies predict outcome in kidney transplantation. Journal of the American Society of Nephrology **21**(8), 1398–1406 (2010)

[24] Unterrainer, C., Döhler, B., Niemann, M., Lachmann, N., Süsal, C.: Can pirche-ii matching outmatch traditional hla matching? Frontiers in immunology **12**, 631246 (2021)

[25] Putter, H., Fiocco, M., Geskus, R.B.: Tutorial in biostatistics: competing risks and multi-state models. Statistics in medicine **26**(11), 2389–2430 (2007)

[26] Sapir-Pichhadze, R., Pintilie, M., Tinckam, K., Laupacis, A., Logan, A., Beyene, J., Kim, S.: Survival analysis in the presence of competing risks: the example of waitlisted kidney transplant candidates. American journal of transplantation **16**(7), 1958–1966 (2016)

[27] Shalit, U., Johansson, F.D., Sontag, D.: Estimating individual treatment effect: generalization bounds and algorithms. In: International Conference on Machine Learning, pp. 3076–3085 (2017). PMLR

[28] Roberts, M.S., Angus, D.C., Bryce, C.L., Valenta, Z., Weissfeld, L.: Survival after liver transplantation in the united states: a disease-specific analysis of the unos database. Liver transplantation **10**(7), 886–897 (2004)

[29] McDonald, S.P., Russ, G.R.: Australian registries—anzdata and anzod. Transplantation Reviews **27**(2), 46–49 (2013)

[30] Byrne, C., Caskey, F., Castledine, C., Dawnay, A., Ford, D., Fraser, S., Williams, A.: Uk renal registry. Nephron **139**(1), 1–12 (2018)

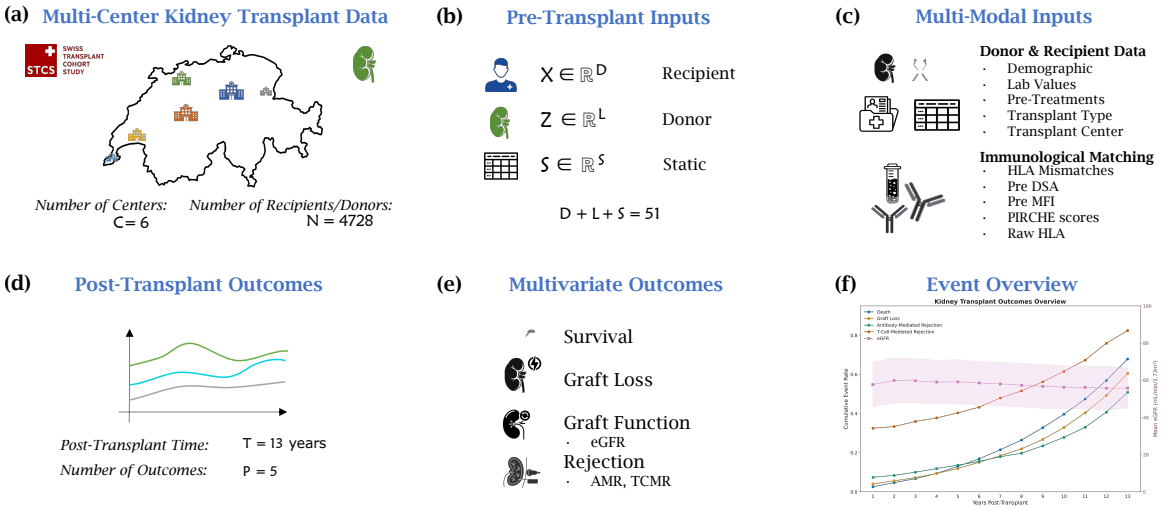


Fig. 1: **Overview of kidney transplant data and outcomes. (a)** *Multi-center cohort.* Swiss Transplant Cohort Study with C=6 transplant centers and N=4,728 recipients/donors depicted. **(b)** *Pre-transplant covariates.* Baseline covariates comprise recipient, donor and static features (in total 51 features), used to predict post-transplant outcomes. **(c)** *Multi-modal inputs.* Donor/recipient data (demographics, labs, pre-treatments, transplant type, center) and immunologic matching features (HLA mismatches, pre-DSA, MFI, PIRCHE, raw HLA). **(d) and (e)** *Post-transplant multivariate outcomes.* Five major post-transplant outcome: death (survival), graft loss, rejection (AMR, TCMR), and kidney function (eGFR), over a horizon of 13 years. **(f)** *Outcome event overview.* Cumulative incidence curves for death, graft loss, AMR, and TCMR (primary y-axis) and mean eGFR with a shaded standard deviation band (secondary y-axis) across t=1:13 years.

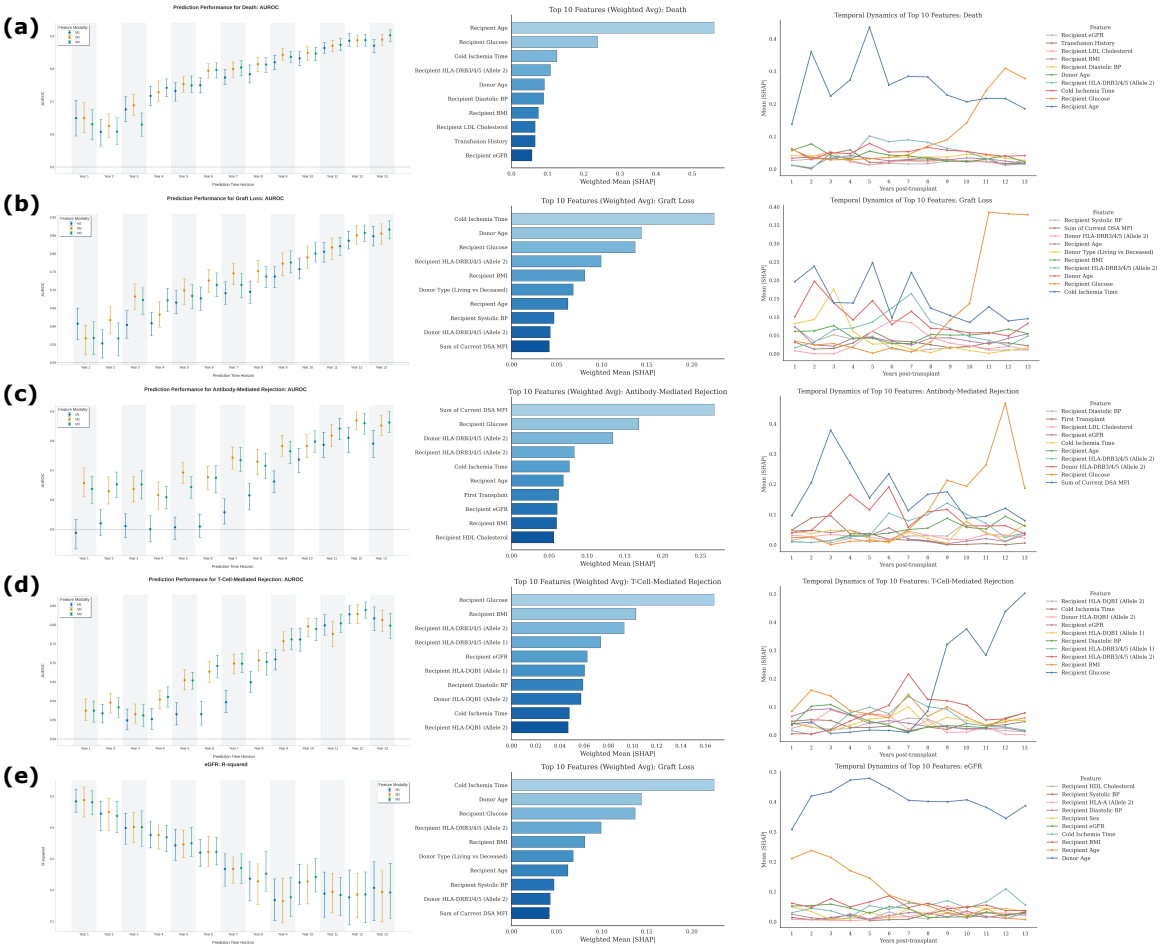


Fig. 2: **Prediction performance and feature attributions across five outcomes and 13 horizons (XGBoost).** For each outcome—**(a)** Death, **(b)** Graft loss, **(c)** Antibody-mediated rejection (AMR), **(d)** T-cell–mediated rejection (TCMR), **(e)** Kidney function (eGFR)—the figure is organized into three columns: *Left*: fixed-horizon performance over t=1 : 13 with mean and 95% CI from 5-fold CV for three feature sets—**M1** (clinical/demographic, blue), **M2** (adds immunologic compatibility, orange), **M3** (adds raw HLA/DSA representations, green); AUROC is shown for binary endpoints and R^2 for eGFR. *Middle*: global feature importance summarized as the average absolute SHAP value for the top-10 features aggregated over the 13 horizons (bars ranked by overall contribution). *Right*: temporal dynamics of the same top-10 features, showing how their mean absolute SHAP values evolve across horizons t=1 : 13. SHAP analyses are computed for the XGBoost models (top-10 features identified by average |SHAP| over horizons, using **M3** feature specification). Labels and analysis sets follow the horizon-specific censoring scheme (Section 2.3); AMR/TCMR panels reflect pre-terminal episodes only. AUPRC (binary) and RMSE (eGFR) are provided in Supp. Figure 1 Supplementary Information.

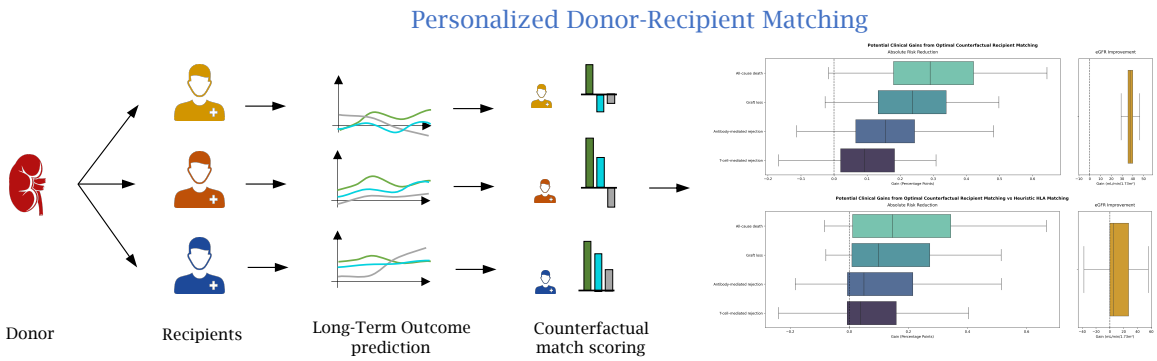


Fig. 3: **Counterfactual donor–recipient matching and comparison to factual and HLA-heuristic baselines.** For a given donor, pre-transplant covariates of all feasible recipients are paired with the donor's features and passed through the M3 predictor to obtain 65 values (5 outcomes × 13 horizons). Per outcome, 13-year averages summarize each trajectory; a composite compatibility score $S(d, r)$ is computed for each candidate pair by rewarding higher eGFR and penalizing higher risks for death, graft loss, AMR, and TCMR (higher is better). **Right, top:** outcome-wise gains for the optimal counterfactual match $r^\star = \arg\max_r S(d, r)$ versus the factual pairing r_f: absolute risk reduction (ARR; factual − optimal) for death, graft loss, AMR, and TCMR, and mean eGFR gain (optimal − factual). **Right, bottom:** additional gains of the optimal counterfactual match versus an HLA-based heuristic that selects the eligible recipient with the lowest total HLA mismatch across class I/II loci; gains are computed analogously by replacing r_f with the heuristic match. Dashed vertical lines indicate zero gain. Labels and horizon construction follow the censoring scheme in Section 2.3; all predictions and SHAP-based interpretation elsewhere in the manuscript are generated with XGBoost (M3).

Variable	Missing (n)	Transplant centre						
		Overall	**BE** *Bern*	**CHUV** *Lausanne*	**HUG** *Geneva*	**SG** *St. Gallen*	**USB** *Basel*	**USZ** *Zurich*
Cohort size								
n	–	4728	721	790	485	306	1118	1308
Recipient characteristics								
Age, Mean ± SD (years)	0	51.4 ± 15.7	51.2 ± 17.0	52.0 ± 16.1	52.8 ± 15.0	52.4 ± 13.4	52.7 ± 13.8	49.2 ± 16.7
Sex, n (%)	0							
Female		1677 (35.5)	258 (35.8)	258 (32.7)	179 (36.9)	100 (32.7)	383 (34.3)	499 (38.1)
Male		3051 (64.5)	463 (64.2)	532 (67.3)	306 (63.1)	206 (67.3)	735 (65.7)	809 (61.9)
Blood group, n (%)	0							
O		1846 (39.0)	273 (37.9)	325 (41.1)	201 (41.4)	106 (34.6)	439 (39.3)	502 (38.4)
A		2147 (45.4)	342 (47.4)	348 (44.1)	199 (41.0)	146 (47.7)	514 (46.0)	598 (45.7)
AB		207 (4.4)	34 (4.7)	33 (4.2)	21 (4.3)	9 (2.9)	46 (4.1)	64 (4.9)
B		528 (11.2)	72 (10.0)	84 (10.6)	64 (13.2)	45 (14.7)	119 (10.6)	144 (11.0)
Donor characteristics								
Age, Mean ± SD (years)	36	52.4 ± 16.1	54.3 ± 15.4	51.9 ± 15.0	52.8 ± 15.6	55.5 ± 14.8	52.6 ± 17.5	50.6 ± 16.0
Sex, n (%)	40							
Female		2331 (49.7)	359 (49.9)	418 (53.3)	242 (50.2)	137 (44.8)	567 (51.1)	608 (47.3)
Male		2357 (50.3)	361 (50.1)	366 (46.7)	240 (49.8)	169 (55.2)	543 (48.9)	678 (52.7)
Blood group, n (%)	0							
O		2075 (43.9)	293 (40.6)	418 (52.9)	213 (43.9)	111 (36.3)	472 (42.2)	568 (43.4)
A		2034 (43.0)	343 (47.6)	298 (37.7)	204 (42.1)	138 (45.1)	495 (44.3)	556 (42.5)
AB		165 (3.5)	21 (2.9)	18 (2.3)	18 (3.7)	14 (4.6)	41 (3.7)	53 (4.1)
B		454 (9.6)	64 (8.9)	56 (7.1)	50 (10.3)	43 (14.1)	110 (9.8)	131 (10.0)
Transplant-related variables								
Donation type, n (%)	0							
Deceased		2983 (63.1)	474 (65.7)	462 (58.5)	281 (57.9)	217 (70.9)	632 (56.5)	917 (70.1)
Living		1745 (36.9)	247 (34.3)	328 (41.5)	204 (42.1)	89 (29.1)	486 (43.5)	391 (29.9)
First transplant, n (%)	0							
No		753 (15.9)	121 (16.8)	130 (16.5)	75 (15.5)	45 (14.7)	164 (14.7)	218 (16.7)
Yes		3975 (84.1)	600 (83.2)	660 (83.5)	410 (84.5)	261 (85.3)	954 (85.3)	1090 (83.3)

Table 1: Baseline characteristics of STCS kidney transplant recipients and donors, overall and stratified by transplant centre.

Model	Category	Variables	Count
M1	Baseline clinical & donor logistics	recipient_age, sex, donage_complete, sex_donor, dontype_livingVSdeath, dontype_living_relatedVSunrelated, dontype_death_DBDVSDCD, coldisch_kidney1, ethnicity	9
	Baseline vitals & labs	bmi_rec0, sbp_rec0, dbp_rec0, eGFR_rec0, ldchol_rec0, hdchol_rec0, gluc_rec0	7
		Total (M1)	**16**
M2	Baseline clinical & donor logistics (M1)	recipient_age, sex, donage_complete, sex_donor, dontype_livingVSdeath, dontype_living_relatedVSunrelated, dontype_death_DBDVSDCD, coldisch_kidney1, ethnicity, bmi_rec0, sbp_rec0, dbp_rec0, eGFR_rec0, ldchol_rec0, hdchol_rec0, gluc_rec0	16
	Immunologic history & sensitization	bg, bg_donor, First_Tpx, past_pregnancy, ec_transfusion	5
	DSA summary metrics	dsaclass_agg_x, mfisubtracted_current_sum	2
	Recipient HLA compatibility (summaries)	s5hlaa_0, s5hlaa_1, s5hlab_0, s5hlab_1, s5hlac_0, s5hlac_1, s5hladpb_0, s5hladpb_1, s5hladqb_0, s5hladqb_1, s5hladrb1_0, s5hladrb1_1, s5hladrb35_0, s5hladrb35_1	14
	Raw donor HLA representations	*(not included in M2)*	0
		Total (M2)	**37**
M3	Baseline clinical & donor logistics (M1)	recipient_age, sex, donage_complete, sex_donor, dontype_livingVSdeath, dontype_living_relatedVSunrelated, dontype_death_DBDVSDCD, coldisch_kidney1, ethnicity, bmi_rec0, sbp_rec0, dbp_rec0, eGFR_rec0, ldchol_rec0, hdchol_rec0, gluc_rec0	16
	Immunologic history & sensitization	bg, bg_donor, First_Tpx, past_pregnancy, ec_transfusion	5
	DSA summary metrics	dsaclass_agg_x, mfisubtracted_current_sum	2
	Recipient HLA compatibility (summaries)	s5hlaa_0, s5hlaa_1, s5hlab_0, s5hlab_1, s5hlac_0, s5hlac_1, s5hladpb_0, s5hladpb_1, s5hladqb_0, s5hladqb_1, s5hladrb1_0, s5hladrb1_1, s5hladrb35_0, s5hladrb35_1	14
	Raw donor HLA representations	s5donor_hlaa_0, s5donor_hlaa_1, s5donor_hlab_0, s5donor_hlab_1, s5donor_hlac_0, s5donor_hlac_1, s5donor_hladpb_0, s5donor_hladpb_1, s5donor_hladqb_0, s5donor_hladqb_1, s5donor_hladrb1_0, s5donor_hladrb1_1, s5donor_hladrb35_0, s5donor_hladrb35_1	14
		Total (M3)	**51**

Supp. Table 1: Detailed feature breakdown for the three hierarchical personalization sets (M1–M3). M1 contains broad, pre-transplant clinical and donor logistics features. M2 augments M1 with immunologic compatibility and sensitization history. M3 further adds raw donor HLA representations to maximize personalization. Counts reflect the number of predictors in each category for the given model.

		Missing	Overall	BE	CHUV	HUG	SG	USB	USZ
n			4728	721	790	485	306	1118	1308
recipient_age, mean (SD)		0	51.4 (15.7)	51.2 (17.0)	52.0 (16.1)	52.8 (15.0)	52.4 (13.4)	52.7 (13.8)	49.2 (16.7)
donage_complete, mean (SD)		36	52.4 (16.1)	54.3 (15.4)	51.9 (15.0)	52.8 (15.6)	55.5 (14.8)	52.6 (17.5)	50.6 (16.0)
bmi_rec0, mean (SD)		574	26.0 (11.1)	27.2 (22.7)	25.6 (4.7)	25.7 (4.3)	26.9 (5.0)	26.4 (10.7)	25.3 (6.1)
sbp_rec0, mean (SD)		622	139.1 (21.4)	136.4 (21.0)	138.4 (20.1)	134.0 (23.5)	143.3 (22.3)	142.9 (21.0)	139.0 (21.1)
dbp_rec0, mean (SD)		624	80.5 (14.2)	78.6 (14.3)	78.5 (13.7)	78.7 (16.4)	79.7 (12.4)	81.4 (13.9)	82.9 (13.8)
eGFR_rec0, mean (SD)		272	9.8 (9.0)	10.3 (6.6)	10.8 (6.2)	10.0 (6.4)	8.8 (3.7)	8.9 (11.6)	9.8 (10.4)
ldchol_rec0, mean (SD)		1795	2.3 (1.0)	2.6 (1.1)	2.2 (1.0)	2.4 (1.0)	2.4 (1.1)	2.3 (1.0)	2.2 (1.0)
hdchol_rec0, mean (SD)		1705	1.3 (0.5)	1.3 (0.5)	1.3 (0.5)	1.2 (0.5)	1.2 (0.5)	1.3 (0.5)	1.3 (0.5)
gluc_rec0, mean (SD)		1688	6.1 (2.5)	6.5 (2.3)	6.2 (2.4)	6.1 (2.5)	6.0 (2.1)	5.9 (2.0)	5.9 (3.2)
coldisch_kidney1, mean (SD)		445	419.1 (309.6)	375.3 (240.7)	376.0 (265.9)	473.1 (394.5)	423.5 (256.3)	403.3 (331.9)	461.3 (316.0)
sumhlamismatch, mean (SD)		4449	7.9 (2.4)	7.3 (2.2)	8.1 (2.7)	8.3 (2.4)	7.5 (1.9)	7.5 (2.1)	8.1 (2.5)
hlaamismatch, mean (SD)		50	1.2 (0.7)	1.2 (0.7)	1.3 (0.7)	1.2 (0.7)	1.2 (0.7)	1.2 (0.7)	1.2 (0.7)
hlabmismatch, mean (SD)		46	1.4 (0.6)	1.4 (0.6)	1.4 (0.6)	1.4 (0.6)	1.5 (0.6)	1.4 (0.7)	1.5 (0.6)
hlacmismatch, mean (SD)		3778	1.2 (0.7)	1.2 (0.7)	1.4 (0.6)	1.1 (0.8)	1.3 (0.6)	1.2 (0.7)	1.3 (0.7)
hladpbmismatch, mean (SD)		4352	1.1 (0.7)	1.0 (0.7)	1.0 (0.7)	0.9 (0.7)	1.2 (0.6)	1.0 (0.8)	1.1 (0.7)
hladqbmismatch, mean (SD)		2142	0.9 (0.7)	0.9 (0.7)	1.0 (0.7)	0.9 (0.7)	0.8 (0.6)	0.9 (0.7)	1.0 (0.7)
hladrb1mismatch, mean (SD)		48	1.1 (0.7)	1.2 (0.7)	1.2 (0.7)	1.2 (0.7)	1.0 (0.6)	1.1 (0.7)	1.1 (0.7)
hladrb35mismatch, mean (SD)		2772	0.5 (0.6)	0.5 (0.6)	0.5 (0.6)	0.4 (0.6)	0.5 (0.6)	0.5 (0.6)	0.5 (0.6)
mfisubtracted_current_sum, mean (SD)		4120	3828.3 (5336.7)	5630.4 (5454.6)	4867.9 (6035.8)	5710.8 (7646.5)	5152.6 (7005.2)	4040.9 (6554.2)	2848.5 (3852.9)
hlaa_MFI_0, mean (SD)		1	54.0 (454.3)	49.2 (411.5)	34.5 (437.5)	37.8 (448.5)	105.4 (679.9)	34.0 (330.8)	79.4 (509.6)
hlaa_MFI_1, mean (SD)		0	1.4 (62.0)	0.0 (0.0)	2.9 (81.7)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	3.3 (99.4)
hlab_MFI_0, mean (SD)		0	47.8 (488.1)	57.0 (506.2)	20.2 (220.3)	27.4 (258.6)	22.9 (319.9)	66.5 (715.2)	56.7 (451.9)
hlab_MFI_1, mean (SD)		0	8.6 (247.0)	7.3 (118.6)	6.9 (118.8)	8.5 (137.5)	4.7 (82.4)	19.2 (475.9)	2.1 (46.4)
hlac_MFI_0, mean (SD)		2	48.4 (567.6)	39.4 (333.8)	39.3 (586.1)	6.1 (96.5)	3.8 (67.0)	88.5 (891.8)	50.6 (458.6)
hlac_MFI_1, mean (SD)		0	6.3 (203.8)	0.0 (0.0)	11.9 (249.6)	0.0 (0.0)	0.0 (0.0)	11.7 (338.8)	5.7 (119.9)
hladpb_MFI_0, mean (SD)		0	63.0 (585.1)	57.7 (615.3)	40.2 (434.2)	34.2 (368.9)	56.2 (434.7)	66.5 (650.0)	88.9 (677.4)
hladpb_MFI_1, mean (SD)		0	5.1 (165.9)	7.6 (203.7)	11.9 (335.3)	0.0 (0.0)	0.0 (0.0)	0.5 (16.9)	6.6 (91.9)
hladqb_MFI_0, mean (SD)		0	132.9 (1018.8)	185.7 (1208.9)	95.8 (914.4)	77.9 (872.7)	134.3 (1101.1)	84.7 (793.6)	187.5 (1154.2)
hladqb_MFI_1, mean (SD)		0	12.3 (306.9)	10.5 (171.7)	4.9 (138.8)	17.6 (342.5)	42.5 (743.2)	11.6 (389.1)	9.4 (104.0)
hladrb1_MFI_0, mean (SD)		0	52.0 (491.1)	64.9 (609.4)	10.7 (210.8)	45.7 (609.3)	142.5 (1081.6)	22.5 (243.6)	76.3 (418.8)
hladrb1_MFI_1, mean (SD)		0	5.0 (143.9)	18.6 (338.5)	0.0 (0.0)	5.8 (126.7)	5.7 (99.5)	0.7 (22.4)	3.9 (54.1)
hladrb35_MFI_0, mean (SD)		2	52.9 (505.8)	81.5 (624.4)	16.5 (292.7)	45.1 (630.1)	4.1 (71.7)	31.1 (317.6)	91.8 (648.4)
hladrb35_MFI_1, mean (SD)		0	2.3 (52.8)	6.2 (100.7)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	4.8 (67.0)
PIRCHE_II, mean (SD)		114	57.9 (26.8)	58.5 (26.3)	59.3 (28.0)	59.3 (26.1)	56.1 (23.7)	55.7 (27.5)	58.4 (26.7)
PIRCHE_A, mean (SD)		114	12.9 (9.8)	13.1 (9.9)	13.3 (9.8)	13.3 (9.4)	12.6 (9.6)	12.2 (9.8)	13.0 (9.8)
PIRCHE_B, mean (SD)		114	10.7 (7.3)	10.7 (7.6)	10.7 (7.4)	11.4 (7.5)	10.9 (6.8)	10.2 (7.3)	10.8 (7.3)
PIRCHE_C, mean (SD)		114	13.6 (9.2)	13.5 (8.6)	13.8 (9.3)	13.4 (9.2)	14.1 (9.0)	13.3 (9.6)	13.8 (9.2)
PIRCHE_DQ, mean (SD)		114	13.8 (10.5)	14.1 (10.4)	14.3 (10.7)	13.9 (10.6)	12.3 (9.7)	13.2 (10.7)	13.9 (10.5)
PIRCHE_DR, mean (SD)		114	7.8 (6.0)	8.0 (6.2)	8.0 (6.0)	8.3 (5.9)	7.2 (5.7)	7.5 (6.1)	7.8 (6.0)
sex, n (%)	Female	0	1677 (35.5)	258 (35.8)	258 (32.7)	179 (36.9)	100 (32.7)	383 (34.3)	499 (38.1)
	Male		3051 (64.5)	463 (64.2)	532 (67.3)	306 (63.1)	206 (67.3)	735 (65.7)	809 (61.9)
sex_donor, n (%)	Female	40	2331 (49.7)	359 (49.9)	418 (53.3)	242 (50.2)	137 (44.8)	567 (51.1)	608 (47.3)
	Male		2357 (50.3)	361 (50.1)	366 (46.7)	240 (49.8)	169 (55.2)	543 (48.9)	678 (52.7)
bg, n (%)	0	0	1846 (39.0)	273 (37.9)	325 (41.1)	201 (41.4)	106 (34.6)	439 (39.3)	502 (38.4)
	A		2147 (45.4)	342 (47.4)	348 (44.1)	199 (41.0)	146 (47.7)	514 (46.0)	598 (45.7)
	AB		207 (4.4)	34 (4.7)	33 (4.2)	21 (4.3)	9 (2.9)	46 (4.1)	64 (4.9)
	B		528 (11.2)	72 (10.0)	84 (10.6)	64 (13.2)	45 (14.7)	119 (10.6)	144 (11.0)
bg_donor, n (%)	0	0	2075 (43.9)	293 (40.6)	418 (52.9)	213 (43.9)	111 (36.3)	472 (42.2)	568 (43.4)
	A		2034 (43.0)	343 (47.6)	298 (37.7)	204 (42.1)	138 (45.1)	495 (44.3)	556 (42.5)
	AB		165 (3.5)	21 (2.9)	18 (2.3)	18 (3.7)	14 (4.6)	41 (3.7)	53 (4.1)
	B		454 (9.6)	64 (8.9)	56 (7.1)	50 (10.3)	43 (14.1)	110 (9.8)	131 (10.0)
dontype_livingVSdeath, n (%)	0.0	0	2983 (63.1)	474 (65.7)	462 (58.5)	281 (57.9)	217 (70.9)	632 (56.5)	917 (70.1)
	1.0		1745 (36.9)	247 (34.3)	328 (41.5)	204 (42.1)	89 (29.1)	486 (43.5)	391 (29.9)
dontype_living_relatedVSunrelated, n (%)	0.0	0	3914 (82.8)	607 (84.2)	651 (82.4)	406 (83.7)	263 (85.9)	878 (78.5)	1109 (84.8)
	1.0		814 (17.2)	114 (15.8)	139 (17.6)	79 (16.3)	43 (14.1)	240 (21.5)	199 (15.2)
dontype_death_DBDVSDCD, n (%)	0.0	0	2380 (50.3)	353 (49.0)	415 (52.5)	266 (54.8)	153 (50.0)	605 (54.1)	588 (45.0)
	1.0		2348 (49.7)	368 (51.0)	375 (47.5)	219 (45.2)	153 (50.0)	513 (45.9)	720 (55.0)
First_Tpx, n (%)	False	0	753 (15.9)	121 (16.8)	130 (16.5)	75 (15.5)	45 (14.7)	164 (14.7)	218 (16.7)
	True		3975 (84.1)	600 (83.2)	660 (83.5)	410 (84.5)	261 (85.3)	954 (85.3)	1090 (83.3)
ethnicity, n (%)	African	457	160 (3.7)	20 (3.1)	42 (5.9)	43 (9.6)	6 (2.0)	20 (2.0)	29 (2.5)
	Asian		216 (5.1)	23 (3.5)	30 (4.2)	38 (8.5)	8 (2.7)	40 (4.0)	77 (6.6)
	Caucasian		3807 (89.1)	604 (92.8)	635 (89.1)	333 (74.3)	282 (94.9)	938 (94.0)	1015 (87.2)
	Other		88 (2.1)	4 (0.6)	6 (0.8)	34 (7.6)	1 (0.3)		43 (3.7)
abocomp, n (%)	No	0	271 (5.7)	48 (6.7)		39 (8.0)	20 (6.5)	102 (9.1)	62 (4.7)
	Yes		4457 (94.3)	673 (93.3)	790 (100.0)	446 (92.0)	286 (93.5)	1016 (90.9)	1246 (95.3)
past_pregnancy, n (%)	0.0	3081	702 (42.6)	147 (55.3)	119 (43.8)	51 (30.0)	27 (27.8)	124 (36.4)	234 (46.7)
	1.0		945 (57.4)	119 (44.7)	153 (56.2)	119 (70.0)	70 (72.2)	217 (63.6)	267 (53.3)
ec_transfusion, n (%)	0.0	1244	2244 (64.4)	307 (62.5)	426 (63.5)	147 (67.1)	181 (61.8)	616 (68.9)	567 (61.9)
	1.0		1240 (35.6)	184 (37.5)	245 (36.5)	72 (32.9)	112 (38.2)	278 (31.1)	349 (38.1)
dsaclass_agg_x, n (%)	I	3743	271 (27.5)	25 (24.8)	28 (28.0)	31 (34.1)	13 (22.0)	80 (41.2)	94 (21.4)
	I+II		160 (16.2)	22 (21.8)	12 (12.0)	15 (16.5)	11 (18.6)	27 (13.9)	73 (16.6)
	II		554 (56.2)	54 (53.5)	60 (60.0)	45 (49.5)	35 (59.3)	87 (44.8)	273 (62.0)

Supp. Table 2: Recipient and Donor Pre-Transplant (pre-selected) data summary statistics.

	AUROC					AUPRC				
Sub-Outcome	xgb	tabpfn	Δ	q	Sig	xgb	tabpfn	Δ	q	Sig
death_y1	0.653	0.587	-0.066	0.081	No	0.044	0.034	-0.010	0.455	No
death_y2	0.620	0.578	-0.042	0.000	Yes	0.075	0.074	-0.001	0.925	No
death_y3	0.683	0.628	-0.055	0.000	Yes	0.131	0.117	-0.014	0.533	No
death_y4	0.740	0.730	-0.010	0.187	No	0.220	0.216	-0.004	0.823	No
death_y5	0.743	0.757	0.014	0.102	No	0.327	0.318	-0.009	0.726	No
death_y6	0.793	0.792	-0.001	0.818	No	0.526	0.525	-0.001	0.925	No
death_y7	0.807	0.797	-0.009	0.022	Yes	0.607	0.599	-0.009	0.455	No
death_y8	0.815	0.803	-0.012	0.102	No	0.665	0.650	-0.014	0.455	No
death_y9	0.836	0.837	0.000	0.920	No	0.743	0.744	0.001	0.925	No
death_y10	0.848	0.842	-0.005	0.238	No	0.803	0.798	-0.005	0.756	No
death_y11	0.868	0.871	0.003	0.448	No	0.870	0.873	0.003	0.756	No
death_y12	0.891	0.893	0.001	0.818	No	0.928	0.928	0.000	0.925	No
death_y13	0.898	0.892	-0.006	0.102	No	0.953	0.948	-0.005	0.455	No

Supp. Table 3: XGBoost vs TabPFN on death risk across 13 horizons. Δ is the metric difference (TabPFN−XGBoost; positive favors TabPFN); q is the Benjamini–Hochberg FDR-adjusted p-value (across horizons); *Sig* marks whether $q < 0.05$.

	AUROC					AUPRC				
Sub-Outcome	xgb	tabpfn	Δ	q	Sig	xgb	tabpfn	Δ	q	Sig
graft_loss_y1	0.587	0.552	-0.035	0.091	No	0.078	0.059	-0.018	0.433	No
graft_loss_y2	0.624	0.578	-0.046	0.091	No	0.146	0.129	-0.017	0.743	No
graft_loss_y3	0.597	0.618	0.021	0.141	No	0.128	0.131	0.003	0.827	No
graft_loss_y4	0.672	0.647	-0.025	0.091	No	0.207	0.196	-0.011	0.743	No
graft_loss_y5	0.651	0.656	0.005	0.812	No	0.243	0.249	0.007	0.827	No
graft_loss_y6	0.701	0.708	0.006	0.774	No	0.436	0.439	0.003	0.830	No
graft_loss_y7	0.713	0.714	0.001	0.935	No	0.509	0.506	-0.003	0.830	No
graft_loss_y8	0.732	0.745	0.013	0.390	No	0.558	0.570	0.012	0.743	No
graft_loss_y9	0.786	0.777	-0.009	0.448	No	0.664	0.659	-0.005	0.827	No
graft_loss_y10	0.786	0.805	0.019	0.091	No	0.720	0.739	0.019	0.228	No
graft_loss_y11	0.813	0.804	-0.009	0.448	No	0.795	0.783	-0.012	0.650	No
graft_loss_y12	0.854	0.826	-0.029	0.000	Yes	0.883	0.859	-0.024	0.000	Yes
graft_loss_y13	0.854	0.857	0.003	0.774	No	0.912	0.914	0.002	0.827	No

Supp. Table 4: XGBoost vs TabPFN on graft loss risk across 13 horizons. Δ is the metric difference (TabPFN−XGBoost); q is the Benjamini–Hochberg FDR-adjusted p-value; *Sig* marks whether $q < 0.05$.

	AUROC					AUPRC				
Sub-Outcome	xgb	tabpfn	Δ	q	Sig	xgb	tabpfn	Δ	q	Sig
AMRCUM2_y1	0.656	0.645	-0.011	0.748	No	0.221	0.203	-0.017	0.899	No
AMRCUM2_y2	0.653	0.616	-0.037	0.065	No	0.206	0.203	-0.002	0.930	No
AMRCUM2_y3	0.669	0.617	-0.052	0.130	No	0.254	0.219	-0.035	0.585	No
AMRCUM2_y4	0.616	0.603	-0.013	0.656	No	0.232	0.228	-0.003	0.930	No
AMRCUM2_y5	0.662	0.620	-0.043	0.065	No	0.272	0.254	-0.018	0.899	No
AMRCUM2_y6	0.701	0.688	-0.013	0.656	No	0.405	0.399	-0.006	0.930	No
AMRCUM2_y7	0.706	0.720	0.014	0.656	No	0.489	0.491	0.003	0.930	No
AMRCUM2_y8	0.757	0.747	-0.010	0.656	No	0.531	0.512	-0.019	0.899	No
AMRCUM2_y9	0.741	0.770	0.029	0.065	No	0.543	0.572	0.029	0.899	No
AMRCUM2_y10	0.771	0.785	0.015	0.130	No	0.649	0.641	-0.008	0.910	No
AMRCUM2_y11	0.835	0.843	0.008	0.656	No	0.777	0.785	0.008	0.899	No
AMRCUM2_y12	0.869	0.872	0.003	0.860	No	0.869	0.875	0.007	0.910	No
AMRCUM2_y13	0.877	0.872	-0.005	0.656	No	0.904	0.904	0.001	0.930	No

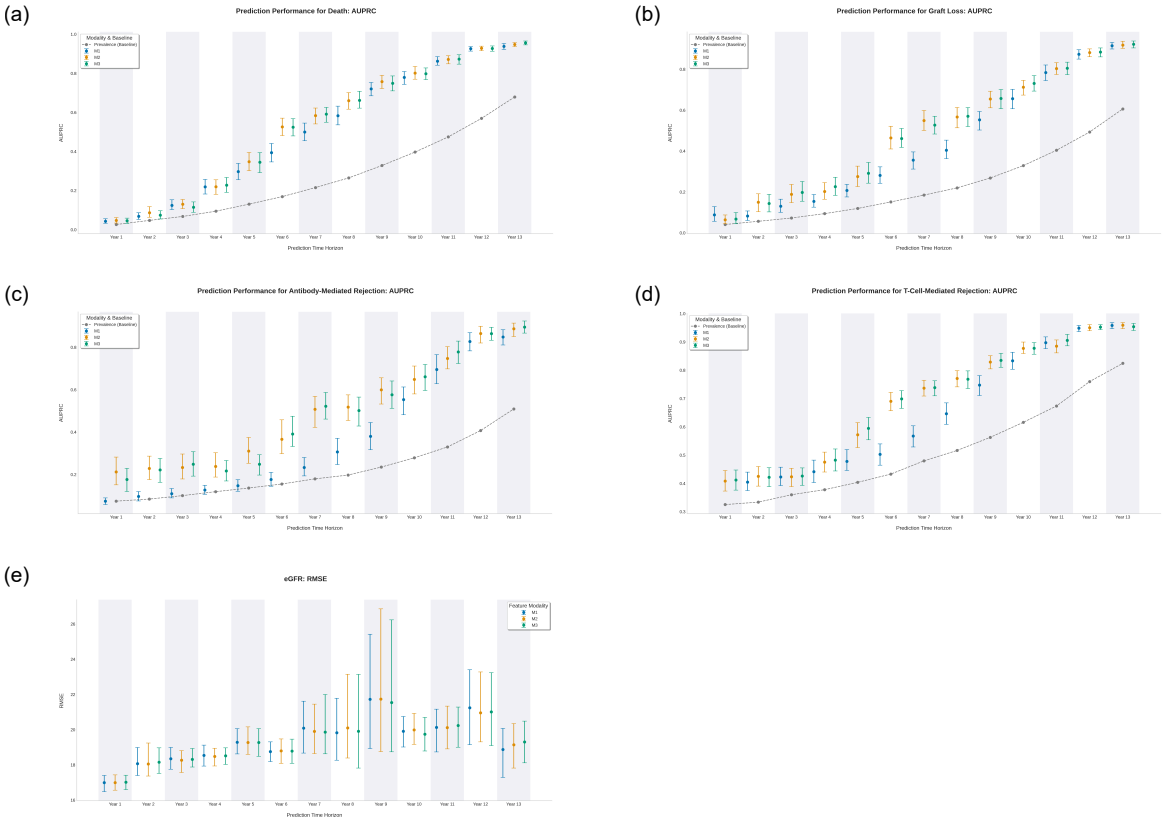
Supp. Table 5: XGBoost vs TabPFN on AMR risk across 13 horizons. Δ is the metric difference (TabPFN−XGBoost; positive favors TabPFN); q is the Benjamini–Hochberg FDR-adjusted p-value; *Sig* marks whether $q < 0.05$.

	AUROC					AUPRC				
Sub-Outcome	xgb	tabpfn	Δ	q	Sig	xgb	tabpfn	Δ	q	Sig
TCMRCUM2_y1	0.575	0.573	-0.002	0.900	No	0.409	0.405	-0.004	0.805	No
TCMRCUM2_y2	0.579	0.583	0.004	0.845	No	0.405	0.416	0.011	0.377	No
TCMRCUM2_y3	0.579	0.589	0.010	0.374	No	0.444	0.456	0.011	0.171	No
TCMRCUM2_y4	0.611	0.618	0.007	0.397	No	0.489	0.487	-0.001	0.805	No
TCMRCUM2_y5	0.673	0.653	-0.021	0.169	No	0.610	0.578	-0.032	0.016	Yes
TCMRCUM2_y6	0.688	0.678	-0.010	0.374	No	0.690	0.675	-0.015	0.039	Yes
TCMRCUM2_y7	0.668	0.669	0.001	0.900	No	0.717	0.710	-0.007	0.420	No
TCMRCUM2_y8	0.725	0.720	-0.006	0.728	No	0.788	0.776	-0.012	0.209	No
TCMRCUM2_y9	0.733	0.747	0.014	0.184	No	0.796	0.819	0.022	0.016	Yes
TCMRCUM2_y10	0.770	0.783	0.014	0.065	No	0.855	0.875	0.020	0.000	Yes
TCMRCUM2_y11	0.810	0.829	0.019	0.065	No	0.906	0.918	0.013	0.054	No
TCMRCUM2_y12	0.818	0.837	0.019	0.087	No	0.940	0.950	0.010	0.000	Yes
TCMRCUM2_y13	0.854	0.833	-0.021	0.169	No	0.968	0.963	-0.005	0.084	No

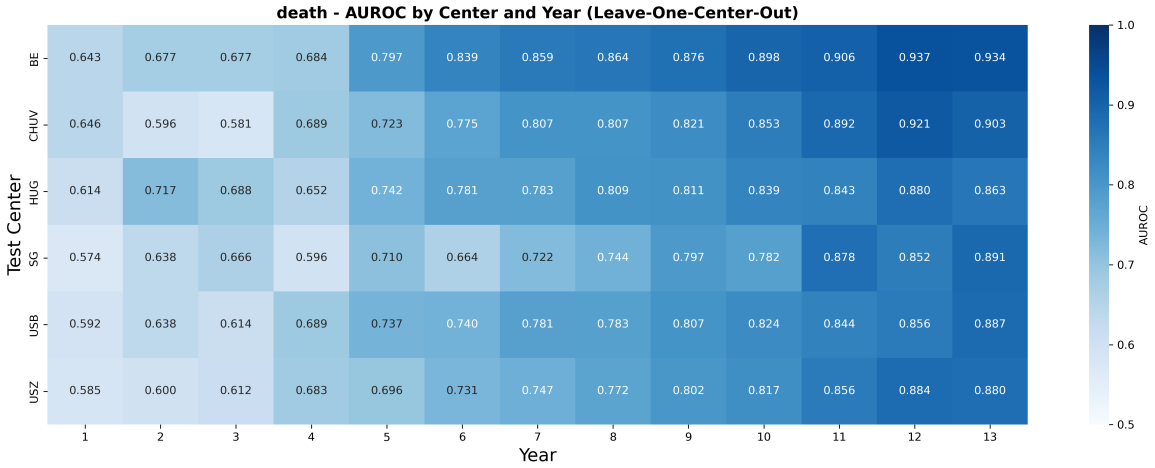
Supp. Table 6: XGBoost vs TabPFN on TCMR risk across 13 horizons. Δ is the metric difference (TabPFN−XGBoost); q is the Benjamini–Hochberg FDR-adjusted p-value; *Sig* marks whether $q < 0.05$.

	R^2					RMSE				
Sub-Outcome	xgb	tabpfn	Δ	q	Sig	xgb	tabpfn	Δ	q	Sig
eGFR_y1	0.484	0.464	-0.020	0.000	Yes	17.031	17.361	0.330	0.000	Yes
eGFR_y2	0.436	0.423	-0.013	0.014	Yes	18.201	18.406	0.205	0.016	Yes
eGFR_y3	0.401	0.384	-0.018	0.000	Yes	18.309	18.579	0.270	0.000	Yes
eGFR_y4	0.372	0.355	-0.018	0.009	Yes	18.528	18.789	0.261	0.016	Yes
eGFR_y5	0.348	0.308	-0.039	0.000	Yes	19.219	19.790	0.571	0.000	Yes
eGFR_y6	0.320	0.296	-0.024	0.014	Yes	18.804	19.138	0.334	0.019	Yes
eGFR_y7	0.275	0.232	-0.043	0.000	Yes	19.941	20.525	0.584	0.000	Yes
eGFR_y8	0.248	0.228	-0.020	0.035	Yes	20.062	20.323	0.261	0.035	Yes
eGFR_y9	0.179	0.127	-0.052	0.033	Yes	21.493	22.161	0.667	0.016	Yes
eGFR_y10	0.243	0.203	-0.040	0.009	Yes	19.769	20.283	0.513	0.019	Yes
eGFR_y11	0.187	0.127	-0.060	0.000	Yes	20.222	20.952	0.729	0.000	Yes
eGFR_y12	0.186	0.148	-0.037	0.065	No	21.181	21.663	0.482	0.076	No
eGFR_y13	0.190	0.159	-0.031	0.240	No	19.289	19.651	0.362	0.185	No

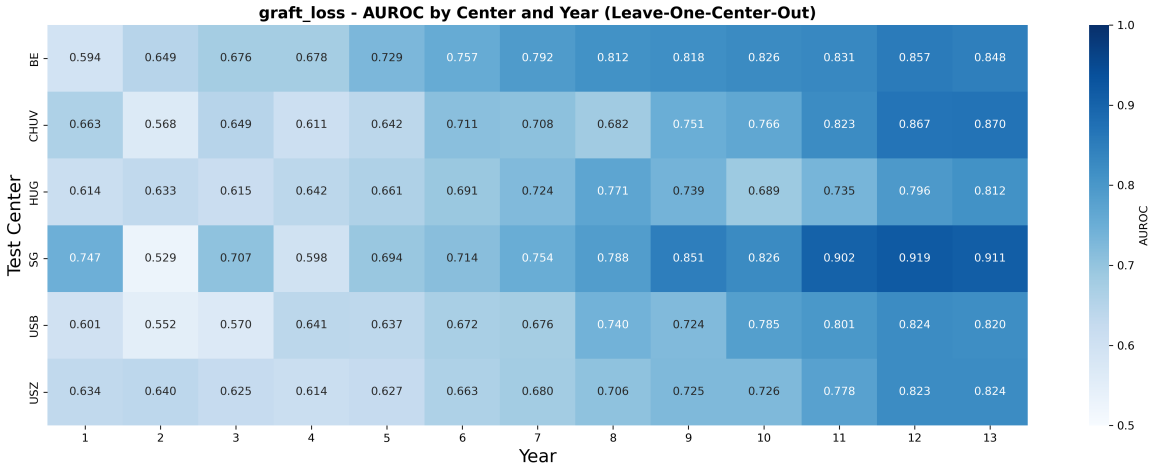
Supp. Table 7: XGBoost vs TabPFN on eGFR regression across 13 horizons. Δ is the metric difference (TabPFN−XGBoost); q is the Benjamini–Hochberg FDR-adjusted p-value; *Sig* marks whether $q < 0.05$.



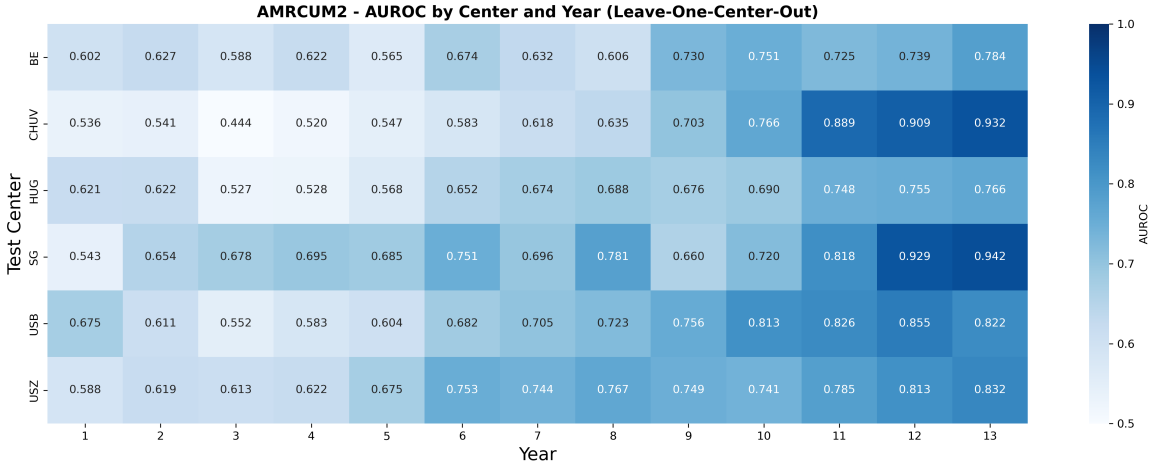
Supp. Fig. 1: **Other prediction performance and feature attributions across five outcomes and 13 horizons (XGBoost). (a):** death; **(b):** graft loss; **(c):** AMR; **(D):** TCMR; **(E):** eGFR.



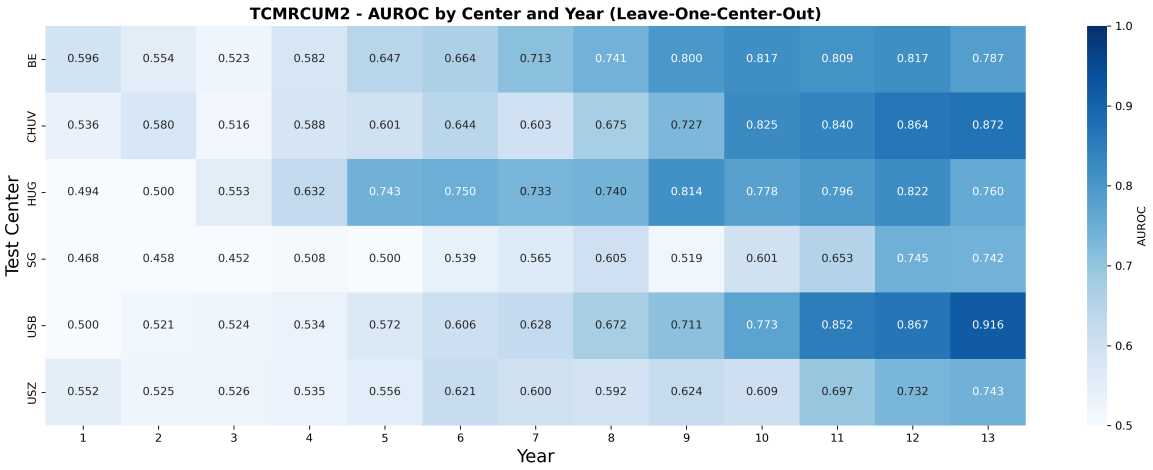
Supp. Fig. 2: **Heatmap comparison of AUROC from the leave-one-center-out validation experiment for death prediction.**



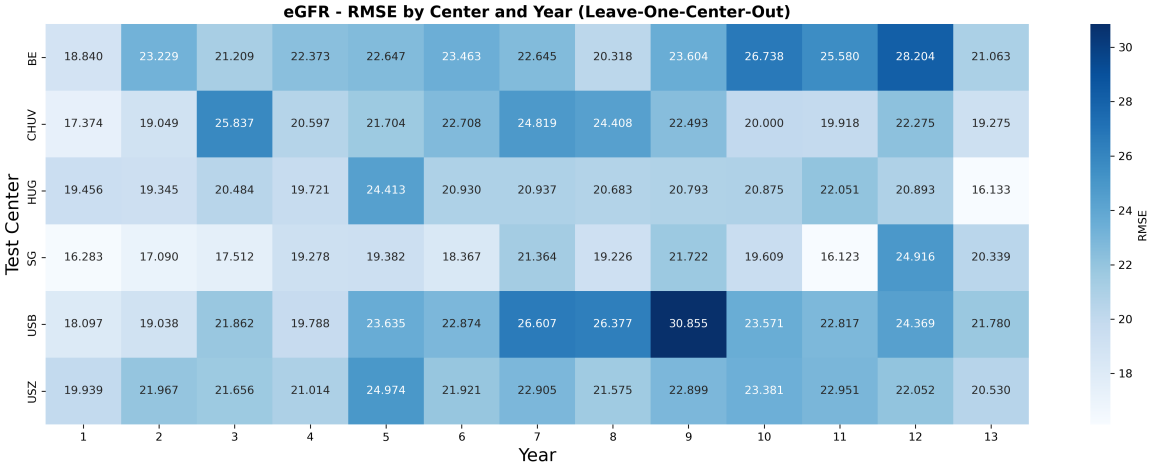
Supp. Fig. 3: **Heatmap comparison of AUROC from leave one center out validation experiment for graft loss prediction.**



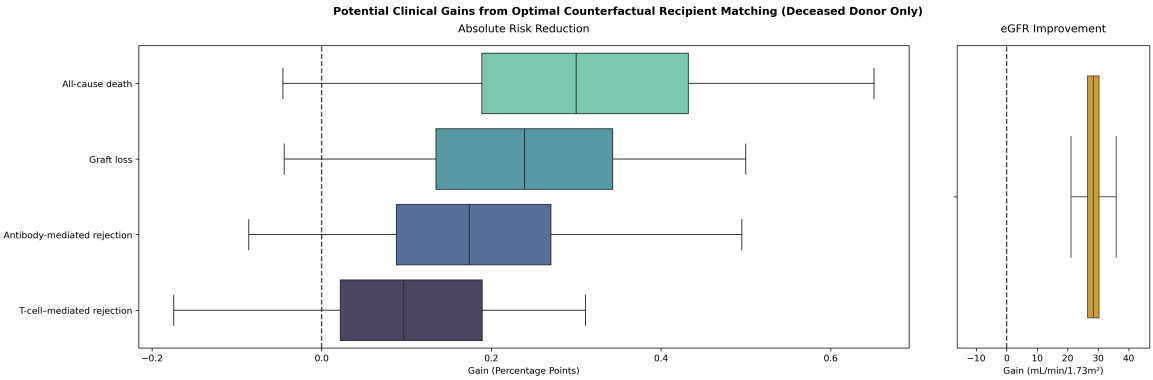
Supp. Fig. 4: **Heatmap comparison of AUROC from leave one center out validation experiment for AMR prediction.**



Supp. Fig. 5: **Heatmap comparison of AUROC from leave one center out validation experiment for TCMR prediction.**



Supp. Fig. 6: **Heatmap comparison of RMSE from leave one center out validation experiment for eGFR prediction.**



Supp. Fig. 7: **Potential clinical gains from optimal counterfactual recipient matching using deceased donors only.** Distribution of per-donor gain when the counterfactual recipient matching procedure is restricted to deceased donors, shown for absolute risk reduction (all-cause death, graft loss, antibody-mediated rejection, and T-cell–mediated rejection) and eGFR improvement.