Research Square

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

Competing risks in machine-learning prediction of kidney graft loss: a benchmark of 14 survival models in 46,559 grafts

Abidur Rahman
Université de Limoges
Marouane Tliba
Université de Limoges
Pierre Marquet
pierre.marquet@unilim.fr

CHU Limoges

Article

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11652048/v1

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

Additional Declarations: Competing interest reported. The DDMoKiT project is funded by Sanofi. PM has received honoraria and/or research funding from Chiesi, Sandoz, Pfizer, Takeda, Womed and Medincell. AR and MT declare no competing interests.

Competing risks in machine-learning prediction of kidney graft loss: a benchmark of 14 survival models in 46,559 grafts

Abidur Rahman[1], Marouane Tliba[1], Pierre Marquet[1,2*]

[1*]Pharmacology & Transplantation (P&T), UMR 1248 INSERM, Université de Limoges, Limoges, France.
[2]Department of Pharmacology, Toxicology and Pharmacovigilance, CHU Limoges, Limoges, France.

*Corresponding author(s). E-mail(s): pierre.marquet@unilim.fr;

Abstract

Prediction models for kidney graft loss often censor death with a functioning graft, which implicitly estimates graft loss in a population where no one dies. We compared single-event and competing-risks models and asked whether standard performance measures detect the resulting bias. Using 46,559 adult kidney grafts transplanted in France, 2005–2020 (CRISTAL registry), we trained nine single-event and five competing-risks models, 11 of them deep learning, on 32,609 grafts with 36 predictors available at discharge, and evaluated them on 13,950 held-out grafts at 1, 3, 5 and 10 years against the Aalen–Johansen cumulative incidence. Mean concordance was 0.715 for single-event and 0.716 for competing-risks models; the families differed only at 10 years (concordance **+0.017**), and net benefit was similar. All nine single-event models over-predicted graft loss (decile bias **+0.013** to **+0.071**), whereas competing-risks models were nearly unbiased (**+0.002** to **+0.007**). At 10 years, the median single-event model predicted 21.6% versus 18.5% observed (competing-risks family, 18.0%). Censoring death inflates the predicted risk of graft loss, yet concordance, index of prediction accuracy and net benefit miss it. Competing-risks models remove most of the bias without losing discrimination. Validation should include calibration against the cumulative incidence function.

Keywords: kidney transplantation; graft loss; competing risks; machine learning; calibration; prediction model

1 Introduction

After kidney transplantation, a graft can be lost in two ways that matter clinically: through return to dialysis or retransplantation, and through the death of the recipient while the graft is still functioning. The two events compete (Figure 1a). A recipient who dies with a functioning graft can no longer lose it, and deaths with a functioning graft are common: in registry and cohort studies they accounted for about 40% of all graft losses [1, 2].

A growing number of prediction models for graft loss have been developed, from regression-based scores [3] to machine-learning survival models [4–6]. Many analyse death-censored graft survival: recipients who die with a functioning graft are treated as censored, and the complement of the Kaplan–Meier estimate [7], or its model-based analogue $1 - S(t \mid x)$, is reported as the risk of graft loss. Under competing risks this quantity is not the probability that a recipient will lose the graft: it describes a hypothetical population in which death cannot occur, and it exceeds the cumulative incidence, increasingly so as the competing event accumulates [8–12]. The appropriate estimand is the cumulative incidence function, estimated by the Aalen–Johansen estimator [13] and modelled by competing-risks regression [14, 15].

The consequences for prediction models are less well characterised. Theory implies that predicted probabilities will be inflated, but not by how much in a contemporary transplant cohort. It is also unclear whether flexible machine-learning models trained on the censored outcome inherit, amplify or partly absorb this error. Nor is it known whether the performance measures that dominate the machine-learning literature would reveal it. Concordance depends only on the ordering of predicted risks, so a bias that preserves the ordering is invisible to it [16, 17]; overall accuracy measures such as the Brier score average errors over a population in which most recipients have a low absolute risk [18, 19]. Guidance on validating prediction models under competing risks exists [20, 21], but deep-learning survival models are still mostly reported with concordance alone.

We compared single-event and competing-risks survival models, including deep-learning models, for the prediction of graft loss in a national cohort of 46,559 kidney grafts, all trained on the same data and evaluated on the same held-out grafts with a single set of competing-risks performance measures, so that differences between the families reflect the formulation rather than the evaluation. We asked whether the families differ in discrimination, calibration and clinical utility, and how much of any calibration difference is attributable to the estimand rather than to the models themselves (Figure 1).

2 Methods

2.1 Data source and study population

We used the CRISTAL data of REXETRIS, a nationwide research programme on kidney transplantation led by CHU Limoges with INSERM UMR 1248 within the French Health Data Hub, which links pseudonymised data from CRISTAL, the national organ transplantation registry of the Agence de la Biomédecine, with other national

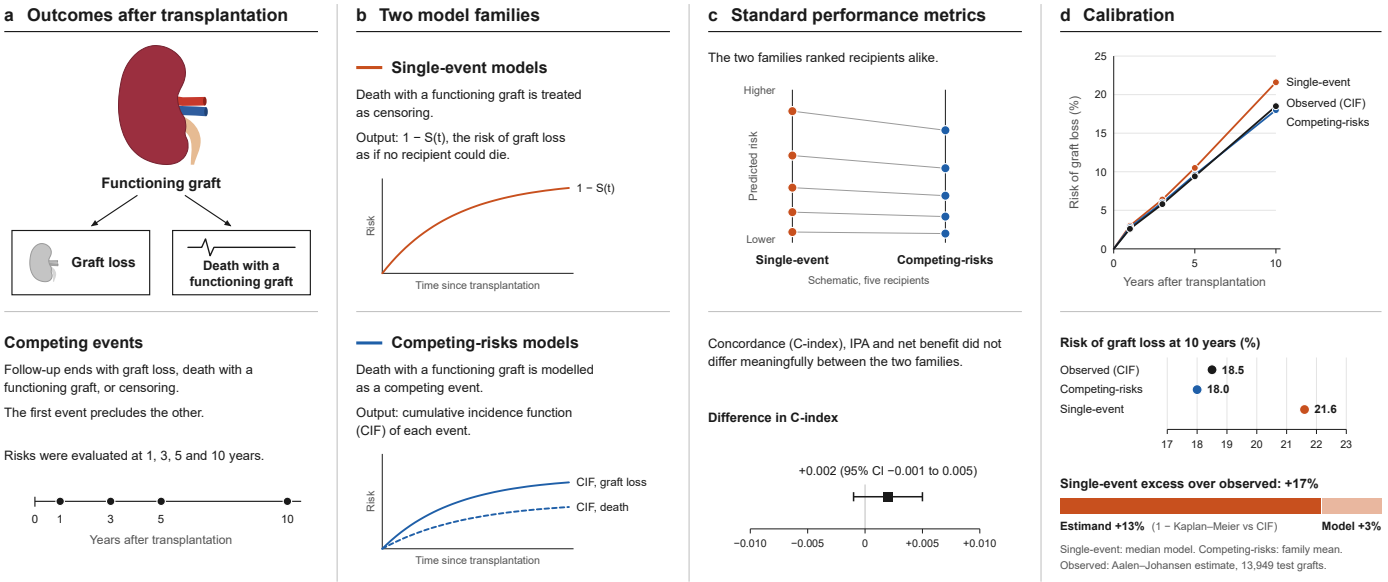


Fig. 1: Study overview and main findings. (a) After kidney transplantation, a graft is lost (return to dialysis or retransplantation) or the recipient dies with a functioning graft; whichever occurs first precludes the other, and grafts with neither event are censored. Risks were evaluated at 1, 3, 5 and 10 years. (b) Single-event models treat death as censoring and estimate $1 - S(t)$, the risk of graft loss as if no recipient could die; competing-risks models estimate the cumulative incidence of each event, $\mathrm{CIF}_k(t)$. (c) The two families ranked recipients alike: concordance, the index of prediction accuracy (IPA) and net benefit did not distinguish them. (d) Mean predicted risk of graft loss by horizon (median single-event model; competing-risks family mean) against the observed Aalen–Johansen cumulative incidence in the test grafts. At 10 years, the median single-event model predicted 21.6% against 18.5% observed (+17%), mostly because of the estimand (+13%: one minus Kaplan–Meier versus the cumulative incidence) rather than miscalibration of the models against their own target; the competing-risks family predicted 18.0%.

health data [22]. The extract covers all adult kidney transplants from adult donors in France from 1 January 2005 to 31 December 2020, with follow-up to 31 December 2021. The unit of analysis was the graft. Of 47,765 grafts, we excluded 1,166 lost within 7 days, most of which were not functioning at discharge from the transplant hospitalisation, when three of the predictors were recorded. We also excluded 40 grafts with a missing outcome field or non-positive follow-up time. The analysis cohort comprised 46,559 grafts in 45,177 recipients (1,366 with two or more grafts; maximum follow-up 17.0 years). The study is reported following TRIPOD+AI reporting guidelines [23] (Supplementary Table S1).

2.2 Outcomes

Two events were considered (Figure 1a): graft loss and death with a functioning graft. Graft loss was the functional arrest of the graft, dated by the transplant team in CRISTAL; when no arrest date was recorded but the recipient received another kidney transplant, graft loss was dated at that retransplantation (fewer than 11 grafts). Each graft was followed from transplantation until graft loss, death with a functioning graft, or the last recorded graft follow-up date, whichever came first. Events occurring after a retransplantation were therefore attributed only to the new graft. Events after 31 December 2021 were censored at that date.

In the competing-risks formulation, each graft was assigned the first observed event, or censoring if neither was observed; a recorded graft loss was retained regardless of any later death. In the single-event formulation, deaths with a functioning graft were treated as censoring. Deaths were ascertained from the national death registry, whether or not the graft was still functioning, so a death can be recorded after the end of graft follow-up. Of the 8,574 deaths in the whole cohort without a recorded graft loss, 90.2% occurred on the last day of graft follow-up and 9.8% after it (median 0.3 years later, interquartile range 0.1–0.8), so graft function at death was unknown for the latter. In the primary analysis, we counted these deaths as deaths with a functioning graft. In sensitivity analyses, deaths more than W days after the end of graft follow-up were instead not counted, and the graft was censored at the end of its follow-up (W from 1 to 365 days; see below).

2.3 Predictors

Thirty-six predictors were used in every model (Table 1; definitions, units and coding in Supplementary Table S2): 11 donor characteristics, 16 recipient characteristics, 6 matching and transplant characteristics and 3 variables recorded at discharge. Recipient characteristics were recorded at waiting-list registration or at transplantation; they included the time from registration to transplantation and the transplant rank, counting only grafts from 2005 onwards. The matching and transplant characteristics were the human leukocyte antigens HLA-A, -B and -DR mismatches, machine perfusion of the graft, cold ischaemia time and year of transplantation. The discharge variables, recorded at discharge from the transplant hospitalisation, were recipient serum creatinine, recipient weight and graft function. Preprocessing was fitted on the training set and applied unchanged to the test set: log-transformation of cold ischaemia time and of

donor and discharge creatinine, imputation of missing values by chained equations [24], and standardisation (Supplementary Methods). No variable selection was performed.

2.4 Data split

Recipients were split at random into a training set (70%) and a test set (30%), so that all grafts of a recipient fell in the same set, with stratification on the outcome crossed with the year of transplantation (32,609 and 13,950 grafts; Supplementary Table S10). All 14 models were evaluated on the same test grafts; one graft lacked a saved prediction and was excluded (13,949 evaluated). Deep-learning models used a stratified 15% of the training set for early stopping; no test data were used for training, tuning or preprocessing.

2.5 Models

Nine single-event models were fitted to graft loss with death treated as censoring: Cox proportional hazards [25] and random survival forest [26] (scikit-survival [27]); DeepSurv [28], Cox-Time [29], logistic-hazard [30, 31], multi-task logistic regression (MTLR) [32], probability mass function (PMF) [31] and DeepHit for a single event [33] (pycox); and Deep Survival Machines [34] (auton-survival [35]). Their predicted risk by horizon h was $1 - \hat{S}(h \mid x)$. They were not compared for death: the registry's patient-survival fields continue after graft loss and describe all-cause mortality, not death with a functioning graft (Supplementary Table S3).

Five competing-risks models were fitted, each predicting the cumulative incidence of cause k at horizon h, $\widehat{\mathrm{CIF}}_k(h \mid x)$. Two modelled both causes jointly: DeepHit with competing risks [33] and Deep Survival Machines with the competing-risks likelihood [34]. Two modelled the cause-specific hazards, with one model per cause, and combined them into cumulative incidence functions: cause-specific DeepSurv [28] and Deep Cox Mixtures [36]. The fifth was the Fine–Gray subdistribution hazard model [14], fitted separately for each cause. No Fine–Gray implementation could be installed on the analysis platform, so we used our own, validated against the Aalen–Johansen estimate (Supplementary Methods). Because the two Fine–Gray incidences are estimated separately, their sum can exceed one; we report how often it did. A covariate-free Aalen–Johansen estimate on the training set served as the reference for the index of prediction accuracy.

Hyperparameters were fixed in advance and were the same for all architectures: two hidden layers of 128 and 64 units, dropout of 0.2 and a learning rate of 10^{-3}, with early stopping where the library supported it. No hyperparameters were tuned (Supplementary Table S4). Each of the 11 deep-learning models was trained three times with different random seeds, and its predicted curves were averaged over the three runs. All results refer to these averaged predictions; run-to-run variation is reported in Supplementary Table S11. The Cox, random survival forest and Fine–Gray models were fitted once.

2.6 Performance evaluation

Every prediction of cause k at horizon h (1, 3, 5 and 10 years) was evaluated against the observed cumulative incidence of cause k, with the other cause treated as a competing event. For single-event models, $1 - \hat{S}(h \mid x)$ was evaluated as it is interpreted in practice: as the probability that the recipient experiences the event by h. Formulas are given in the Supplementary Methods. Measures requiring censoring weights used the reverse Kaplan–Meier estimate of the censoring distribution in the test grafts. *Discrimination*: fixed-horizon concordance with inverse-probability-of-censoring weights [37] and competing-risks comparability [16], and Antolini's time-dependent concordance over the whole predicted curve [38]. *Overall accuracy*: the censoring-weighted Brier score [18, 39] and the index of prediction accuracy (IPA) [19]. *Calibration*: within deciles of predicted risk, the mean prediction was compared with the Aalen–Johansen cumulative incidence [20, 21]. Differences were summarised over deciles and horizons as the mean absolute difference (calibration error) and the mean signed difference (bias; positive for over-prediction). The mean predicted risk was also compared with the pooled Aalen–Johansen estimate. For single-event models, its relative excess was split into an *estimand* component, the ratio of one minus the death-censored Kaplan–Meier estimate to the cumulative incidence, which every single-event model inherits, and a *model* component, the miscalibration against the model's own target. *Clinical utility*: decision curves with censoring-adjusted net benefit [40] at thresholds from 1% to 50% at 5 and 10 years, compared with two default strategies: classifying every recipient as high risk, as if all grafts were expected to be lost (treat all), and classifying no recipient as high risk (treat none).

The evaluation was repeated within subgroups: era of transplantation (2005–2008, 2009–2012, 2013–2016 and 2017–2020), transplant rank (first graft since 2005 versus second or later), recipient age at transplantation (18–44, 45–59, 60–69 and 70 years or more) and the four combinations of donor and recipient sex. Horizons at which fewer than 50 grafts remained at risk were omitted.

2.7 Statistical uncertainty

Confidence intervals (95%) were obtained from 200 bootstrap replicates of the test set. Each replicate resampled recipients, with all of their grafts, and the same replicates were used for every model, so differences between models and between families are paired. For each model, we report concordance and IPA averaged over horizons. For each comparison, we report two differences, both as competing-risks minus single-event: the difference between the family means, and the difference between the best model in each family. Calibration measures, Antolini's concordance and subgroup results are reported as point estimates.

2.8 Sensitivity analysis

Under the alternative definition, a death occurring more than W days after the end of graft follow-up was not counted as an event; the graft was censored at the end of its follow-up. We used $W = 365$, 90, 30 and 1 day. For each window, all competing-risk models were retrained, and every model was re-evaluated; the observed incidences

and the estimand component were also recomputed (Supplementary Table S7). The single-event graft-loss models were unaffected, because their outcome does not use the death fields.

2.9 Software and reproducibility

Analyses were performed in a dedicated project space on the technological platform of the French Health Data Hub (Plateforme des données de santé), under the security framework of the national health data system (SNDS) and without internet access, in Python (versions in the Supplementary Methods). Every measure was computed for every model from the saved test-set predictions with a single implementation. Only aggregate results were exported, with counts below 11 suppressed. The code is available from the authors on request.

2.10 Ethics and data access

REXETRIS was reviewed by the Comité d'expertise pour les recherches, les études et les évaluations dans le domaine de la santé (CEREES; the French national expert committee for research, studies and evaluations in health) and authorised by the Commission Nationale de l'Informatique et des Libertés (CNIL; the French data protection authority; decision DR-2020-359 [22]; authorisation 920315 of 20 November 2020). It is registered by the Health Data Hub (T11068872019112). CHU Limoges is the data controller of this research in the public interest. The data are pseudonymised, and the correspondence tables between codes and identities are held separately by CHU Limoges, the Agence de la Biomédecine and the national health insurance fund (CNAM). The study was conducted in accordance with the relevant guidelines and regulations. Individual informed consent was not required for this research under French law: owing to the population size and time depth, the CNIL authorisation provided for informing patients through their transplant physicians and the websites of CHU Limoges and the Health Data Hub, rather than individually; patients may object to the use of their data without consequence for their care [22]. The present analysis falls within the authorised purposes of REXETRIS.

2.11 Use of AI tools

During the preparation of this work, the authors used Claude (Anthropic) to assist with drafting and editing the manuscript text and with writing analysis code. Following use of this tool, the authors reviewed and edited the content as necessary and accept full responsibility for the final manuscript.

3 Results

3.1 Cohort and observed incidence

Of the 46,559 grafts, 32,609 formed the training set and 13,950 the test set (Table 1; Supplementary Table S9). Recipients had a median age of 53 years (interquartile range 42–63), 37.1% were women and 20.2% had diabetes at registration on the waiting list;

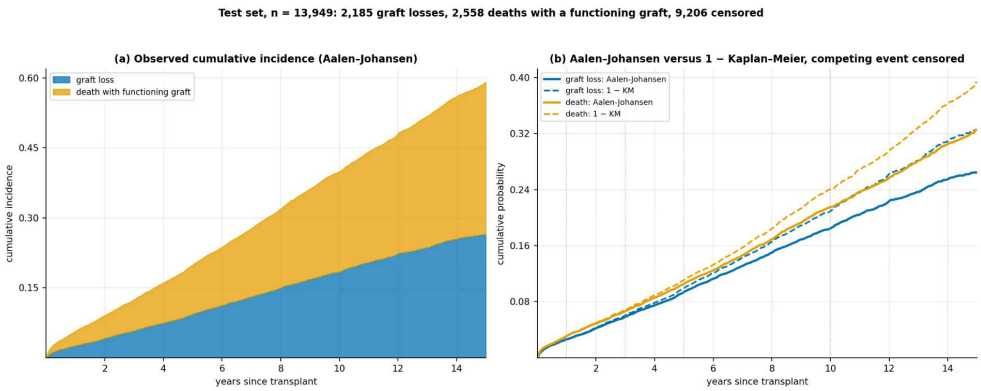


Fig. 2: Observed cumulative incidence of graft loss and of death with a functioning graft in the test grafts. (a) Aalen–Johansen estimates, stacked. (b) For each cause, the Aalen–Johansen cumulative incidence (solid) and one minus the Kaplan–Meier estimate with the competing event censored (dashed), the quantity a single-event model estimates. Dotted vertical lines mark the evaluation horizons; curves stop where fewer than 50 grafts remain at risk.

donors had a median age of 54 years (43–65) and 12.4% were living donors. The median cold ischaemia time was 15 hours (11–20), 25.0% of grafts were machine perfused before transplantation, and the median year of transplantation was 2013 (2009–2017). At discharge from the transplant hospitalisation, 97.2% of grafts were functioning. Second or later grafts since 2005 made up 4.1% of the cohort. These characteristics agree with the published description of the full CRISTAL extract (62.7% men, 12.4% living donors) [41]. Values were missing for 25–32% of grafts for the recipient comorbidities, 21.9% for recipient diabetes, 25.2% for machine perfusion and 15–18% for the discharge variables.

The stratified split reproduced the outcome distribution and the transplant years: in both sets 15.7% of grafts were lost, 18.3–18.4% ended with the death of the recipient with a functioning graft and 66% were censored, and the test set held 29.8–30.4% of the grafts of every transplant year (Supplementary Table S10). Deaths with a functioning graft outnumbered graft losses (2,559 versus 2,185 in the test set).

The observed cumulative incidence of graft loss in the test set was 2.6%, 5.8%, 9.4% and 18.5% at 1, 3, 5 and 10 years, and that of death with a functioning graft 3.1%, 6.7%, 10.5% and 21.5% (Figure 2a). One minus the Kaplan–Meier estimate with the competing event censored exceeded the cumulative incidence for both causes, by an amount that grew with the horizon: for graft loss 2.6%, 6.1%, 9.9% and 21.0% (relative excess 1.6%, 3.6%, 6.0% and 13.4%), and for death 3.1%, 6.9%, 11.1% and 24.0% (1.3%, 3.0%, 5.1% and 11.9%) (Figure 2b).

3.2 Discrimination

For graft loss, the mean concordance over the four horizons ranged from 0.694 to 0.725 among single-event models and from 0.705 to 0.724 among competing-risks models (Table 2, Figure 3; Supplementary Figure S1 for cause-specific comparability).

Table 1: Baseline characteristics of the training and test grafts. Continuous variables: median [interquartile range]; binary variables: n (%). Counts between 1 and 10 are suppressed. Dialysis vintage: time from registration on the waiting list to transplantation. Missing: percentage of all grafts with a missing value; missing values were imputed for modelling (Methods).

Characteristic	Training ($n = 32{,}609$)	Test ($n = 13{,}950$)	Missing (%)
Donor			
Donor age, years	54 [43–65]	54 [43–64]	0
Female donor	14,476 (44.4%)	6,129 (43.9%)	0
Donor height, cm	170 [163–176]	170 [163–176]	0.2
Donor weight, kg	72.5 [63.0–83.0]	72.0 [63.0–83.3]	0.2
Donor diabetes	2,239 (7.1%)	943 (7.0%)	3.3
Donor hypertension	8,975 (27.9%)	3,867 (28.1%)	1.3
Donor CMV seropositive	17,186 (53.6%)	7,295 (53.2%)	1.7
Donor EBV seropositive	31,056 (96.9%)	13,298 (97.0%)	1.7
Donor kidney disease	810 (2.5%)	327 (2.4%)	2.1
Donor creatinine, μmol/L	74.0 [58.0–97.4]	74.0 [58.0–97.0]	1.8
Living donor	4,042 (12.4%)	1,737 (12.5%)	0
Recipient			
Recipient age at transplant, years	53 [42–63]	53 [42–62]	0
Female recipient	12,104 (37.1%)	5,185 (37.2%)	0
Recipient height, cm	170 [162–175]	170 [163–176]	5.2
Recipient weight, kg	70.0 [60.0–81.0]	70.0 [60.0–81.0]	4.0
Recipient diabetes	5,198 (20.4%)	2,133 (19.6%)	21.9
Recipient history of hypertension	16,950 (70.1%)	7,199 (69.8%)	25.9
Recipient CMV seropositive	20,806 (64.1%)	8,949 (64.5%)	0.5
Recipient EBV seropositive	31,340 (97.6%)	13,372 (97.3%)	1.5
Recipient ever-smoker	10,839 (48.8%)	4,577 (48.3%)	31.9
Recipient cardiovascular disease	4,431 (18.8%)	1,850 (18.3%)	27.6
Other disease: dyslipidaemia	9,064 (39.7%)	3,700 (38.1%)	30.1
Other disease: cirrhosis	511 (2.1%)	195 (1.9%)	25.3
Other disease: neuropathy	1,441 (5.9%)	563 (5.4%)	24.9
Other disease: urological pathology	3,507 (14.3%)	1,440 (13.8%)	24.8
Dialysis vintage, years	1 [1–3]	1 [1–3]	0
Second or later graft since 2005	1,346 (4.1%)	553 (4.0%)	0
Donor–recipient matching			
HLA-A mismatches	1 [1–2]	1 [1–2]	0
HLA-B mismatches	1 [1–2]	1 [1–2]	0
HLA-DR mismatches	1 [0–1]	1 [0–1]	0
Transplant: process			
Machine perfusion of the graft	6,128 (25.1%)	2,582 (24.8%)	25.2
Cold ischaemia time, h	15.0 [11.0–19.6]	14.9 [10.8–19.7]	3.1
Year of transplant	2013 [2009–2017]	2013 [2009–2017]	0
Transplant: at discharge			
Creatinine at discharge, μmol/L	144.0 [110.0–201.0]	144.0 [110.0–202.0]	18.4
Weight at discharge, kg	71.0 [61.3–82.0]	71.3 [61.5–82.0]	16.5
Functioning graft at discharge	26,912 (97.2%)	11,489 (97.3%)	15.2

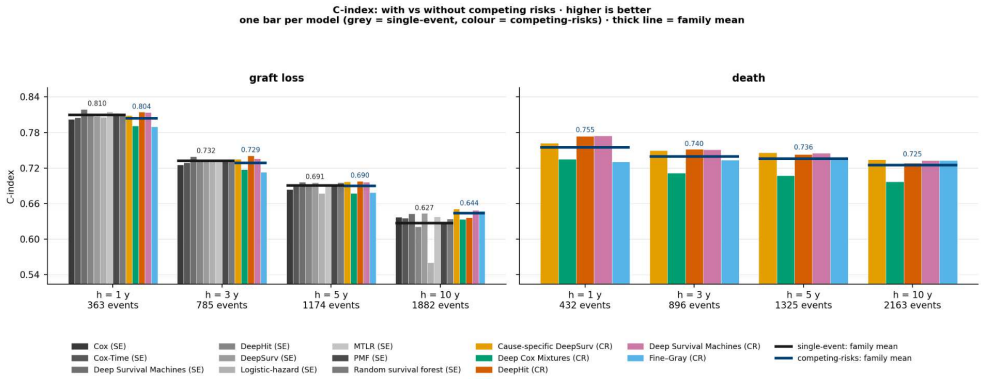


Fig. 3: Fixed-horizon concordance by model family at 1, 3, 5 and 10 years, for graft loss (left; single-event models in grey, competing-risks models in colour) and death with a functioning graft (right; competing-risks models only). One bar per model; thick horizontal lines show family means. Concordance with inverse probability of censoring weights and competing-risks comparability, on the same test grafts.

The family means were almost identical (0.715 versus 0.716; difference +0.002, 95% confidence interval or CI −0.001 to 0.005), as were the best models of each family: Deep Survival Machines in both formulations (single-event 0.725, 0.709–0.737; competing-risks 0.724, 0.708–0.737), a difference of −0.001 (−0.004 to 0.003). The small family differences depended on the horizon (Supplementary Table S5): single-event models were marginally ahead at 1 year (difference −0.006, −0.009 to −0.002), the families were indistinguishable at 3 and 5 years, and competing-risks models were ahead at 10 years (+0.017, 0.013 to 0.020; best against best +0.008, 0.004 to 0.011). In both families concordance fell with the horizon, from about 0.81 at 1 year to 0.63–0.64 at 10 years.

Antolini's time-dependent concordance, which evaluates the whole predicted curve and keeps grafts lost to the competing event as controls, separated the families more clearly: 0.642 to 0.645 for the three best competing-risks models (cause-specific DeepSurv, Deep Survival Machines and Fine–Gray) against 0.552 to 0.611 for the single-event models (Table 2; point estimates; Supplementary Figure S2). Only competing-risks DeepHit (0.582) was within the single-event range.

3.3 Calibration

The single-event models predicted a mean risk of graft loss of 2.9%, 6.3%, 10.6% and 24.2% at 1, 3, 5 and 10 years, against an observed 2.6%, 5.8%, 9.4% and 18.5% (relative excess +13%, +8%, +13% and +31%; Figure 4). The competing-risks models predicted 2.8%, 6.0%, 9.6% and 18.0% (+8%, +3%, +3% and −3%). Both families over-predicted slightly at 1 year, but the gap widened with the horizon only for the single-event models (Figure 1d). The 10-year family mean was inflated by one model: logistic-hazard predicted a mean 10-year risk of 43.4%, more than twice the observed incidence. The median single-event model predicted 21.6% (+17%), and the mean without logistic-hazard was 21.8% (+18%; Supplementary Table S6).

Table 2: Performance of the 14 models on the 13,949 evaluated test grafts. Deep-learning models: predictions averaged over three training runs. C-index and IPA: mean over the 1-, 3-, 5- and 10-year horizons with 95% bootstrap CIs (recipients resampled). Calibration error and bias: mean absolute and mean signed difference between predicted and observed (Aalen–Johansen) risk across deciles of predicted risk and horizons; positive bias indicates over-prediction. Antolini C: time-dependent concordance over the whole predicted curve (point estimate). Bold: best value in each column for each endpoint.

Model	C-index (95% CI)	IPA (95% CI)	Cal. error	Cal. bias	Antolini C
A. Graft loss					
Single-event models (death censored)					
Deep Survival Machines	**0.725** (0.709–0.737)	0.059 (0.040–0.077)	0.018	+0.017	0.608
DeepSurv	0.720 (0.703–0.732)	0.042 (0.024–0.059)	0.018	+0.015	0.611
MTLR	0.720 (0.704–0.731)	0.047 (0.021–0.068)	0.017	+0.014	0.594
Random survival forest	0.718 (0.703–0.730)	0.063 (0.047–0.080)	0.013	+0.013	0.595
PMF	0.717 (0.700–0.730)	0.052 (0.029–0.073)	0.019	+0.019	0.580
Cox-Time	0.715 (0.699–0.727)	0.052 (0.035–0.067)	0.016	+0.014	0.594
DeepHit	0.714 (0.697–0.727)	0.055 (0.047–0.062)	0.022	+0.014	0.574
Cox	0.712 (0.697–0.726)	0.048 (0.033–0.063)	0.017	+0.014	0.597
Logistic-hazard	0.694 (0.678–0.705)	0.037 (0.012–0.057)	0.074	+0.071	0.552
Competing-risks models					
Deep Survival Machines	0.724 (0.708–0.737)	**0.072** (0.057–0.086)	**0.006**	+0.002	0.644
Cause-specific DeepSurv	0.723 (0.707–0.735)	0.060 (0.046–0.072)	0.009	+0.003	0.642
DeepHit	0.722 (0.708–0.733)	0.055 (0.048–0.062)	0.020	+0.006	0.582
Fine–Gray	0.708 (0.693–0.720)	0.041 (0.031–0.050)	0.011	+0.002	**0.645**
Deep Cox Mixtures	0.705 (0.690–0.718)	0.033 (0.013–0.053)	0.027	+0.007	0.632
B. Death with a functioning graft					
Competing-risks models					
Deep Survival Machines	**0.751** (0.739–0.763)	0.110 (0.099–0.121)	0.011	+0.001	0.676
DeepHit	0.749 (0.737–0.762)	0.099 (0.091–0.106)	0.029	−0.007	0.654
Cause-specific DeepSurv	0.748 (0.736–0.760)	**0.119** (0.107–0.131)	0.008	+0.001	0.671
Fine–Gray	0.733 (0.721–0.746)	0.089 (0.079–0.099)	**0.005**	+0.001	**0.679**
Deep Cox Mixtures	0.713 (0.697–0.724)	0.066 (0.049–0.080)	0.032	+0.008	0.650

Calibration by decile of predicted risk confirmed the pattern (Figure 5, Table 2). All nine single-event models over-predicted graft loss, with mean bias from +0.013 to +0.071 (median +0.014) and calibration error from 0.013 to 0.074. For most of them the bias was almost equal to the error, so that the miscalibration was almost entirely systematic; the clearest exception was DeepHit (error 0.022, bias +0.014). The competing-risks models were close to unbiased (bias +0.002 to +0.007, median +0.003); their calibration error ranged from 0.006 for Deep Survival Machines and 0.009 for cause-specific DeepSurv to 0.020 for DeepHit and 0.027 for Deep Cox Mixtures, and for these last two the error was mostly non-directional. The best-calibrated single-event model, the random survival forest, had an error (0.013) comparable to that of Fine–Gray (0.011), but its error was entirely over-prediction (bias +0.013, against +0.002).

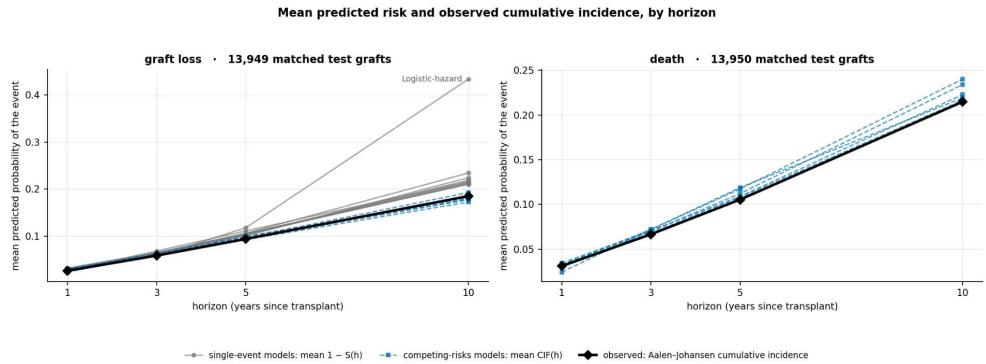


Fig. 4: Mean predicted risk and observed cumulative incidence at 1, 3, 5 and 10 years. Grey: single-event models (mean of $1 - \hat{S}(h)$), one line per model; blue: competing-risks models (mean $\widehat{\mathrm{CIF}}(h)$); black: observed Aalen–Johansen estimate.

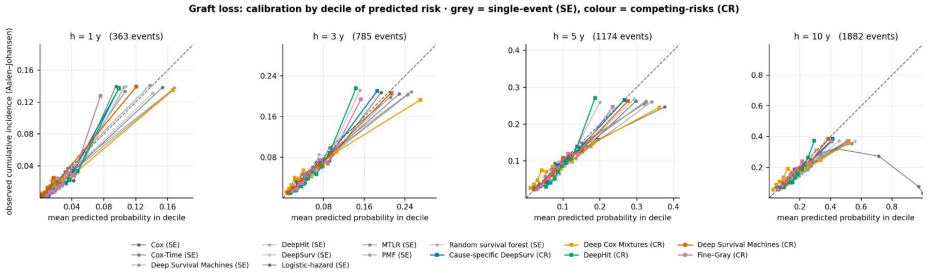


Fig. 5: Calibration of graft-loss predictions by decile of predicted risk at 1, 3, 5 and 10 years. The observed risk in each decile is the Aalen–Johansen cumulative incidence of its grafts. Grey: single-event models; colour: competing-risks models; dashed line: perfect calibration.

3.4 Estimand and model components of the single-event excess

For graft loss, the estimand component (the excess of one minus Kaplan–Meier over the cumulative incidence) grew with the horizon (+2%, +4%, +6% and +13%), as expected from the accumulation of deaths with a functioning graft, while the model component did not follow a trend (+11%, +4%, +7% and +15%; Table 3). The model component at 10 years came largely from logistic-hazard; without it, the 10-year mean prediction was 21.8%, a model component of +4%, so that the 10-year excess of the remaining single-event models (+18%) was mostly the estimand (+13%). At 1 year, the model component (+11%) was close to the over-prediction of the competing-risks models (+8%), which suggests that it partly reflects a calibration offset shared by both families rather than the single-event formulation itself.

3.5 Overall accuracy

The IPA of every model was low and fell with the horizon: the family means went from about 0.09 at 1 year to 0.004 (single-event) and 0.017 (competing-risks) at 10 years,

Table 3: Graft loss: observed cumulative incidence, one minus Kaplan–Meier with death censored, and mean predicted risk of each model family, with the single-event excess split into an estimand and a model component. The single-event mean includes logistic-hazard (43.4% at 10 years); without it, the 10-year mean was 0.218 and the model component +4%. Relative excess; components multiply: $(1 + \text{estimand})(1 + \text{model}) = 1 + \text{total}$.

	Cumulative probability by horizon				Single-event excess over observed		
Horizon	Observed (AJ)	1 − KM	Single-event	Competing-risks	Estimand	Model	Total
1 year	0.026	0.026	0.029	0.028	+2%	+11%	+13%
3 years	0.058	0.061	0.063	0.060	+4%	+4%	+8%
5 years	0.094	0.099	0.106	0.096	+6%	+7%	+13%
10 years	0.185	0.210	0.242	0.180	+13%	+15%	+31%

when three models, including logistic-hazard and competing-risks Deep Cox Mixtures, did worse than the null model (negative IPA; Supplementary Table S5; Supplementary Figures S3–S5). Over horizons, the mean IPA was 0.051 for single-event and 0.052 for competing-risks models (difference +0.002, 95% CI −0.004 to 0.010), and the mean Brier score 0.076 in both families. Overall accuracy separated the families only at 10 years, and by a small amount: IPA 0.004 against 0.017 (difference +0.013, 0.007 to 0.021) and Brier score 0.150 against 0.148 (−0.002, −0.003 to −0.001). The best model of each family by IPA was the random survival forest (0.063) and competing-risks Deep Survival Machines (0.072; difference +0.008, 0.004 to 0.013).

3.6 Clinical utility

Net benefit did not separate the families (Supplementary Figures S6 and S7; Supplementary Table S18). For graft loss at 5 years, the mean net benefit of both families exceeded that of classifying every recipient as high risk (treat all) and that of classifying none as high risk (treat none) at every threshold from 5% to 40%. At 10 years, the competing-risks family did so from 15% to 30% and the single-event family also at 3% and 10%; at lower thresholds, intervening in every recipient was about as good, because the 10-year risk was high. The single-event family had a slightly higher mean net benefit at every threshold shown, by 0.0004 to 0.0022 (for example, at a 20% threshold, 0.0079 versus 0.0064 at 5 years and 0.0225 versus 0.0220 at 10 years), that is, at most two additional net true positives per 1,000 grafts. For death with a functioning graft, the competing-risks models had a higher net benefit than both default strategies at every threshold from 3% to 50% at 5 years and from 5% to 50% at 10 years.

3.7 Subgroups

The over-prediction of the single-event models was present in every era of transplantation (Supplementary Tables S12 and S13). At the longest horizon estimable in each era, the single-event family predicted 22.6% against 20.5% observed at 10 years for

2005–2008 (competing-risks 21.4%), 22.4% against 18.8% at 10 years for 2009–2012 (19.6%), 10.2% against 9.3% at 5 years for 2013–2016 (9.3%) and 5.6% against 5.1% at 3 years for 2017–2020 (5.2%). Discrimination was weakest for grafts transplanted in 2005–2008 (5-year concordance 0.62 in both families, against 0.69–0.71 later). In the whole cohort, the 5-year cumulative incidence of death with a functioning graft rose from 8.5% for grafts transplanted in 2005–2008 to 11.1% in 2013–2016, while that of graft loss fell slightly, from 10.1% to 9.5%: the competing event became more frequent in later eras. For death with a functioning graft, the competing-risks models over-predicted for 2005–2008 (18.2% against 16.4% at 10 years) and under-predicted for 2017–2020 (8.7% against 9.6% at 3 years), consistent with this temporal trend.

For first grafts (13,396 test grafts), the results matched the whole test set; the single-event family predicted 23.8% at 10 years against 18.4% observed (competing-risks 17.9%). For the 553 second or later grafts, both families discriminated well (1-year concordance 0.85–0.86), and at 5 years the single-event family predicted 13.2% and the competing-risks family 11.8% against 12.6% observed (Supplementary Table S14).

The over-prediction of graft loss grew with recipient age, as did the competing event (Supplementary Table S15). The 10-year cumulative incidence of death with a functioning graft rose from 6.4% for recipients aged 18–44 to 54.6% for those aged 70 or more, and the 10-year prediction of the single-event family exceeded the observed incidence of graft loss by 13% (22.3% against 19.7%) at 18–44, 32% (21.9% against 16.6%) at 45–59, 34% (26.9% against 20.0%) at 60–69 and 67% (32.1% against 19.2%) at 70 or more. The competing-risks family stayed within 1.3 percentage points of the observed incidence in every age group (18.6%, 16.8%, 18.7% and 19.4%). The advantage of the competing-risks family in 10-year concordance also grew with age, from +0.008 at 18–44 to +0.029 at 70 or more, where the 10-year IPA of the single-event family was negative (−0.064, against +0.038). The competing-risks models under-predicted death with a functioning graft for recipients aged 70 or more (50.6% against 54.6% at 10 years) and over-predicted it for those aged 18–44 (7.6% against 6.4%). The results did not depend on the sex of the donor or of the recipient (Supplementary Table S16): at 10 years the single-event family over-predicted graft loss by 26% to 33% in each of the four combinations, and the competing-risks family was within 1 percentage point of the observed incidence; discrimination was lowest for grafts from female donors to female recipients (1-year concordance 0.73–0.74, against 0.79–0.84 for the other combinations).

3.8 Run-to-run variation of the deep-learning models

Across the three training runs, the mean concordance of ten of the 11 deep-learning models varied by at most 0.011 and their calibration bias by at most 0.019 (Supplementary Table S11). The exception was Deep Cox Mixtures, whose concordance for graft loss ranged from 0.674 to 0.718 across the three runs and whose IPA ranged from −0.067 to 0.062. Averaging the three runs raised the concordance of every deep-learning model above the mean of its single runs, most markedly Deep Cox Mixtures (graft-loss concordance 0.705 against 0.689, IPA 0.033 against −0.021). The single-event over-prediction was present in every run of every single-event model.

3.9 Sensitivity to the outcome coding

The alternative definition with $W = 365$ reclassified 38 of the 2,559 deaths with a functioning graft in the test set as censored. In the test set, the observed cumulative incidence of graft loss and the estimand component were virtually unchanged (10-year excess +13.4% under the alternative definition versus +13.2% under the primary). After retraining, the competing-risks models other than Deep Cox Mixtures changed by at most 0.001 in concordance, 0.006 in IPA and 0.009 in calibration bias, all within their run-to-run variation. Deep Cox Mixtures changed more (graft-loss IPA changes from −0.058 to +0.063 across horizons), but by about as much as its own run-to-run variation (Supplementary Table S8). In the whole cohort, shortening the window from 365 days to 90, 30 and 1 day lowered the 10-year estimand component from +13.7% to +13.0%, +12.7% and +12.4% (Supplementary Table S7).

3.10 Death with a functioning graft

Among competing-risks models, concordance for death with a functioning graft ranged from 0.713 to 0.751 and Antolini's concordance from 0.650 to 0.679 (Table 2, Figure 3). Calibration bias ranged from −0.007 to +0.008; Fine–Gray and cause-specific DeepSurv were the best calibrated (error 0.005 and 0.008), and DeepHit under-predicted slightly (bias −0.007; Supplementary Figures S8–S10). The pooled mean predicted risk was 3.1%, 7.0%, 11.3% and 22.7% at 1, 3, 5 and 10 years, against 3.1%, 6.7%, 10.5% and 21.5% observed (0%, +6%, +7% and +5%).

No single-event comparison was made for this cause, because the registry's patient-survival fields measure a different event. Patient follow-up continued for more than 30 days after graft loss for 85.9% of the lost test grafts, and 614 recipients (28.1% of graft losses) died after losing their graft, a median of 1.9 years later (interquartile range 0.7–4.2). One minus Kaplan–Meier on these fields therefore estimated all-cause mortality (3.3%, 7.6%, 12.2% and 27.3% at 1, 3, 5 and 10 years) rather than death with a functioning graft, whose observed cumulative incidence was 3.1%, 6.7%, 10.5% and 21.5% (Supplementary Table S3).

3.11 Fine–Gray

The sum of the Fine–Gray cumulative incidences of the two causes never exceeded one at 1 or 3 years; it exceeded one for fewer than 11 test grafts ($< 0.1\%$) at 5 years and for 146 (1.05%) at 10 years, with a maximum of 1.59 (Supplementary Table S17).

4 Discussion

In 46,559 kidney grafts, single-event models that censored death and competing-risks models that treated it as a competing event performed almost identically on the two measures most often reported, concordance and the index of prediction accuracy. Yet they produced clearly different probabilities. All nine single-event models over-predicted the risk of graft loss, and at 10 years the median model over-stated it by 17%. The five competing-risks models, four of them deep-learning models, were close to unbiased at every horizon and within 3% of the observed incidence at 10 years. Concordance over

the whole predicted curve and fixed-horizon concordance at 10 years did favour the competing-risks models, but up to 5 years the families ranked recipients alike, the IPA differed only at 10 years and by about 0.01, and net benefit was, if anything, slightly higher for the single-event family, by at most two net true positives per 1,000 grafts. Neither discrimination at short horizons, overall accuracy nor net benefit can therefore be taken as evidence that a graft-loss model is fit for estimating absolute risk.

Why concordance misses the problem.

Concordance depends only on the ordering of predicted risks [16, 37]. The single-event prediction and the cumulative incidence rank recipients nearly identically when competing events are rare, and exactly so when the competing hazard does not depend on the covariates and the predicted curves do not cross. Deaths with a functioning graft are rare early after transplantation, which is consistent with the two families coinciding at short horizons. As deaths with a functioning graft accumulate, recipients at high risk of dying (older, more comorbid) have a cumulative incidence of graft loss lower than their graft-loss hazard alone suggests, because many of them die first, and a single-event model ranks them too high; this is how competing-risks models gained concordance at 10 years and over the whole curve.

Why overall accuracy misses it too.

The Brier score penalises a calibration error by its square: shifting a calibrated model up by 0.014, the median single-event bias, adds about 0.0002 to a Brier score of 0.076 and about 0.003 to the IPA, less than the differences between models within each family [18, 19]. Net benefit behaved similarly, because an over-predicting model that ranks well adds true and false positives that largely offset each other. These summaries do not isolate the error that matters when a probability is communicated to a patient or compared with a threshold. Calibration against the cumulative incidence function compares predicted with observed risk on the probability scale, was sensitive to the bias at every horizon, and should accompany any transplant prediction model intended to communicate absolute risk [17, 20, 21].

Estimand and model.

Separating the single-event excess into an estimand and a model component clarifies what can be fixed. The estimand component is a property of the target: any single-event model, however flexible, estimates one minus the Kaplan–Meier estimate with death censored, which at 10 years exceeded the observed incidence by 13%; better architectures or more data cannot remove it, nor can recalibration against the model's own target. The model component was largest at 1 year, where it was close to the over-prediction of the competing-risks models themselves, so that it is partly a general calibration offset that recalibration could address; at 10 years it was small for every model except logistic-hazard.

A second pitfall: the event definition.

A second pitfall concerns the event definition. The registry's patient-survival fields continue after graft loss: 28% of recipients who lost their graft later died, a median of

1.9 years afterwards, so that a single-event model trained on these fields predicts all-cause mortality (27.3% at 10 years), not death with a functioning graft (21.5%). The death outcome of a graft-loss study should be derived from the same competing-risks coding as the graft-loss outcome, which is why we compared the two families for graft loss only.

Transplant era and recipient age.

The competing event became more frequent over the study period: the 5-year cumulative incidence of death with a functioning graft rose by about a third between grafts transplanted in 2005–2008 and in 2013–2016, while that of graft loss fell slightly. The bias of a single-event model grows with the competing event: across recipient age, where the 10-year incidence of death with a functioning graft rose from 6% to 55%, the single-event excess at 10 years rose from 13% to 67% while the competing-risks models stayed within 1.3 points of the observed incidence. The bias is therefore likely to be larger for contemporary recipients than in older cohorts, and it was present in every era we could examine. The competing-risks models, trained on all eras, over-predicted death for the earliest grafts and under-predicted it for the most recent ones: calibration also drifts with time and should be checked in the population to which a model is applied.

Individual models and training runs.

Deep Survival Machines was the most discriminating model in both formulations and, with competing risks, the best calibrated for graft loss. Fine–Gray, with a linear predictor, was well calibrated and had the highest Antolini concordance for both causes. DeepHit, trained partly on a ranking loss, had larger and less directional calibration errors [33]. Deep Cox Mixtures was the only model whose performance depended strongly on the random seed (graft-loss IPA from -0.067 to 0.062 across three runs); averaging runs stabilised it and raised the concordance of every deep-learning model, which argues for reporting deep survival models as ensembles of runs, or at least with their run-to-run variation. Among single-event models, the random survival forest was the best calibrated but still over-predicted. Logistic-hazard (43.4% at 10 years against 18.5% observed) explained most of the gap between the family mean and the median.

Comparison with previous work.

That the complement of the Kaplan–Meier estimate overstates the cumulative incidence under competing risks is well established [8–11, 42]; a recent analysis of two transplant registries found that censoring the competing event overstated the absolute risk of graft failure by up to 8.4 percentage points [43]. Our contribution is to quantify the consequence for machine-learning prediction models at registry scale, across classical and deep-learning methods trained and evaluated identically, and to show which validation measures expose it: deep survival models [33, 34] are usually reported with concordance and, increasingly, the Brier score or IPA, neither of which would have revealed that every single-event model over-stated 10-year risk.

Implications.

Models for graft loss should therefore target the cumulative incidence through a competing-risks formulation, which in our data cost at most 0.006 in fixed-horizon concordance at 1 year and gained 0.017 at 10 years, and should be validated with calibration against the Aalen–Johansen estimate at each horizon, alongside concordance and decision curves. Existing single-event scores can still rank recipients, for example to prioritise surveillance, but their probabilities should not be communicated to patients or used against absolute-risk thresholds without recalibration to the cumulative incidence.

Strengths and limitations.

The study used a large national cohort; all 14 models were trained on the same data and evaluated on the same held-out grafts with one implementation of every measure; the split was made by recipient and stratified on outcome and transplant year; uncertainty was estimated with paired bootstrap replicates that resampled recipients; and deep-learning models were averaged over three training runs, with their run-to-run variation reported.

The study has limitations. It used a single national registry: the analysis by transplant era describes performance across time within one split, and external validation in other populations is needed. Three predictors were recorded at discharge from the transplant hospitalisation, so the models predict from discharge; early graft losses that were partly known at that time raise the 1-year concordance of both families alike (about 0.81), and therefore do not bias the comparison between them, which is less affected at 3 to 10 years. Some limitations come from how the registry records data: graft and patient follow-up are recorded separately, transplant rank counts grafts since 2005 only; censoring the deaths that occurred after the end of graft follow-up, the alternative definition, did not change the results beyond run-to-run variation. Missing values, frequent for recipient comorbidities and the discharge variables, were imputed under a missing-at-random assumption. Hyperparameters were fixed a priori and identical across models rather than tuned, which favours a fair comparison over the best attainable performance of each model, and subgroups were limited to transplant era, rank, recipient age and donor and recipient sex; the transplant centre is not available in the extract. Finally, predicting outcomes up to 10 years ahead from data available around transplantation is inherently imprecise, which explains the modest concordance and the near-zero IPA of every model at 10 years; the direction of the single-event bias, however, was the same at every horizon, in every era and in every training run.

Conclusion.

Censoring death when predicting kidney graft loss inflates predicted probabilities, while short-horizon concordance and overall accuracy remain almost unchanged. Competing-risks formulations, including deep-learning models, remove most of the bias at no meaningful cost in discrimination. Calibration against the cumulative incidence function should be a standard part of validating transplant prediction models.

Data availability

Individual-level data cannot be shared: they are accessible only to the members of the research team and of the Health Data Hub involved in the study, within the secured Health Data Hub environment, and the CRISTAL extract may not be reused outside the approved research programme [22]. CRISTAL data can be requested from the Agence de la Biomédecine and SNDS data from the Assurance Maladie. Aggregate results, all figures and the evaluation code are available from the authors on request.

Funding

The REXETRIS database received funding and material support from the French Health Data Hub and CHU Limoges. The present analysis was carried out within the DDMoKiT project (Disease–Drug Modelling in Kidney Transplantation), funded by Sanofi through the Sanofi Innovative Technology Awards Program (Sanofi reference 251034, award #134129). The funder had no role in the design of the study, the analysis of the data, or the decision to publish.

Competing interests

The DDMoKiT project is funded by Sanofi. PM has received honoraria and/or research funding from Chiesi, Sandoz, Pfizer, Takeda, Womed and Medincell. A.R. and M.T. declare no competing interests.

Author contributions

A.R.: conceptualisation, data curation, writing and revision. M.T.: methodology, software, formal analysis, visualisation, writing (original draft). P.M.: conceptualisation, supervision, funding acquisition, writing (review and editing). All authors approved the final manuscript.

Acknowledgements

We thank the Agence de la Biomédecine for the CRISTAL registry data and all French kidney transplant centres that contribute to it, the national health insurance fund (CNAM) for the SNDS data, and the French Health Data Hub for hosting the database and providing the secured analysis environment.

References

[1] Ojo, A. O. *et al.* Long-term survival in renal transplant recipients with graft function. *Kidney Int.* **57**, 307–313 (2000).

[2] El-Zoghby, Z. M. *et al.* Identifying specific causes of kidney allograft loss. *Am. J. Transplant.* **9**, 527–535 (2009).

[3] Loupy, A. *et al.* Prediction system for risk of allograft loss in patients receiving kidney transplants: international derivation and validation study. *BMJ* **366**, l4923 (2019).

[4] Truchot, A. *et al.* Machine learning does not outperform traditional statistical modelling for kidney allograft failure prediction. *Kidney Int.* **103**, 936–948 (2023).

[5] Salaün, A., Knight, S., Wingfield, L. & Zhu, T. Predicting graft and patient outcomes following kidney transplantation using interpretable machine learning models. *Sci. Rep.* **14**, 17356 (2024).

[6] Ali, H. *et al.* Improved survival prediction for kidney transplant outcomes using artificial intelligence-based models: development of the UK Deceased Donor Kidney Transplant Outcome Prediction (UK-DTOP) Tool. *Ren. Fail.* **46**, 2373273 (2024).

[7] Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958).

[8] Gooley, T. A., Leisenring, W., Crowley, J. & Storer, B. E. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Stat. Med.* **18**, 695–706 (1999).

[9] Putter, H., Fiocco, M. & Geskus, R. B. Tutorial in biostatistics: competing risks and multi-state models. *Stat. Med.* **26**, 2389–2430 (2007).

[10] Austin, P. C., Lee, D. S. & Fine, J. P. Introduction to the analysis of survival data in the presence of competing risks. *Circulation* **133**, 601–609 (2016).

[11] Noordzij, M. *et al.* When do we need competing risks methods for survival analysis in nephrology? *Nephrol. Dial. Transplant.* **28**, 2670–2677 (2013).

[12] Schuster, N. A., Hoogendijk, E. O., Kok, A. A. L., Twisk, J. W. R. & Heymans, M. W. Ignoring competing events in the analysis of survival data may lead to biased results: a nonmathematical illustration of competing risk analysis. *J. Clin. Epidemiol.* **122**, 42–48 (2020).

[13] Aalen, O. O. & Johansen, S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand. J. Stat.* **5**, 141–150 (1978).

[14] Fine, J. P. & Gray, R. J. A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* **94**, 496–509 (1999).

[15] Wolbers, M., Koller, M. T., Witteman, J. C. M. & Steyerberg, E. W. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology* **20**, 555–561 (2009).

[16] Wolbers, M., Blanche, P., Koller, M. T., Witteman, J. C. M. & Gerds, T. A. Concordance for prognostic models with competing risks. *Biostatistics* **15**, 526–539

(2014).

[17] Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L. & Steyerberg, E. W. Calibration: the Achilles heel of predictive analytics. *BMC Med.* **17**, 230 (2019).

[18] Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.* **18**, 2529–2545 (1999).

[19] Kattan, M. W. & Gerds, T. A. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagn. Progn. Res.* **2**, 7 (2018).

[20] van Geloven, N. *et al.* Validation of prediction models in the presence of competing risks: a guide through modern methods. *BMJ* **377**, e069249 (2022).

[21] Gerds, T. A., Andersen, P. K. & Kattan, M. W. Calibration plots for risk prediction models in the presence of competing risks. *Stat. Med.* **33**, 3191–3203 (2014).

[22] CHU de Limoges & Health Data Hub. REXETRIS: information du patient. Patient information notice, version of 25 February 2021 (2021). URL https://www.health-data-hub.fr/sites/default/files/2021-03/REXETRIS%20affiche%20information%20patient%20v012%2020210225.pdf. Accessed 21 October 2026.

[23] Collins, G. S. *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).

[24] van Buuren, S. & Groothuis-Oudshoorn, K. mice: multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).

[25] Cox, D. R. Regression models and life-tables. *J. R. Stat. Soc. Ser. B (Methodol.)* **34**, 187–202 (1972).

[26] Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *Ann. Appl. Stat.* **2**, 841–860 (2008).

[27] Pölsterl, S. scikit-survival: a library for time-to-event analysis built on top of scikit-learn. *J. Mach. Learn. Res.* **21**, 1–6 (2020).

[28] Katzman, J. L. *et al.* DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 24 (2018).

[29] Kvamme, H., Borgan, Ø. & Scheel, I. Time-to-event prediction with neural networks and Cox regression. *J. Mach. Learn. Res.* **20**, 1–30 (2019).

[30] Gensheimer, M. F. & Narasimhan, B. A scalable discrete-time survival model for neural networks. *PeerJ* **7**, e6257 (2019).

[31] Kvamme, H. & Borgan, Ø. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Anal.* **27**, 710–736 (2021).

[32] Yu, C.-N., Greiner, R., Lin, H.-C. & Baracos, V. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Adv. Neural Inf. Process. Syst.* **24** (2011).

[33] Lee, C., Zame, W., Yoon, J. & van der Schaar, M. DeepHit: a deep learning approach to survival analysis with competing risks. *Proc. AAAI Conf. Artif. Intell.* **32** (2018).

[34] Nagpal, C., Li, X. & Dubrawski, A. Deep survival machines: fully parametric survival regression and representation learning for censored data with competing risks. *IEEE J. Biomed. Health Inform.* **25**, 3163–3175 (2021).

[35] Nagpal, C., Potosnak, W. & Dubrawski, A. auton-survival: an open-source package for regression, counterfactual estimation, evaluation and phenotyping with censored time-to-event data. *Proc. Mach. Learn. Res.* **182**, 585–608 (2022).

[36] Nagpal, C., Yadlowsky, S., Rostamzadeh, N. & Heller, K. Deep Cox mixtures for survival regression. *Proc. Mach. Learn. Res.* **149**, 674–708 (2021).

[37] Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. & Wei, L. J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* **30**, 1105–1117 (2011).

[38] Antolini, L., Boracchi, P. & Biganzoli, E. A time-dependent discrimination index for survival data. *Stat. Med.* **24**, 3927–3944 (2005).

[39] Gerds, T. A. & Schumacher, M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biom. J.* **48**, 1029–1040 (2006).

[40] Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making* **26**, 565–574 (2006).

[41] Marquet, P. *et al.* A pharmaco-epidemiological research program leveraging a nationwide database of kidney transplantation. *Pharmacoepidemiol. Drug Saf.* **35**, e70442 (2026).

[42] Sapir-Pichhadze, R. *et al.* Survival analysis in the presence of competing risks: the example of waitlisted kidney transplant candidates. *Am. J. Transplant.* **16**, 1958–1966 (2016).

[43] Coemans, M. *et al.* A competing risks model to estimate the risk of graft failure and patient death after kidney transplantation using continuous donor-recipient age combinations. *Am. J. Transplant.* **25**, 355–367 (2025).

Supplementary Files

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

- supplementarymaterialrev1.pdf