Supplementary Information

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] and 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: pierre.marquet@unilim.fr

Supplementary Methods

Outcome coding. With graft time T^g, graft-loss indicator δ^g, patient time T^d and death indicator δ^d, each graft received a time T and cause $\delta \in \{0, 1, 2\}$: $(T, \delta) = (T^g, 1)$ if $\delta^g = 1$; otherwise $(T^d, 2)$ if $\delta^d = 1$; otherwise $(T^g, 0)$. Grafts with a non-positive T were excluded. Graft loss $\delta^g = 1$ is a recorded functional arrest of the graft or, without a recorded arrest date, a retransplantation (dated at retransplantation); T^g ends at graft loss, death or the last date recorded for the patient, and T^d stops at retransplantation; both are capped at 31 December 2021.

Single-event outcomes. The single-event graft-loss models used (T^g, δ^g): graft loss, with death censored. No single-event death model is reported: the registry's patient-survival fields (T^d, δ^d) continue after graft loss and therefore describe all-cause mortality, not death with a functioning graft (Supplementary Table S3).

Preprocessing. Fitted on the training set and applied unchanged to the test set: binary variables encoded as 0/1; cold ischaemia time and donor and discharge creatinine log-transformed; missing values imputed by chained equations (scikit-learn IterativeImputer, 10 iterations), with imputed binary values clipped to $[0, 1]$; all variables then standardised to zero mean and unit variance. Ordinal variables (transplant rank, year of transplantation, mismatch counts) were kept as numeric.

Cumulative incidence. The Aalen–Johansen estimate of the cumulative incidence of cause k is $\widehat{\mathrm{CIF}}_k(t) = \sum_{t_j \le t} \hat{S}(t_{j-1})\, d_{kj}/n_j$, where t_j are the distinct event times, n_j the number at risk, d_{kj} the number of cause-k events and $\hat{S}$ the all-cause Kaplan–Meier estimate. One minus the Kaplan–Meier estimate with the competing event censored, $1 - \widehat{\mathrm{KM}}_k(t)$, is the quantity estimated by a single-event model for cause k.

Censoring weights. $\hat{G}(t)$ is the reverse Kaplan–Meier estimate of the censoring distribution on the test grafts. Weights $1/\hat{G}(\cdot)$ were floored at $\hat{G} \ge 0.05$.

Fixed-horizon concordance. For cause k, horizon h and predicted risks $\hat{\pi}_i$, cases are grafts with $\delta_i = k$ and $T_i \le h$. The comparable controls of case i are $\mathcal{R}_i = \{j : T_j > T_i\} \cup \{j : T_j \le T_i,\ \delta_j \notin \{0, k\}\}$ (competing-risks comparability; grafts lost to the competing event remain controls). With weights $\omega_i = 1/\hat{G}(T_i^-)^2$,

$$C_k(h) = \frac{\sum_i \omega_i \sum_{j \in \mathcal{R}_i} \left[\mathbb{1}(\hat{\pi}_i > \hat{\pi}_j) + \frac{1}{2}\mathbb{1}(\hat{\pi}_i = \hat{\pi}_j) \right]}{\sum_i \omega_i \, |\mathcal{R}_i|}.$$

The cause-specific variant restricts $\mathcal{R}_i$ to $\{j : T_j > T_i\}$ (Supplementary Figure S1).

Antolini's concordance. As above, but each case is compared with its controls on the predicted cumulative incidence evaluated at the case's own event time, $\widehat{\mathrm{CIF}}_k(T_i \mid x)$, over all cases with T_i up to the last time at which $\hat{G} \geq 0.05$.

Brier score and IPA. With $N_{ik}(h) = \mathbb{1}(T_i \leq h, \delta_i = k)$ and the usual inverse-probability-of-censoring weights, $\mathrm{BS}_k(h) = n^{-1} \sum_i w_i(h)\{N_{ik}(h) - \hat{\pi}_i\}^2$; $\mathrm{IPA}_k(h) = 1 - \mathrm{BS}_k(h)/\mathrm{BS}_k^0(h)$, where BS^0 is the Brier score of the covariate-free Aalen–Johansen estimate fitted on the training set.

Calibration. At each horizon, test grafts were grouped into deciles of $\hat{\pi}$; for decile d, $\bar{\pi}_d$ is the mean prediction and $\widehat{\mathrm{CIF}}_{k,d}(h)$ the Aalen–Johansen estimate among its grafts. Calibration error $= \mathrm{mean}_{d,h}|\bar{\pi}_d - \widehat{\mathrm{CIF}}_{k,d}(h)|$; calibration bias $= \mathrm{mean}_{d,h}(\bar{\pi}_d - \widehat{\mathrm{CIF}}_{k,d}(h))$.

Decomposition of the single-event excess. For a single-event model with mean predicted risk $\bar{\pi}(h)$ over the test grafts,

$$\frac{\bar{\pi}(h)}{\widehat{\mathrm{CIF}}(h)} = \underbrace{\frac{1 - \widehat{\mathrm{KM}}(h)}{\widehat{\mathrm{CIF}}(h)}}_{\text{estimand}} \times \underbrace{\frac{\bar{\pi}(h)}{1 - \widehat{\mathrm{KM}}(h)}}_{\text{model}},$$

where $1 - \widehat{\mathrm{KM}}(h)$, one minus the Kaplan–Meier estimate with the competing event censored, is the target of every single-event model, so that the estimand component is common to all of them, and the model component is the miscalibration of the model against its own target. The relative excesses reported are these ratios minus one; they multiply: $(1 + \text{estimand})(1 + \text{model}) = 1 + \text{total}$.

Net benefit. At threshold p_t, $\mathrm{NB}(p_t) = \mathrm{TP}/n - \mathrm{FP}/n \cdot p_t/(1 - p_t)$, with true and false positives weighted by the censoring weights above.

Fine–Gray implementation. For cause k, a subject with the competing event at T_l remains in the risk set after T_l with weight $\hat{G}(t)/\hat{G}(T_l)$. The weighted partial likelihood was maximised by Newton–Raphson (ridge 10^{-4}, step length capped at 1, linear predictor clipped at ± 30, at most 50 iterations, tolerance 10^{-6}); risk-set sums were computed by forward and reverse cumulative sums, so that each iteration is $O(n \log n)$. The baseline cumulative subdistribution hazard was estimated by the Breslow estimator and $\widehat{\mathrm{CIF}}_k(t \mid x) = 1 - \exp\{-\hat{\Lambda}_{0k}(t) e^{\hat{\beta}^\top x}\}$. With no covariates, the estimate reproduced the Aalen–Johansen cumulative incidence to within 2×10^{-4}.

Bootstrap. 200 replicates of the test set, drawn by recipient with replacement (each sampled recipient contributes all of their test grafts); every model was scored on the same replicate; $\hat{G}$ was held fixed. Percentile intervals. Replicate-level results are stored with the analysis (`boot_replicates.npz`).

Outcome-coding sensitivity. The registry records graft and patient follow-up separately. A death without a recorded graft loss was classified by its date relative to the end of graft follow-up T^g: on the same day ($|T^d - T^g| \leq 1$ day), before it, or after it. In the alternative definition, a patient whose graft follow-up ends before death is censored at T^g, because the state of the graft at death is unknown. We recomputed the cumulative incidences and the estimand component under this definition, censoring at T^g every death occurring more than W days after T^g, for $W = 365$, 90, 30 and 1 day (Supplementary Table S7). For $W = 365$, all competing-risks models were retrained on the recoded outcome (with the same test grafts) and every model was re-evaluated (Supplementary Table S8).

Training runs. Each deep-learning model was trained three times with random seeds 42, 43 and 44 (Deep Cox Mixtures, whose library fixes its own seed, with 0, 43 and 44); the seed controls the weight initialisation, the mini-batch order and the internal validation split. For each test graft, the predicted cumulative incidence curves (competing-risks models) or survival curves (single-event

models) were averaged pointwise over the three runs on a common time grid. The Cox, random survival forest and Fine–Gray models were fitted once. Supplementary Table S11 reports each deep-learning model's performance in each run.

Subgroups. Era of transplantation (2005–2008, 2009–2012, 2013–2016, 2017–2020), transplant rank (first graft since 2005 versus second or later), recipient age at transplantation (18–44, 45–59, 60–69, 70 or more) and the donor–recipient sex combination were evaluated on the test grafts of each subgroup with the same measures, without bootstrap. Horizons at which fewer than 50 grafts of a subgroup remained at risk were omitted, so that the 10-year horizon is available only for 2005–2012 and the 5-year horizon not for 2017–2020.

Robustness of family summaries. Family means can be dominated by a single model. For the mean predicted risk of the single-event family we therefore also report the median model and the mean without the model with the highest prediction (Supplementary Table S6).

Software. Python 3.10; PyTorch 1.13.1; pycox; auton-survival 0.1.0; scikit-survival 0.21; scikit-learn 1.2.2; NumPy 1.26.4; pandas 1.5.3. All evaluation measures were computed by one module from the saved test-set predictions of every model.

Supplementary Tables

Table S1: TRIPOD+AI checklist [1] (Table 2 of the statement; item numbers as published) and where each item is reported. M, Methods; R, Results; D, Discussion; S, Supplementary Methods.

Section	Item	Where reported
Title	1	Title
Abstract	2	Structured abstract (checked against TRIPOD+AI for Abstracts: title, background, objectives, data source, participants, outcome, predictors, sample size and events, model building, evaluation, results with CI, limitations, interpretation; no registration)
Background	3a	Introduction, paragraphs 1–3
	3b	Introduction; D: Implications (risk communication and thresholds in transplant follow-up)
	3c	Not addressed
Objectives	4	Introduction, last paragraph (development and internal evaluation of 14 models)
Data	5a	M: Data source (CRISTAL, via REXETRIS)
	5b	M: Data source (transplants 2005–2020; follow-up to 31 December 2021)
Participants	6a	M: Data source (all kidney transplant centres in France)
	6b	M: Data source (adult recipients and donors; grafts lost within 7 days and incomplete records excluded); R: Cohort
	6c	Not modelled: treatments after transplantation were not used as predictors
Data preparation	7	M: Predictors (encoding, transforms, standardisation, fitted on training data)
Outcome	8a	M: Outcomes, Performance evaluation (horizons 1, 3, 5, 10 years); S: Outcome coding
	8b	Not applicable (registry-recorded outcomes)
	8c	Not applicable (registry-recorded outcomes)
Predictors	9a	M: Predictors (all available variables at transplantation and discharge; no selection)
	9b	M: Predictors; Table 1
	9c	Not applicable
Sample size	10	M: Data split (whole national cohort; no formal calculation); Supplementary Table S9 (events)
Missing data	11	M: Predictors (chained-equation imputation); Table 1 (missing %)
Analytical methods	12a	M: Data split (70/30 by recipient, stratified on outcome and transplant year; same test grafts for all models)

Section	Item	Where reported
	12b	M: Predictors
	12c	M: Models, Shared settings, Training runs (three seeds, averaged); Supplementary Tables S4 (no tuning) and S11
	12d	Not examined (centre not available); D: Limitations
	12e	M: Performance evaluation; S: formulas
	12f	Not performed (no recalibration); D: Implications
	12g	M: Models (risk $= 1 - \hat{S}(h)$ or $\widehat{\mathrm{CIF}}_k(h)$ from saved test-set predictions)
Class imbalance	13	Not applicable (time-to-event models; no resampling)
Fairness	14	Partly: transplant era, rank, recipient age and donor and recipient sex (Supplementary Tables S12–S16); centre not available (D: Limitations)
Model output	15	M: Models (predicted probabilities by horizon); decision thresholds 1–50%
Training vs evaluation	16	M: Data split (random split of one population by recipient); Supplementary Table S10
Ethical approval	17	M: Ethics and data access
Open science	18a	Funding statement
	18b	Competing interests statement
	18c	No study-specific protocol; the database and research programme are described in [2]
	18d	Registered by the Health Data Hub (T11068872019112)
	18e	Data availability statement
	18f	Data availability statement; M: Software and reproducibility (code available on request)
Patient and public involvement	19	Programme level: two patient-association representatives sit on the programme's scientific committee [2]; none for this analysis
Participants	20a	R: Cohort; Supplementary Table S9
	20b	Table 1
	20c	Table 1 (training and test sets)
Model development	21	Supplementary Table S9; M: Data split
Model specification	22	No single model is proposed for use; settings in Supplementary Table S4
Model performance	23a	Table 2; Supplementary Table S5 (95% CIs); subgroups in Supplementary Tables S13–S16
	23b	Supplementary Tables S12, S13 (era), S14 (rank), S15 (age) and S16 (sex); point estimates
Model updating	24	Not applicable (no updating)
Interpretation	25	D
Limitations	26	D: Strengths and limitations
Usability	27a	Not applicable (benchmark; no model proposed for implementation)
	27b	Not applicable
	27c	D: Implications, Limitations (external and temporal validation)

Table S2: The 36 predictors: definition, extract column, and unit or, for binary variables, the meaning of TRUE. Unknown values of binary variables were missing and imputed with the other variables. Recipient characteristics are those recorded at registration on the waiting list unless stated otherwise.

Group	Variable	Column	Unit / TRUE means
Donor	Age	`age_donor`	years
	Sex	`sex_donor`	female
	Height	`height_donor`	cm
	Weight	`weight_donor`	kg
	Diabetes	`diabetes_donor`	diabetes
	Hypertension	`hbp_donor`	hypertension
	CMV serostatus	`cmv_donor_2l`	positive
	EBV serostatus	`ebv_donor_2l`	positive
	Kidney disease	`kidney_disease_donor`	kidney disease
	Serum creatinine	`creat_donor`	µmol/L
	Living donation	`donor_living`	living donor
Recipient	Age at transplantation	`age_tx`	years
	Sex	`sex`	female
	Height at registration	`height_regist`	cm
	Weight at registration	`weight_regist`	kg
	Diabetes	`diabetes_regist`	diabetes
	Hypertension	`history_hbp_regist`	history of hypertension
	CMV serostatus	`cmv`	positive at registration or transplantation
	EBV serostatus	`ebv`	positive at registration or transplantation
	Smoking history	`smoking_ever_regist`	former or current smoker
	Cardiovascular disease	`cvd_regist`	any of arrhythmia, coronary insufficiency, heart failure, myocardial infarction, unstable angina, stroke or transient ischaemic attack
	Dyslipidaemia	`dyslipidemia_regist`	dyslipidaemia
	Cirrhosis	`cirrhosis_regist`	cirrhosis
	Neuropathy	`neuropath_regist`	neuropathy
	Urological pathology	`uro_patho_regist`	urological pathology
	Registration to transplantation	`TxDelayYrs`	years
	Transplant rank	`rexetris_tx_rank`	rank (1–3) since 2005
Matching	HLA-A mismatches	`MCHA`	0–2
	HLA-B mismatches	`MCHB`	0–2
	HLA-DR mismatches	`MCHDR`	0–2
Transplant	Machine perfusion	`kidney_machine`	machine-perfused
	Cold ischaemia time	`cit`	minutes
	Year of transplantation	`YearofTx`	year
Discharge	Serum creatinine	`creat_dc_hosp`	µmol/L
	Weight	`weight_discharge_hosp`	kg
	Functioning graft	`functionnal_discharge_hosp`	functioning graft

Table S3: Endpoint check, test grafts. AJ: observed cumulative incidence (Aalen–Johansen). 1−KM (CR-coded): one minus Kaplan–Meier with the competing event censored, the target of a single-event model trained on the competing-risks coding. 1−KM (registry fields): the same on the registry's own fields — graft survival for graft loss; patient survival for death, which continues after graft loss and therefore measures all-cause mortality. Last two columns: mean prediction of each model family (single-event: graft loss only).

Endpoint	Horizon	AJ	1−KM (CR-coded)	1−KM (registry fields)	Single-event mean	Competing-risks mean
Graft loss	1 y	0.026	0.026	0.026	0.029	0.028
Graft loss	3 y	0.058	0.061	0.061	0.063	0.060
Graft loss	5 y	0.094	0.099	0.099	0.106	0.096
Graft loss	10 y	0.185	0.210	0.210	0.242	0.180
Death	1 y	0.031	0.031	0.033	–	0.031
Death	3 y	0.067	0.069	0.076	–	0.070
Death	5 y	0.105	0.111	0.122	–	0.113
Death	10 y	0.215	0.240	0.273	–	0.227

Table S4: Model settings. All settings were fixed a priori and identical across runs; none were tuned. Neural networks: two hidden layers (128, 64 units), dropout 0.2, batch size 256, Adam with learning rate 10^{-3}; pycox models used early stopping (patience 20 epochs, at most 256 epochs) on a stratified 15% of the training set.

Model	Library	Settings
Single-event (graft loss, death censored)		
Cox proportional hazards	scikit-survival	Efron ties; ridge penalty 10^{-4}; at most 200 iterations
Random survival forest	scikit-survival	500 trees; minimum 10 grafts to split, 15 per leaf; $\sqrt{p}$ features per split
DeepSurv, Cox-Time	pycox	network and early stopping as above
Logistic-hazard, MTLR, PMF	pycox	50 discrete intervals (quantiles of training event times)
DeepHit	pycox	50 intervals; $\alpha = 0.2$, $\sigma = 0.1$
Deep Survival Machines	auton-survival	3 Weibull components; 256 iterations, no early stopping
Competing-risks		
Fine–Gray	NumPy (this study)	IPCW partial likelihood; Newton–Raphson, ridge 10^{-4}, at most 50 iterations; Breslow baseline
Cause-specific DeepSurv	pycox	one network per cause; CIFs by the discrete Aalen–Johansen product
DeepHit	pycox	50 intervals; $\alpha = 0.2$, $\sigma = 0.1$
Deep Survival Machines	auton-survival	competing-risks likelihood; 3 Weibull components; 50 iterations
Deep Cox Mixtures	auton-survival	fitted per cause and combined as for cause-specific DeepSurv; 50 iterations
Aalen–Johansen (reference)	NumPy	covariate-free; training set; reference for the IPA

Table S5: Fixed-horizon concordance, IPCW Brier score and IPA for graft loss by horizon: family means, competing-risks minus single-event difference, and difference between the best model of each family (highest concordance or IPA, lowest Brier score, on the point estimate), with 95% paired bootstrap CIs. For concordance and IPA a positive difference favours the competing-risks models; for the Brier score a negative one does.

Endpoint	Horizon	Single-event	Competing-risks	Difference (95% CI)	Best vs best (95% CI)
C-index					
Graft loss	1 y	0.810	0.804	−0.006 (−0.009 to −0.002)	−0.005 (−0.014 to +0.003)
	3 y	0.732	0.729	−0.004 (−0.007 to 0.000)	+0.001 (−0.004 to +0.006)
	5 y	0.691	0.690	−0.001 (−0.005 to +0.003)	+0.002 (−0.004 to +0.006)
	10 y	0.627	0.644	+0.017 (+0.013 to +0.020)	+0.008 (+0.004 to +0.011)
	mean	0.715	0.716	+0.002 (−0.001 to +0.005)	−0.001 (−0.004 to +0.003)
IPCW Brier score					
Graft loss	1 y	0.023	0.023	0.000 (0.000 to 0.000)	0.000 (0.000 to 0.000)
	3 y	0.052	0.052	0.000 (0.000 to 0.000)	0.000 (0.000 to 0.000)
	5 y	0.081	0.081	0.000 (0.000 to +0.001)	0.000 (0.000 to 0.000)
	10 y	0.150	0.148	−0.002 (−0.003 to −0.001)	−0.002 (−0.003 to −0.001)
	mean	0.076	0.076	0.000 (−0.001 to 0.000)	−0.001 (−0.001 to −0.001)
IPA					
Graft loss	1 y	0.093	0.090	−0.003 (−0.012 to +0.008)	0.000 (−0.007 to +0.007)
	3 y	0.061	0.059	−0.002 (−0.008 to +0.006)	+0.003 (−0.002 to +0.008)
	5 y	0.044	0.043	−0.001 (−0.007 to +0.005)	0.000 (−0.004 to +0.005)
	10 y	0.004	0.017	+0.013 (+0.007 to +0.021)	+0.013 (+0.005 to +0.021)
	mean	0.051	0.052	+0.002 (−0.004 to +0.010)	+0.008 (+0.004 to +0.013)

Table S6: Mean predicted risk of graft loss by horizon: robustness of the single-event family summary to a single model. Last single-event column: family mean without the model with the highest mean prediction at that horizon (named, with its prediction). Relative excess over the observed incidence (AJ) in parentheses.

			Single-event			Competing-risks
Horizon	AJ	1−KM	Mean	Median	Mean without highest	Mean
1 year	0.026	0.026	0.029 (+13%)	0.030 (+14%)	0.029 (+12%; PMF 0.032)	0.028 (+8%)
3 years	0.058	0.061	0.063 (+8%)	0.064 (+9%)	0.062 (+7%; Deep Survival Machines 0.068)	0.060 (+3%)
5 years	0.094	0.099	0.106 (+13%)	0.105 (+12%)	0.105 (+12%; logistic-hazard 0.118)	0.096 (+3%)
10 years	0.185	0.210	0.242 (+31%)	0.216 (+17%)	0.218 (+18%; logistic-hazard 0.434)	0.180 (−3%)

Table S7: Outcome-coding sensitivity, all 46,559 grafts (7,309 graft losses under every coding). Original: every death without a recorded graft loss is a death with a functioning graft. Alternative: a death more than W days after the end of graft follow-up is censored there. Estimand gap: relative excess of one minus Kaplan–Meier over the cumulative incidence of graft loss at 10 years. Of the 8,574 deaths without a recorded graft loss, 90.2% occurred on the last day of graft follow-up, fewer than 11 before it and 9.8% after it (median 0.3 years, interquartile range 0.1–0.8; more than 1 year for 14.7% of these).

Coding	Deaths with a functioning graft	AJ graft loss 10 years	AJ death 10 years	1−KM graft loss 10 years	Estimand gap 10 years
Original	8,574	0.187	0.216	0.213	+13.7%
$W = 365$ days	8,450	0.188	0.213	0.213	+13.5%
$W = 90$ days	8,109	0.189	0.206	0.213	+13.0%
$W = 30$ days	7,950	0.189	0.202	0.213	+12.7%
$W = 1$ day	7,731	0.190	0.198	0.213	+12.4%

Table S8: Sensitivity analysis: registry coding of deaths after the end of graft follow-up ($W = 365$ days), with all competing-risks models retrained (single runs; primary → alternative coding). Test set: 38 of 2,558 evaluated deaths with a functioning graft recoded as censored; 2,185 graft losses unchanged. Single-event graft-loss models do not use the death fields and changed by at most 0.001 in every measure. The change of Deep Cox Mixtures is within its run-to-run variation (Supplementary Table S11).

Endpoint	Model	C-index	IPA	Cal. error	Cal. bias	Antolini C
Death	Cause-specific DeepSurv	0.747 → 0.747	0.119 → 0.115	0.008 → 0.007	+0.001 → +0.001	0.669 → 0.664
	Deep Cox Mixtures	0.692 → 0.708	−0.001 → 0.031	0.054 → 0.041	+0.017 → −0.005	0.640 → 0.656
	DeepHit	0.746 → 0.745	0.096 → 0.100	0.027 → 0.026	−0.007 → −0.004	0.648 → 0.651
	Deep Survival Machines	0.750 → 0.751	0.108 → 0.109	0.011 → 0.015	+0.004 → +0.013	0.676 → 0.678
	Fine–Gray	0.733 → 0.734	0.089 → 0.089	0.005 → 0.005	+0.001 → +0.001	0.679 → 0.679
Graft loss	Cause-specific DeepSurv	0.717 → 0.718	0.058 → 0.052	0.011 → 0.011	+0.002 → +0.003	0.640 → 0.636
	Deep Cox Mixtures	0.676 → 0.717	−0.058 → 0.063	0.051 → 0.015	+0.007 → −0.002	0.604 → 0.615
	DeepHit	0.720 → 0.719	0.051 → 0.052	0.022 → 0.018	+0.008 → +0.006	0.601 → 0.604
	Deep Survival Machines	0.716 → 0.717	0.067 → 0.067	0.007 → 0.011	+0.002 → +0.007	0.642 → 0.643
	Fine–Gray	0.708 → 0.708	0.041 → 0.041	0.011 → 0.011	+0.002 → +0.002	0.645 → 0.646

Table S9: Outcomes in the training and test sets. The split was made by recipient and stratified on outcome and year of transplantation.

	Training	Test
Grafts, n	32,609	13,950
Graft loss, n (%)	5,124 (15.7)	2,185 (15.7)
Death with a functioning graft, n (%)	6,015 (18.4)	2,559 (18.3)
Censored, n (%)	21,470 (65.8)	9,206 (66.0)
Second or later graft since 2005, n (%)	1,346 (4.1)	553 (4.0)
Test set: observed cumulative incidence (Aalen–Johansen) at 1 / 3 / 5 / 10 years, %		
Graft loss		2.6 / 5.8 / 9.4 / 18.5
Death with a functioning graft		3.1 / 6.7 / 10.5 / 21.5

Table S10: Training and test sets by year of transplantation. The split was made by recipient and stratified on outcome (graft loss, death with a functioning graft, censored) crossed with year of transplantation. Outcomes by set are given in Supplementary Table S9.

Year	Training	Test	Test share (%)
2005	1,680	719	30.0
2006	1,773	760	30.0
2007	1,869	797	29.9
2008	1,921	825	30.0
2009	1,852	799	30.1
2010	1,881	802	29.9
2011	1,939	827	29.9
2012	1,999	850	29.8
2013	2,015	858	29.9
2014	2,115	903	29.9
2015	2,300	976	29.8
2016	2,353	1,002	29.9
2017	2,475	1,055	29.9
2018	2,334	1,018	30.4
2019	2,398	1,021	29.9
2020	1,705	738	30.2
Total	32,609	13,950	30.0

Table S11: Run-to-run variation of the deep-learning models: mean over three training runs (minimum to maximum), each run scored alone on the test grafts (pooled over horizons). Cox, random survival forest and Fine–Gray were fitted once (single value). The main results use the averaged predictions of the three runs, which differ from the mean of the single runs.

Model	C-index	IPA	Cal. error	Cal. bias	Antolini C
Graft loss					
Single-event					
Cox	0.712	0.048	0.017	+0.014	0.597
Cox-Time	0.709 (0.702 to 0.713)	0.046 (0.043 to 0.049)	0.017 (0.016 to 0.018)	+0.014 (+0.014 to +0.015)	0.589 (0.576 to 0.598)
Deep Survival Machines	0.721 (0.720 to 0.721)	0.056 (0.053 to 0.058)	0.019 (0.017 to 0.021)	+0.017 (+0.014 to +0.021)	0.606 (0.598 to 0.614)
DeepHit	0.708 (0.706 to 0.713)	0.053 (0.051 to 0.054)	0.021 (0.018 to 0.022)	+0.013 (+0.011 to +0.015)	0.572 (0.568 to 0.576)
DeepSurv	0.717 (0.715 to 0.719)	0.038 (0.037 to 0.039)	0.019 (0.018 to 0.020)	+0.015 (+0.015 to +0.015)	0.608 (0.603 to 0.613)
Logistic-hazard	0.689 (0.688 to 0.691)	0.029 (0.023 to 0.039)	0.072 (0.064 to 0.081)	+0.067 (+0.058 to +0.077)	0.551 (0.549 to 0.553)
MTLR	0.710 (0.706 to 0.713)	0.036 (0.029 to 0.041)	0.020 (0.014 to 0.024)	+0.014 (+0.010 to +0.019)	0.589 (0.580 to 0.595)
PMF	0.711 (0.709 to 0.715)	0.046 (0.038 to 0.050)	0.021 (0.019 to 0.024)	+0.019 (+0.017 to +0.022)	0.577 (0.567 to 0.582)
Random survival forest	0.718	0.063	0.013	+0.013	0.595
Competing-risks					
Cause-specific DeepSurv	0.718 (0.717 to 0.720)	0.057 (0.055 to 0.058)	0.010 (0.009 to 0.011)	+0.003 (+0.002 to +0.003)	0.638 (0.630 to 0.644)
Deep Cox Mixtures	0.689 (0.674 to 0.718)	−0.021 (−0.067 to 0.062)	0.039 (0.011 to 0.054)	+0.006 (+0.003 to +0.007)	0.622 (0.602 to 0.660)
DeepHit	0.718 (0.716 to 0.720)	0.054 (0.051 to 0.058)	0.019 (0.017 to 0.022)	+0.006 (+0.003 to +0.008)	0.581 (0.566 to 0.601)
Deep Survival Machines	0.719 (0.716 to 0.721)	0.068 (0.065 to 0.072)	0.009 (0.007 to 0.013)	+0.002 (−0.003 to +0.007)	0.640 (0.628 to 0.650)
Fine–Gray	0.708	0.041	0.011	+0.002	0.645
Death with a functioning graft					
Competing-risks					
Cause-specific DeepSurv	0.746 (0.744 to 0.747)	0.116 (0.114 to 0.119)	0.007 (0.006 to 0.008)	+0.002 (+0.001 to +0.002)	0.670 (0.669 to 0.671)
Deep Cox Mixtures	0.691 (0.675 to 0.707)	0.004 (−0.005 to 0.018)	0.051 (0.050 to 0.054)	+0.009 (+0.000 to +0.017)	0.637 (0.624 to 0.646)
DeepHit	0.747 (0.746 to 0.748)	0.098 (0.096 to 0.099)	0.027 (0.026 to 0.028)	−0.006 (−0.007 to −0.005)	0.652 (0.648 to 0.657)
Deep Survival Machines	0.748 (0.746 to 0.750)	0.108 (0.106 to 0.109)	0.011 (0.011 to 0.011)	+0.002 (−0.003 to +0.004)	0.674 (0.671 to 0.676)
Fine–Gray	0.733	0.089	0.005	+0.001	0.679

Table S12: Observed cumulative incidence (Aalen–Johansen) by era of transplantation, all 46,559 grafts. –: horizon not estimable (follow-up ends on 31 December 2021).

		Graft loss				Death with a functioning graft			
Era	Grafts	1 y	3 y	5 y	10 y	1 y	3 y	5 y	10 y
2005–2008	10,344	0.028	0.063	0.101	0.207	0.027	0.053	0.085	0.171
2009–2012	10,949	0.030	0.063	0.099	0.191	0.030	0.061	0.096	0.209
2013–2016	12,522	0.027	0.058	0.095	–	0.031	0.067	0.111	–
2017–2020	12,744	0.024	0.051	–	–	0.038	0.093	–	–

Table S13: Performance by era of transplantation, test grafts (number in parentheses): family means of concordance and IPA, mean predicted risk and observed cumulative incidence (AJ). SE, single-event; CR, competing-risks. Point estimates; horizons with fewer than 50 grafts at risk omitted.

Era	Endpoint	Horizon	Family	C-index	IPA	Mean predicted	AJ
2005–2008 (3,100)	Death	1 y	CR	0.744	0.027	0.023	0.026
	Death	3 y	CR	0.717	0.038	0.052	0.050
	Death	5 y	CR	0.719	0.059	0.086	0.082
	Death	10 y	CR	0.715	0.114	0.182	0.164
	Graft loss	1 y	CR	0.707	0.073	0.030	0.027
	Graft loss	1 y	SE	0.707	0.085	0.030	0.027
	Graft loss	3 y	CR	0.642	0.033	0.068	0.064
	Graft loss	3 y	SE	0.635	0.037	0.063	0.064
	Graft loss	5 y	CR	0.622	0.014	0.110	0.101
	Graft loss	5 y	SE	0.617	0.016	0.108	0.101
	Graft loss	10 y	CR	0.596	−0.011	0.214	0.205
	Graft loss	10 y	SE	0.593	−0.017	0.226	0.205
2009–2012 (3,278)	Death	1 y	CR	0.753	0.046	0.028	0.027
	Death	3 y	CR	0.740	0.046	0.064	0.057
	Death	5 y	CR	0.734	0.051	0.103	0.089
	Death	10 y	CR	0.733	0.144	0.211	0.209
	Graft loss	1 y	CR	0.831	0.123	0.031	0.031
	Graft loss	1 y	SE	0.837	0.140	0.033	0.031
	Graft loss	3 y	CR	0.741	0.083	0.066	0.064
	Graft loss	3 y	SE	0.749	0.093	0.065	0.064
	Graft loss	5 y	CR	0.706	0.060	0.104	0.099
	Graft loss	5 y	SE	0.713	0.067	0.109	0.099
	Graft loss	10 y	CR	0.627	0.026	0.196	0.188
	Graft loss	10 y	SE	0.625	0.013	0.224	0.188
2013–2016 (3,739)	Death	1 y	CR	0.740	0.051	0.032	0.031
	Death	3 y	CR	0.730	0.035	0.074	0.064
	Death	5 y	CR	0.738	0.059	0.119	0.111
	Graft loss	1 y	CR	0.829	0.080	0.027	0.025
	Graft loss	1 y	SE	0.834	0.077	0.030	0.025
	Graft loss	3 y	CR	0.755	0.056	0.059	0.055
	Graft loss	3 y	SE	0.761	0.058	0.060	0.055
	Graft loss	5 y	CR	0.690	0.041	0.093	0.093
	Graft loss	5 y	SE	0.694	0.037	0.102	0.093
2017–2020 (3,832)	Death	1 y	CR	0.776	0.087	0.039	0.039
	Death	3 y	CR	0.749	0.122	0.087	0.096
	Graft loss	1 y	CR	0.816	0.079	0.025	0.022
	Graft loss	1 y	SE	0.825	0.063	0.028	0.022
	Graft loss	3 y	CR	0.756	0.064	0.052	0.051
	Graft loss	3 y	SE	0.760	0.054	0.056	0.051

Table S14: Performance by transplant rank (first graft since 2005 versus second or later), test grafts (number in parentheses). Same measures as Supplementary Table S13.

Rank	Endpoint	Horizon	Family	C-index	IPA	Mean predicted	AJ
First (13,396)	Death	1 y	CR	0.755	0.055	0.031	0.031
	Death	3 y	CR	0.737	0.065	0.070	0.066
	Death	5 y	CR	0.734	0.091	0.112	0.105
	Death	10 y	CR	0.725	0.166	0.225	0.215
	Graft loss	1 y	CR	0.802	0.088	0.028	0.026
	Graft loss	1 y	SE	0.808	0.091	0.030	0.026
	Graft loss	3 y	CR	0.729	0.058	0.060	0.058
	Graft loss	3 y	SE	0.732	0.061	0.060	0.058
	Graft loss	5 y	CR	0.689	0.043	0.095	0.093
	Graft loss	5 y	SE	0.691	0.045	0.104	0.093
	Graft loss	10 y	CR	0.642	0.015	0.179	0.184
	Graft loss	10 y	SE	0.626	0.003	0.238	0.184
Second or later (553)	Death	1 y	CR	0.781	0.115	0.039	0.033
	Death	3 y	CR	0.812	0.125	0.085	0.067
	Death	5 y	CR	0.787	0.160	0.133	0.119
	Graft loss	1 y	CR	0.857	0.133	0.036	0.031
	Graft loss	1 y	SE	0.849	0.124	0.039	0.031
	Graft loss	3 y	CR	0.738	0.076	0.075	0.068
	Graft loss	3 y	SE	0.737	0.055	0.075	0.068
	Graft loss	5 y	CR	0.679	0.054	0.118	0.126
	Graft loss	5 y	SE	0.669	0.026	0.132	0.126

Table S15: Performance by recipient age at transplantation (years), test grafts (number in parentheses). Same measures as Supplementary Table S13.

Recipient age	Endpoint	Horizon	Family	C-index	IPA	Mean predicted	AJ
18–44 (4,158)	Death	1 y	CR	0.759	0.070	0.008	0.012
	Death	3 y	CR	0.685	0.113	0.020	0.022
	Death	5 y	CR	0.669	0.178	0.034	0.032
	Death	10 y	CR	0.652	0.294	0.076	0.064
	Graft loss	1 y	CR	0.831	0.127	0.024	0.022
	Graft loss	1 y	SE	0.836	0.151	0.023	0.022
	Graft loss	3 y	CR	0.733	0.069	0.056	0.055
	Graft loss	3 y	SE	0.735	0.078	0.050	0.055
	Graft loss	5 y	CR	0.685	0.044	0.092	0.097
	Graft loss	5 y	SE	0.685	0.054	0.092	0.097
	Graft loss	10 y	CR	0.639	0.008	0.186	0.197
	Graft loss	10 y	SE	0.631	0.029	0.223	0.197
45–59 (5,261)	Death	1 y	CR	0.707	0.043	0.024	0.025
	Death	3 y	CR	0.668	0.034	0.055	0.051
	Death	5 y	CR	0.654	0.036	0.090	0.081
	Death	10 y	CR	0.627	0.053	0.195	0.178
	Graft loss	1 y	CR	0.816	0.105	0.024	0.019
	Graft loss	1 y	SE	0.816	0.111	0.027	0.019
	Graft loss	3 y	CR	0.727	0.057	0.054	0.045
	Graft loss	3 y	SE	0.727	0.061	0.054	0.045
	Graft loss	5 y	CR	0.684	0.039	0.087	0.075
	Graft loss	5 y	SE	0.684	0.042	0.093	0.075
	Graft loss	10 y	CR	0.639	0.017	0.168	0.166
	Graft loss	10 y	SE	0.627	0.013	0.219	0.166
60–69 (3,210)	Death	1 y	CR	0.661	0.051	0.051	0.049
	Death	3 y	CR	0.630	0.042	0.115	0.105
	Death	5 y	CR	0.617	0.065	0.182	0.177
	Death	10 y	CR	0.589	0.142	0.358	0.369
	Graft loss	1 y	CR	0.774	0.070	0.034	0.035
	Graft loss	1 y	SE	0.777	0.065	0.039	0.035
	Graft loss	3 y	CR	0.710	0.051	0.069	0.071
	Graft loss	3 y	SE	0.710	0.044	0.074	0.071
	Graft loss	5 y	CR	0.682	0.042	0.106	0.108
	Graft loss	5 y	SE	0.682	0.031	0.124	0.108
	Graft loss	10 y	CR	0.645	0.022	0.187	0.200
	Graft loss	10 y	SE	0.625	−0.019	0.269	0.200
70 or more (1,320)	Death	1 y	CR	0.659	0.082	0.084	0.073
	Death	3 y	CR	0.633	0.127	0.185	0.176
	Death	5 y	CR	0.613	0.179	0.284	0.271
	Death	10 y	CR	0.571	0.335	0.506	0.546
	Graft loss	1 y	CR	0.724	0.043	0.041	0.044
	Graft loss	1 y	SE	0.719	0.021	0.048	0.044
	Graft loss	3 y	CR	0.701	0.058	0.081	0.095
	Graft loss	3 y	SE	0.706	0.060	0.090	0.095
	Graft loss	5 y	CR	0.688	0.054	0.121	0.127
	Graft loss	5 y	SE	0.689	0.053	0.151	0.127
	Graft loss	10 y	CR	0.657	0.038	0.194	0.192
	Graft loss	10 y	SE	0.628	−0.064	0.321	0.192

Table S16: Performance by donor and recipient sex, test grafts (number in parentheses). Same measures as Supplementary Table S13. F, female; M, male.

Donor / recipient	Endpoint	Horizon	Family	C-index	IPA	Mean predicted	AJ
Donor F / recipient F (2,300)	Death	1 y	CR	0.749	0.060	0.032	0.033
	Death	3 y	CR	0.727	0.062	0.071	0.065
	Death	5 y	CR	0.727	0.092	0.114	0.101
	Death	10 y	CR	0.716	0.166	0.229	0.223
	Graft loss	1 y	CR	0.734	0.068	0.028	0.024
	Graft loss	1 y	SE	0.740	0.064	0.030	0.024
	Graft loss	3 y	CR	0.691	0.041	0.063	0.057
	Graft loss	3 y	SE	0.689	0.035	0.062	0.057
	Graft loss	5 y	CR	0.655	0.027	0.101	0.097
	Graft loss	5 y	SE	0.649	0.024	0.109	0.097
	Graft loss	10 y	CR	0.636	0.011	0.192	0.190
	Graft loss	10 y	SE	0.615	−0.001	0.248	0.190
Donor F / recipient M (3,829)	Death	1 y	CR	0.750	0.059	0.035	0.034
	Death	3 y	CR	0.755	0.086	0.079	0.077
	Death	5 y	CR	0.753	0.120	0.126	0.121
	Death	10 y	CR	0.737	0.207	0.249	0.241
	Graft loss	1 y	CR	0.788	0.092	0.030	0.030
	Graft loss	1 y	SE	0.802	0.098	0.033	0.030
	Graft loss	3 y	CR	0.718	0.055	0.064	0.070
	Graft loss	3 y	SE	0.728	0.058	0.065	0.070
	Graft loss	5 y	CR	0.685	0.044	0.100	0.103
	Graft loss	5 y	SE	0.693	0.049	0.111	0.103
	Graft loss	10 y	CR	0.635	0.010	0.183	0.193
	Graft loss	10 y	SE	0.621	−0.005	0.249	0.193
Donor M / recipient F (2,885)	Death	1 y	CR	0.757	0.066	0.027	0.026
	Death	3 y	CR	0.728	0.062	0.060	0.053
	Death	5 y	CR	0.729	0.090	0.096	0.082
	Death	10 y	CR	0.724	0.158	0.196	0.169
	Graft loss	1 y	CR	0.822	0.083	0.027	0.027
	Graft loss	1 y	SE	0.832	0.076	0.029	0.027
	Graft loss	3 y	CR	0.735	0.053	0.058	0.052
	Graft loss	3 y	SE	0.735	0.044	0.057	0.052
	Graft loss	5 y	CR	0.688	0.036	0.094	0.088
	Graft loss	5 y	SE	0.685	0.029	0.100	0.088
	Graft loss	10 y	CR	0.631	0.004	0.180	0.185
	Graft loss	10 y	SE	0.613	−0.007	0.234	0.185
Donor M / recipient M (4,935)	Death	1 y	CR	0.756	0.052	0.030	0.031
	Death	3 y	CR	0.730	0.057	0.069	0.068
	Death	5 y	CR	0.720	0.074	0.112	0.109
	Death	10 y	CR	0.715	0.141	0.226	0.217
	Graft loss	1 y	CR	0.837	0.103	0.027	0.024
	Graft loss	1 y	SE	0.833	0.112	0.030	0.024
	Graft loss	3 y	CR	0.749	0.075	0.058	0.054
	Graft loss	3 y	SE	0.750	0.086	0.060	0.054
	Graft loss	5 y	CR	0.710	0.054	0.092	0.089
	Graft loss	5 y	SE	0.709	0.059	0.103	0.089
	Graft loss	10 y	CR	0.662	0.033	0.173	0.177
	Graft loss	10 y	SE	0.646	0.020	0.236	0.177

Table S17: Fine–Gray: test grafts whose predicted cumulative incidences of graft loss and of death with a functioning graft sum to more than one.

Horizon	Grafts with $\hat{F}_1 + \hat{F}_2 > 1$	% of test grafts	Maximum $\hat{F}_1 + \hat{F}_2$
1 year	0	0.00	0.596
3 years	0	0.00	0.990
5 years	<11	–	1.242
10 years	146	1.05	1.587

Table S18: Net benefit (IPCW) at selected thresholds, mean of each model family, test grafts. Useful thresholds: those at which the family mean exceeded both intervening in every recipient (treat all) and in none (treat none), among 1, 2, 3, 5, 7.5, 10, 15, 20, 25, 30, 40 and 50%.

Endpoint, horizon	Strategy	5%	10%	20%	30%	Useful thresholds
Graft loss, 5 years	Treat all	0.0460	−0.0070	−0.1329	−0.2947	
	Single-event	0.0487	0.0218	0.0079	0.0032	5–40%
	Competing-risks	0.0473	0.0205	0.0064	0.0017	5–40%
Graft loss, 10 years	Treat all	0.1422	0.0945	−0.0186	−0.1641	
	Single-event	0.1420	0.0953	0.0225	0.0046	3, 10–30%
	Competing-risks	0.1410	0.0931	0.0220	0.0042	15–30%
Death, 5 years	Treat all	0.0582	0.0059	−0.1183	−0.2781	
	Competing-risks	0.0656	0.0425	0.0172	0.0050	3–50%
Death, 10 years	Treat all	0.1734	0.1275	0.0185	−0.1217	
	Competing-risks	0.1745	0.1416	0.0936	0.0612	5–50%

Supplementary Figures

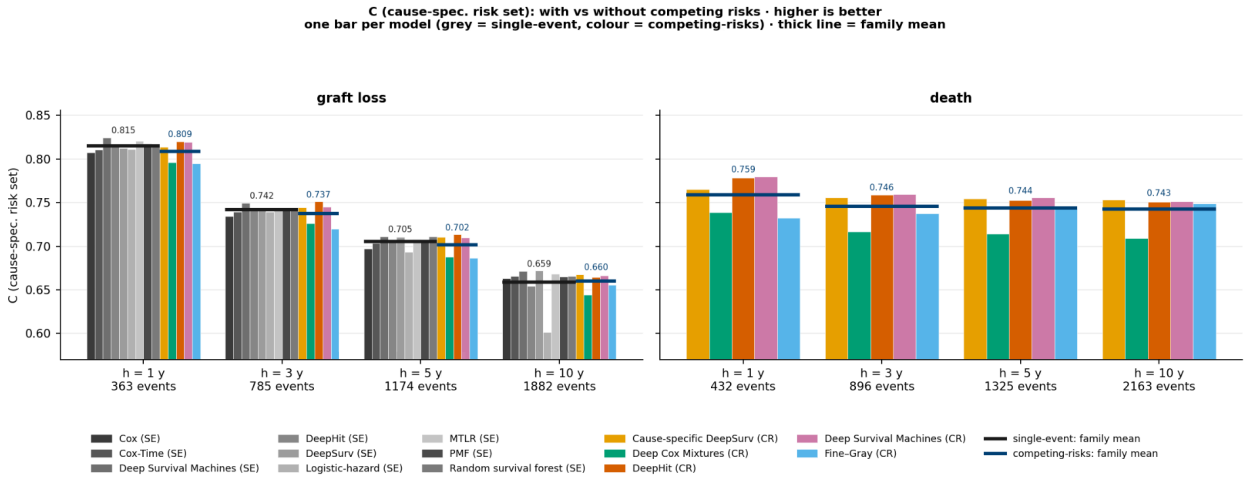

Figure S1: Fixed-horizon concordance with cause-specific comparability (grafts lost to the competing event excluded from the controls), by model and family.

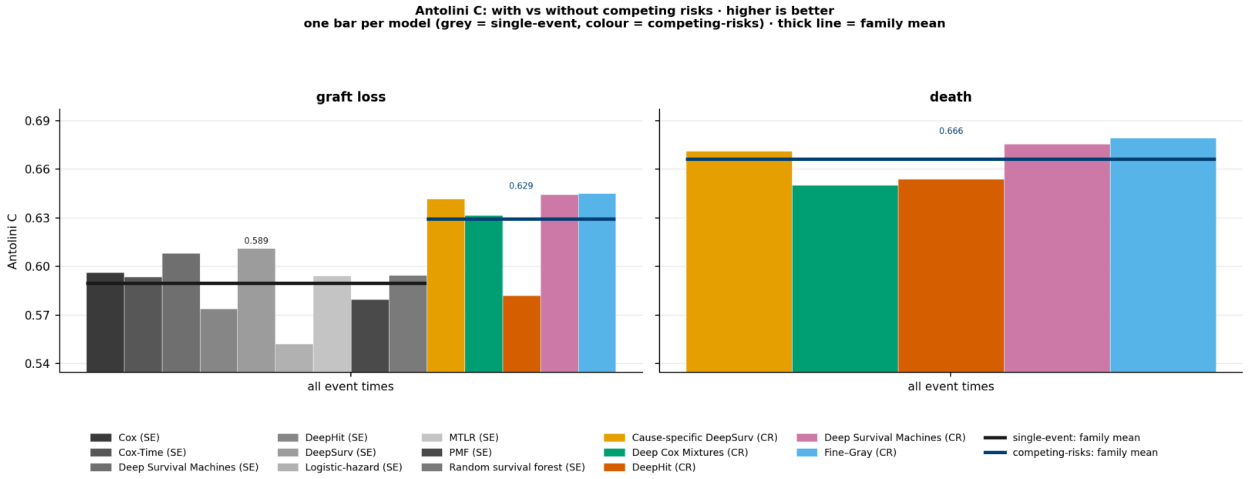

Figure S2: Antolini's time-dependent concordance by model and family.

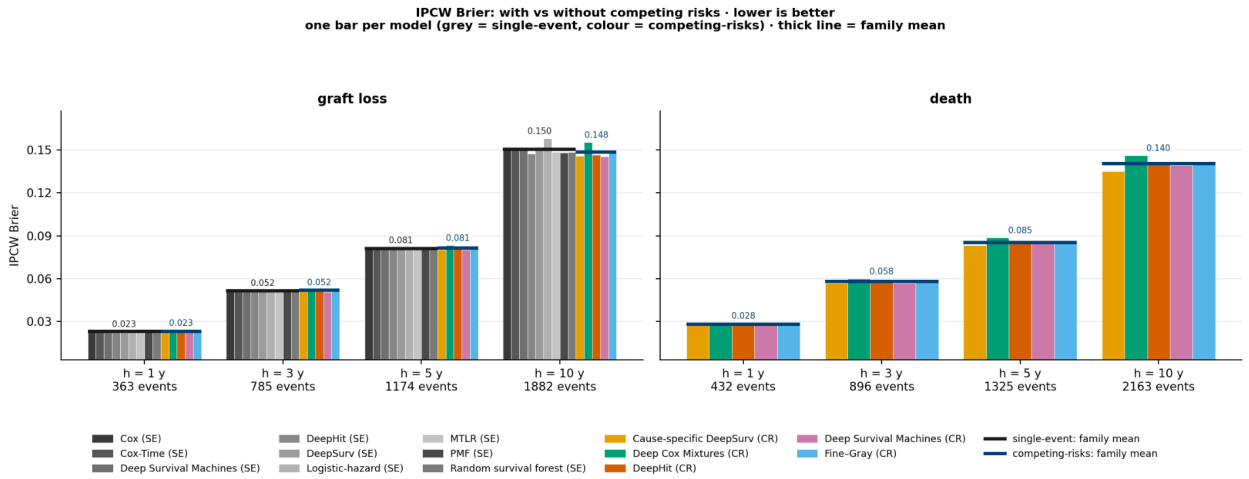

Figure S3: IPCW Brier score by model, family and horizon (lower is better).

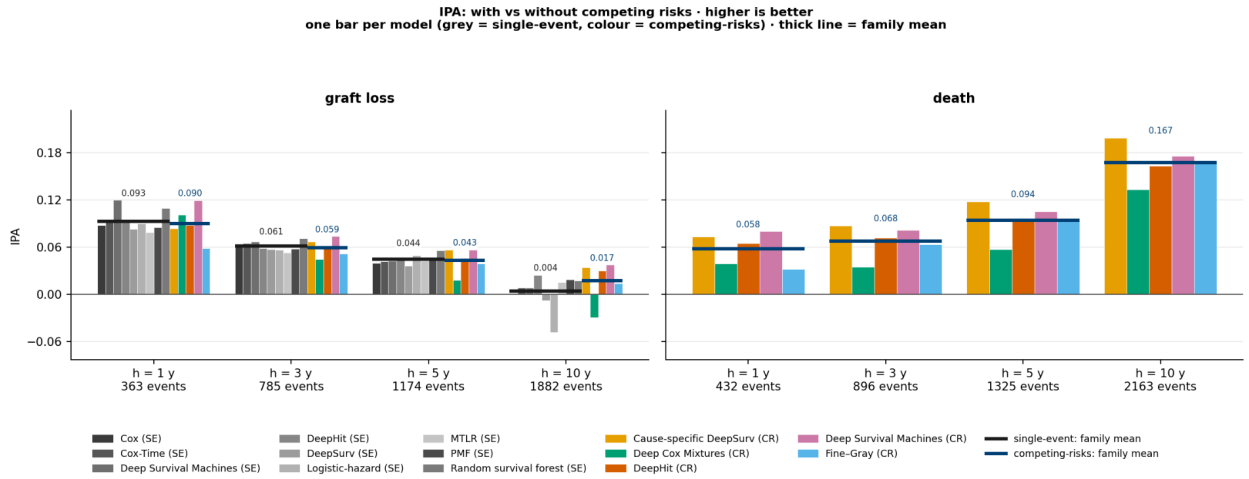

Figure S4: Index of prediction accuracy by model, family and horizon.

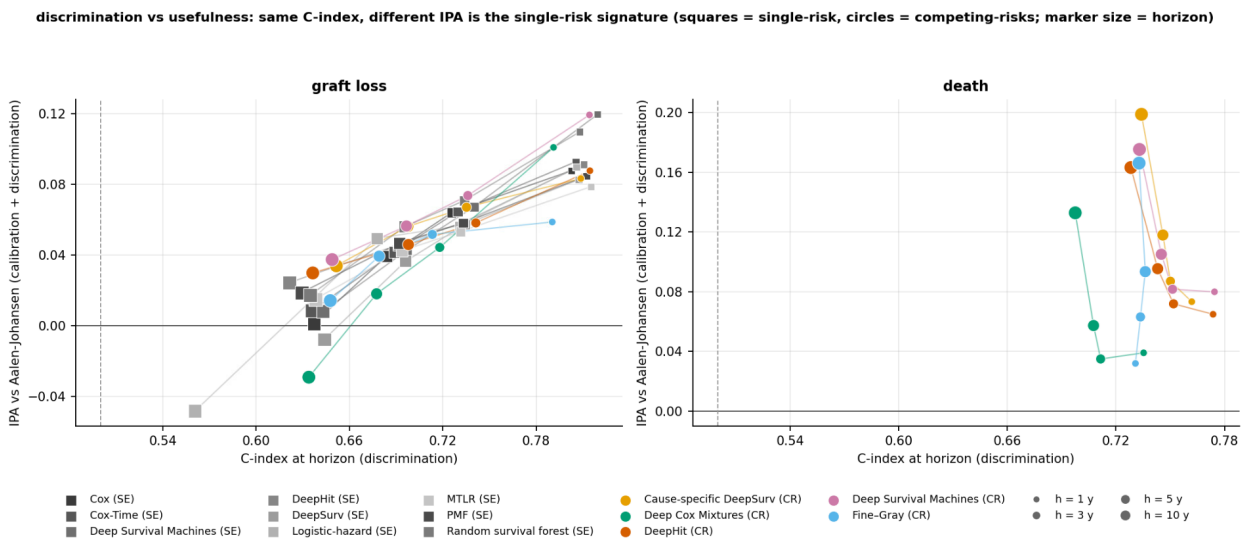

Figure S5: Concordance against IPA for every model and horizon; marker size encodes the horizon. Squares, single-event; circles, competing-risks.

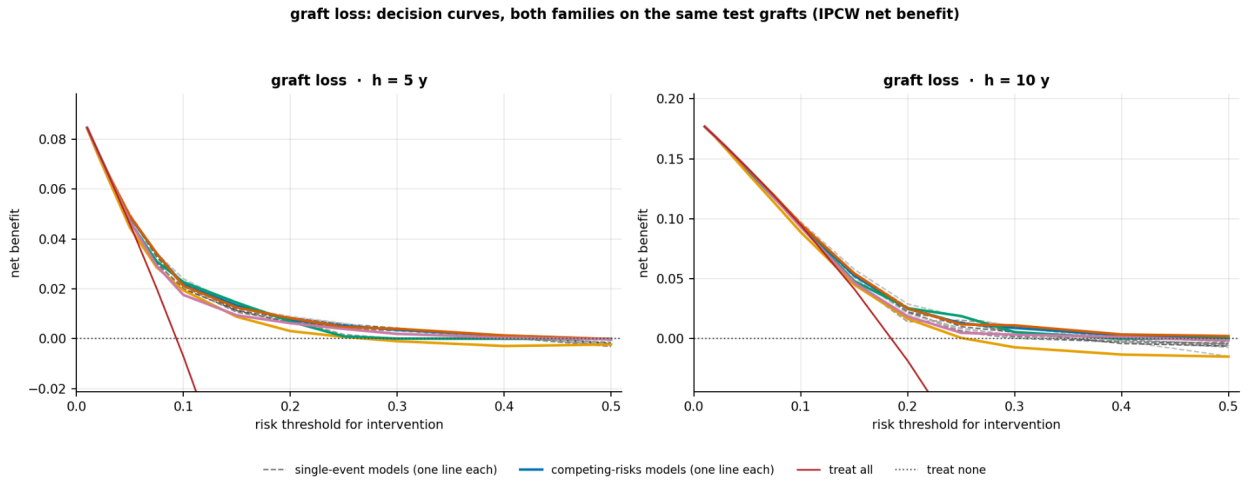

Figure S6: Decision curves for graft loss at 5 and 10 years. Censoring-adjusted net benefit of classifying a recipient as high risk when the predicted risk exceeds the threshold, for single-event models (grey, dashed) and competing-risks models (colour). Reference strategies: classifying every recipient as high risk (treat all, red) and classifying none as high risk (treat none, dotted).

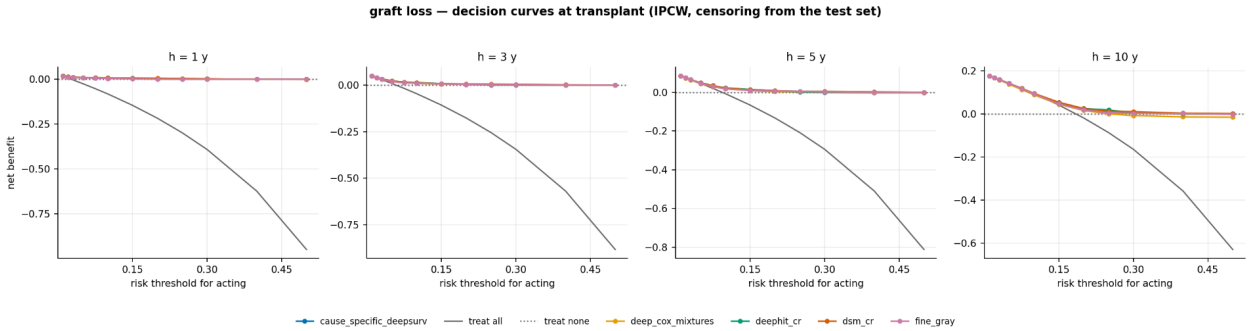

Figure S7: Decision curves for graft loss at every horizon, competing-risks models only.

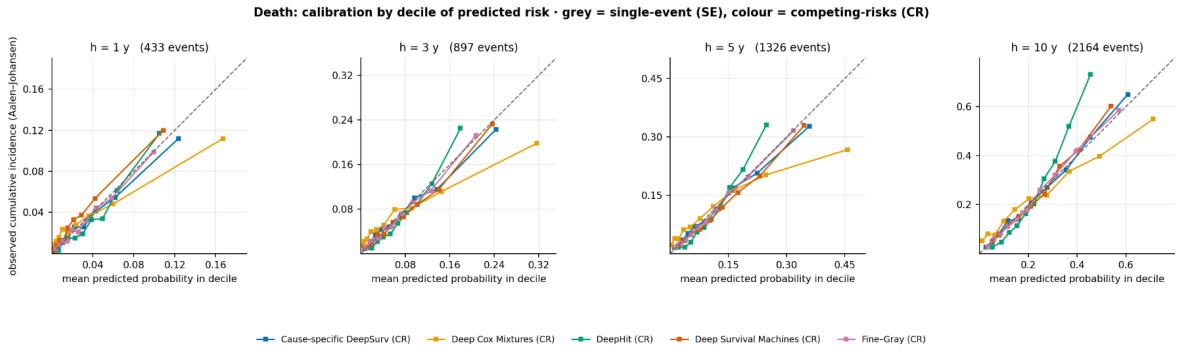

Figure S8: Calibration by decile of predicted risk for death with a functioning graft, competing-risks models.

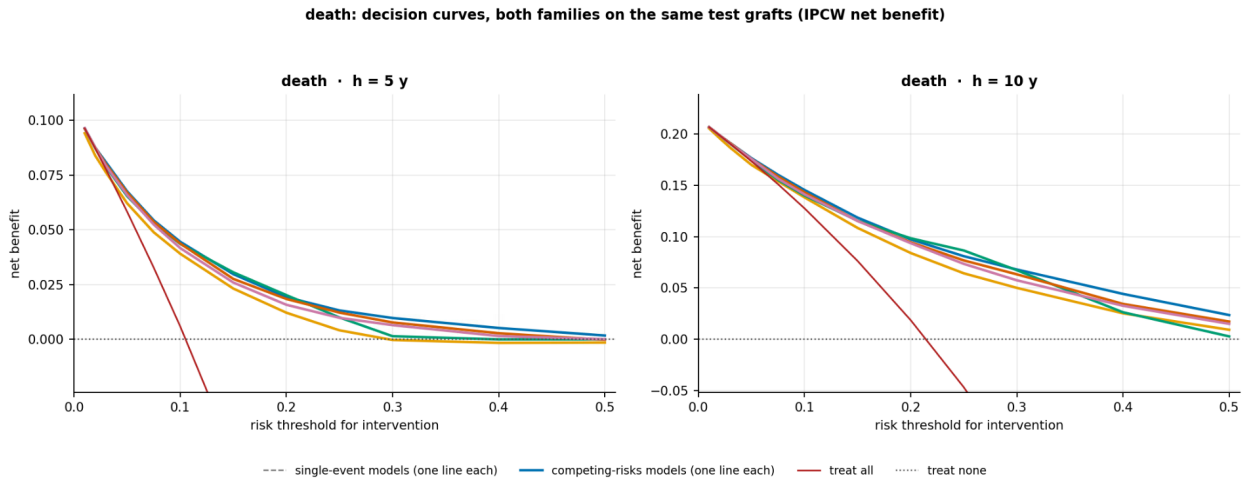

Figure S9: Decision curves for death with a functioning graft at 5 and 10 years, competing-risks models.

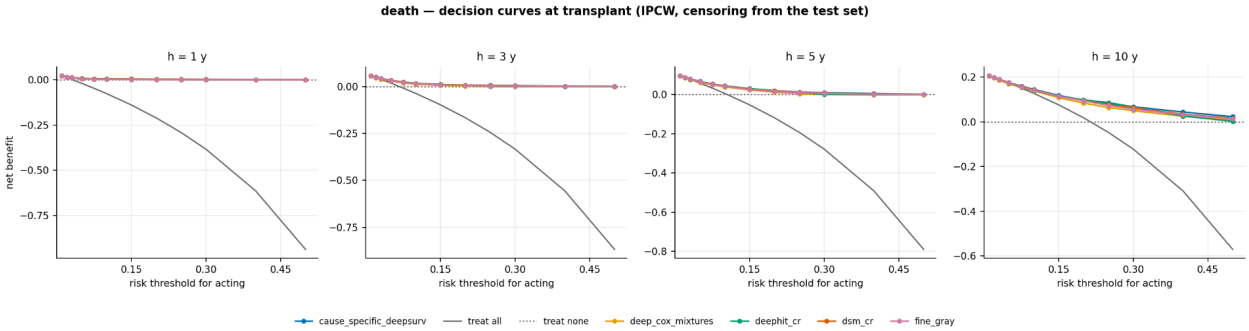

Figure S10: Decision curves for death with a functioning graft at every horizon, competing-risks models.

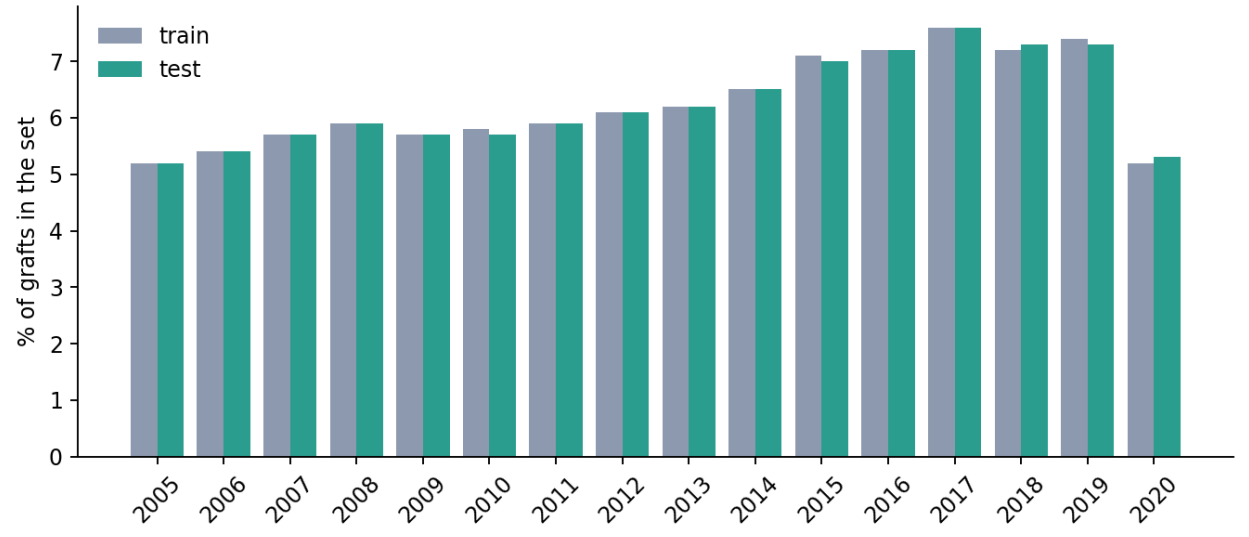

Figure S11: Share of the training and test grafts by year of transplantation (Supplementary Table S10).

Supplementary references

[1] Gary S Collins, Karel G M Moons, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B Reitsma, Maarten van Smeden, Anne-Laure Boulesteix, Jennifer Catherine Camaradou, Leo Anthony Celi, Spiros Denaxas,

Alastair K Denniston, Ben Glocker, Robert M Golub, Hugh Harvey, Georg Heinze, Michael M Hoffman, André Pascal Kengne, Emily Lam, Naomi Lee, Elizabeth W Loder, Lena Maier-Hein, Bilal A Mateen, Melissa D McCradden, Lauren Oakden-Rayner, Johan Ordish, Richard Parnell, Sherri Rose, Karandeep Singh, Laure Wynants, and Patricia Logullo. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, 2024. doi: 10.1136/bmj-2023-078378.

[2] Pierre Marquet, Antoine Humeau, Sabrina Crépin, Clément Benoist, Sylvain Couderc, Caroline Monchaud, Marc Labriffe, and Franck Saint-Marcoux. A pharmaco-epidemiological research program leveraging a nationwide database of kidney transplantation. *Pharmacoepidemiol. Drug Saf.*, 35(8):e70442, 2026. doi: 10.1002/pds.70442.