Supplementary Materials

Machine-learning risk stratification for 12-month mortality and SGLT2 inhibitor-related absolute risk differences in elderly patients with heart failure and type 2 diabetes

Supplementary Table 1. Algorithms, settings and repeat-level performance

Algorithm	Fixed configuration
LogisticL2	L2; C=1; lbfgs; balanced weights; max_iter=5000
ElasticNet	L1 ratio=0.35; C=0.6; saga; balanced weights; max_iter=8000
SVM_RBF	RBF; C=1; gamma=scale; balanced weights; probability=True
KNN	k=21; distance weights; Minkowski p=2
GaussianNB	var_smoothing=1e-9; empirical class priors
DecisionTree	depth=4; min_samples_leaf=12; balanced weights
RandomForest	500 trees; depth=5; leaf=8; max_features=sqrt; balanced_subsample
ExtraTrees	600 trees; depth=5; leaf=8; max_features=sqrt; balanced weights
GradientBoosting	150 trees; learning_rate=0.035; depth=2; leaf=12; subsample=0.8
AdaBoost	220 estimators; learning_rate=0.035; default depth-1 tree; SAMME.R (sklearn 1.5.1)

No hyperparameter search was performed. Numeric imputation/scaling and categorical completion/encoding were fitted within each training fold. Full default configurations and seeds are in estimator_parameters.json. A common set of inner splits was used for all candidates. Values below are mean (SD) across five repeat-level outer-fold evaluations; they are descriptive candidate benchmarks, not independent test-set rankings.

Feature set	Algorithm	AUROC	AUPRC	Brier
CompleteImputed	RandomForest	0.767 (0.005)	0.661 (0.009)	0.197 (0.002)
CompleteImputed	GradientBoosting	0.766 (0.015)	0.667 (0.019)	0.189 (0.006)
GuidelineCore	ElasticNet	0.759 (0.004)	0.649 (0.011)	0.201 (0.002)
CompleteImputed	SVM_RBF	0.759 (0.007)	0.635 (0.016)	0.192 (0.003)
PrimaryAnalysis	RandomForest	0.758 (0.010)	0.637 (0.012)	0.200 (0.003)
CompleteImputed	AdaBoost	0.757 (0.011)	0.649 (0.018)	0.231 (0.001)
GuidelineCore	LogisticL2	0.757 (0.006)	0.644 (0.011)	0.202 (0.002)
CompleteImputed	ExtraTrees	0.754 (0.005)	0.662 (0.006)	0.204 (0.001)
PrimaryAnalysis	GradientBoosting	0.753 (0.016)	0.632 (0.019)	0.195 (0.006)
PrimaryAnalysis	AdaBoost	0.749 (0.011)	0.630 (0.010)	0.232 (0.001)
PrimaryAnalysis	SVM_RBF	0.748 (0.007)	0.605 (0.005)	0.197 (0.002)
GuidelineCore	RandomForest	0.744 (0.008)	0.609 (0.009)	0.202 (0.002)
CompleteImputed	ElasticNet	0.744 (0.017)	0.653 (0.020)	0.217 (0.009)
GuidelineCore	AdaBoost	0.739 (0.010)	0.608 (0.012)	0.233 (0.001)
PrimaryAnalysis	ExtraTrees	0.738 (0.005)	0.631 (0.012)	0.209 (0.001)

Feature set	Algorithm	AUROC	AUPRC	Brier
PrimaryAnalysis	ElasticNet	0.736 (0.019)	0.630 (0.022)	0.217 (0.010)
CompleteImputed	LogisticL2	0.730 (0.020)	0.642 (0.024)	0.230 (0.011)
GuidelineCore	GradientBoosting	0.728 (0.003)	0.590 (0.002)	0.205 (0.001)
GuidelineCore	ExtraTrees	0.726 (0.007)	0.601 (0.006)	0.219 (0.001)
CompleteImputed	KNN	0.725 (0.005)	0.624 (0.018)	0.208 (0.002)
PrimaryAnalysis	KNN	0.725 (0.009)	0.629 (0.023)	0.208 (0.003)
PrimaryAnalysis	LogisticL2	0.720 (0.022)	0.619 (0.023)	0.229 (0.012)
GuidelineCore	SVM_RBF	0.717 (0.010)	0.555 (0.017)	0.206 (0.003)
GuidelineCore	KNN	0.710 (0.015)	0.618 (0.010)	0.205 (0.003)
GuidelineCore	DecisionTree	0.695 (0.016)	0.550 (0.023)	0.229 (0.006)
CompleteImputed	DecisionTree	0.658 (0.007)	0.532 (0.027)	0.261 (0.003)
GuidelineCore	GaussianNB	0.655 (0.012)	0.563 (0.011)	0.296 (0.007)
PrimaryAnalysis	DecisionTree	0.638 (0.012)	0.496 (0.025)	0.272 (0.010)
CompleteImputed	GaussianNB	0.535 (0.004)	0.414 (0.006)	0.365 (0.005)
PrimaryAnalysis	GaussianNB	0.535 (0.004)	0.415 (0.006)	0.364 (0.005)

Nested selection frequency across 25 outer training samples

FeatureSet	Model	SelectedFolds	SelectionFrequency
CompleteImputed	GradientBoosting	6	0.24
CompleteImputed	RandomForest	6	0.24
GuidelineCore	ElasticNet	3	0.12
CompleteImputed	AdaBoost	2	0.08
CompleteImputed	ExtraTrees	2	0.08
CompleteImputed	SVM_RBF	2	0.08
GuidelineCore	RandomForest	1	0.04
PrimaryAnalysis	AdaBoost	1	0.04
PrimaryAnalysis	GradientBoosting	1	0.04
PrimaryAnalysis	RandomForest	1	0.04

Supplementary Table 2. Six-variable collinearity audit

Variable	VIF
Age	1.030
BMI	1.018
NtProBNP	1.322
Cardiac_function_grade	1.176
Albumin	1.068
LVEF	1.155

VIF = 1/(1-R2) from regressing each display variable on the other five. NYHA is encoded 1-4; NT-proBNP uses ln(1+x); numeric gaps are median-completed in the 434-patient cohort for this descriptive audit. This is not a VIF audit of the entire primary feature set. VIF, variance inflation factor.

Supplementary Table 3. Predictor rosters and source missingness

Source missingness uses all 500 supplied records. The retained historical primary roster uses <30% missingness; completion is fitted separately in each training fold. Core is investigator-defined and guideline-informed, not a guideline-mandated seven-variable score.

Predictor	Missing n (%)	Core	Primary	Full
Age	0 (0.0)	Yes	Yes	Yes
Temperature	5 (1.0)		Yes	Yes
BMI	104 (20.8)		Yes	Yes
Admission_pulse	7 (1.4)		Yes	Yes
Admission_SBP	6 (1.2)		Yes	Yes
Admission_DBP	6 (1.2)		Yes	Yes
Gender	0 (0.0)	Yes	Yes	Yes
Cardiac_function_grade	0 (0.0)		Yes	Yes
Hypertension	0 (0.0)		Yes	Yes
Hyperlipidemia	0 (0.0)		Yes	Yes
CKD	0 (0.0)		Yes	Yes
Stroke	0 (0.0)		Yes	Yes
Malignant_tumor	0 (0.0)		Yes	Yes
Atrial_fibrillation	0 (0.0)		Yes	Yes
Arterial_plaque	0 (0.0)		Yes	Yes
Myocardial_infarction	0 (0.0)	Yes	Yes	Yes
Stable_angina	0 (0.0)		Yes	Yes
Prior_revascularization	0 (0.0)		Yes	Yes
Cardiac_function	0 (0.0)		Yes	Yes
WBC_count	4 (0.8)		Yes	Yes
Neutrophil_count	4 (0.8)		Yes	Yes
Lymphocyte_count	4 (0.8)		Yes	Yes
Monocyte_count	4 (0.8)		Yes	Yes

Predictor	Missing n (%)	Core	Primary	Full
Hemoglobin	4 (0.8)		Yes	Yes
Hematocrit	4 (0.8)		Yes	Yes
Platelet_count	5 (1.0)		Yes	Yes
Fibrinogen	7 (1.4)		Yes	Yes
D_dimer	11 (2.2)		Yes	Yes
Potassium	5 (1.0)		Yes	Yes
Sodium	5 (1.0)		Yes	Yes
Chloride	5 (1.0)		Yes	Yes
Phosphorus	13 (2.6)		Yes	Yes
Calcium	7 (1.4)		Yes	Yes
ALT	11 (2.2)		Yes	Yes
AST	9 (1.8)		Yes	Yes
Total_bilirubin	11 (2.2)		Yes	Yes
Albumin	10 (2.0)	Yes	Yes	Yes
Uric_acid	7 (1.4)		Yes	Yes
BUN	8 (1.6)		Yes	Yes
Creatinine	7 (1.4)		Yes	Yes
eGFR	8 (1.6)	Yes	Yes	Yes
NtProBNP	12 (2.4)	Yes	Yes	Yes
Troponin	18 (3.6)		Yes	Yes
Fasting_glucose	86 (17.2)		Yes	Yes
Triglyceride	43 (8.6)		Yes	Yes
Total_cholesterol	43 (8.6)		Yes	Yes
LDL_C	44 (8.8)		Yes	Yes
HDL_C	44 (8.8)		Yes	Yes
Apolipoprotein_A1	262 (52.4)			Yes
Apolipoprotein_B	262 (52.4)			Yes
Lipoprotein_a	260 (52.0)			Yes
HbA1c	83 (16.6)		Yes	Yes
CRP	161 (32.2)			Yes
PCT	159 (31.8)			Yes
LVEF	32 (6.4)	Yes	Yes	Yes
LAD	98 (19.6)		Yes	Yes
LVDD	158 (31.6)			Yes
IVSTD	195 (39.0)			Yes

Supplementary Table 4. Retained original descriptive treatment comparison

N=358; discharge survivors observed for at least 30 days; fitted-reference-score tertiles; mortality over all available follow-up. This original descriptive analysis is separate from the adjusted 365-day estimates in main Table 3.

Risk stratum	N	SGLT2i users, n	Non-users, n	User mortality, % (95% CI)	Non-user mortality, % (95% CI)	ARR, pp (95% CI)	NNT
Low	120	71	49	2.8 (0.8–9.7)	6.1 (2.1–16.5)	3.3 (-4.7 to 13.9)	31
Intermediate	119	62	57	12.9 (6.7–23.4)	26.3 (16.6–39.0)	13.4 (-0.9 to 27.5)	8
High	119	60	59	41.7 (30.1–54.3)	61.0 (48.3–72.4)	19.4 (1.4 to 35.6)	6

Original ARR point estimates were 3.3, 13.4 and 19.4 pp, with NNT 31, 8 and 6. Wilson and Newcombe intervals are conditional on the original strata. The revised cohort, score source, grouping and time horizon differ, so changes in estimates are not attributable to weighting alone. ARR and NNT summarize observed differences, not established causal treatment effects.

Supplementary Table 5. Adjusted discharge-treatment analysis

New analysis: 381 patients, including early post-discharge deaths; 365 days from discharge; overlap weights; out-of-sample fixed-reference risk tertiles. Original fitted-score/full-follow-up summaries are different analyses. The bootstrap refits propensity models in 1,000 treatment-stratified patient samples, conditional on fixed scores and strata.

Stratum	Users / controls	User risk % (95% CI)	Control risk % (95% CI)	RD pp (95% CI)
Overall	198/183	16.1 (10.4 to 22.5)	31.1 (24.0 to 38.8)	15.0 (5.9 to 23.9)
Low	70/57	8.0 (0.4 to 17.0)	14.5 (4.3 to 23.1)	6.5 (-6.5 to 17.4)
Intermediate	71/56	17.3 (6.5 to 30.2)	35.4 (20.0 to 53.5)	18.1 (-4.5 to 39.6)
High	57/70	32.9 (17.5 to 47.9)	52.6 (38.8 to 65.8)	19.7 (-1.7 to 40.6)

Balance and effective sample size

Stratum	Max SMD before	Max SMD after	ESS users / controls	Propensity range
Overall	0.531	0.0007	156.6/147.1	0.0622-0.9206
Low	0.611	0.0009	52.5/45.7	0.0597-0.9559
Intermediate	0.649	0.0011	46.5/37.5	0.0142-0.9999
High	0.750	0.0015	39.5/42.5	0.0003-0.9468

The overall difference was 15.0 pp (95% CI 5.9-23.9). All three stratum-specific intervals included zero. The high-minus-low contrast was 13.2 pp (95% CI -10.0 to 36.2). SMD, standardized mean difference; ESS, effective sample size; RD, non-user minus user risk difference. Overlap weighting addresses recorded covariates, not unmeasured prescribing reasons.

Covariate balance in the propensity models

Covariate	Overall	Low	Intermediate	High
Admission year	0.531 / 0.0000	0.611 / 0.0003	0.649 / 0.0002	0.296 / 0.0003
Age	0.166 / 0.0000	0.009 / 0.0000	0.350 / 0.0002	0.031 / 0.0001
BMI	0.028 / 0.0000	0.135 / 0.0004	0.088 / 0.0000	0.020 / 0.0002
Admission SBP	0.088 / 0.0001	0.121 / 0.0003	0.340 / 0.0004	0.025 / 0.0001
Admission pulse	0.101 / 0.0001	0.057 / 0.0002	0.108 / 0.0001	0.416 / 0.0001
LVEF	0.365 / 0.0001	0.439 / 0.0003	0.233 / 0.0003	0.583 / 0.0001
ln(1+NT-proBNP)	0.055 / 0.0002	0.232 / 0.0003	0.359 / 0.0000	0.200 / 0.0004
eGFR	0.448 / 0.0001	0.059 / 0.0002	0.561 / 0.0001	0.750 / 0.0003
Albumin	0.048 / 0.0000	0.073 / 0.0001	0.044 / 0.0002	0.121 / 0.0000
HbA1c	0.277 / 0.0000	0.226 / 0.0001	0.081 / 0.0002	0.508 / 0.0000
Gender Male	0.196 / 0.0001	0.323 / 0.0002	0.174 / 0.0002	0.132 / 0.0010
NYHA I	0.047 / 0.0002	0.122 / 0.0007	0.241 / 0.0005	0.243 / 0.0003
NYHA II	0.203 / 0.0002	0.254 / 0.0002	0.249 / 0.0004	0.088 / 0.0015
NYHA III	0.123 / 0.0001	0.073 / 0.0003	0.407 / 0.0008	0.010 / 0.0009
NYHA IV	0.040 / 0.0003	0.255 / 0.0003	0.134 / 0.0011	0.101 / 0.0001
CKD Yes	0.420 / 0.0001	0.193 / 0.0004	0.417 / 0.0008	0.607 / 0.0004
Atrial fibrillation Yes	0.107 / 0.0002	0.086 / 0.0001	0.183 / 0.0006	0.059 / 0.0004
Myocardial infarction Yes	0.218 / 0.0007	0.179 / 0.0009	0.429 / 0.0002	0.197 / 0.0000

Each cell gives absolute SMD before / after overlap weighting. A common original pooled SD was used before and after weighting; categorical NYHA levels are encoded separately. These diagnostics describe measured-covariate balance, not balance in unrecorded treatment eligibility.

Supplementary Table 6. Surrogate formula and validation

For transformed input z_j, standardized input u_j=(z_j-mean_j)/scale_j. Linear predictor eta=intercept+sum(coefficient_j*u_j); predicted risk=1/(1+exp(-eta)). Intercept=-0.616021940291. Ridge alpha=2. NT-proBNP uses z=ln(1+raw pg/mL); NYHA I/II/III/IV maps to 1/2/3/4. All other variables use their entered values.

Variable / unit	Median z	Mean z	Scale z	Coefficient
Age (years)	80	80.723502	4.3437598	0.2150712954
BMI (kg/m2)	23.845	24.031313	3.6997051	-0.2332695811
NtProBNP (ln(1+pg/mL))	8.3437917	8.2842735	1.361855	0.5360860905
Cardiac_function_grade (I-IV -> 1-4)	3	3.124424	0.81642289	0.2697389563
Albumin (g/L)	36.8	36.348157	4.8284066	-0.2172960513
LVEF (%)	52	50.175115	11.263158	-0.09621950221

The machine-readable payload retains full precision. These are final deployment-fit coefficients trained on full-cohort out-of-fold teacher targets; they do not produce the held-out validation predictions by resubstitution. Validation repeated the complete teacher-target and student-fitting process inside outer training folds.

Held-out measure	Value
AUROC	0.757855
AUPRC	0.666913
Brier	0.192568
LogLoss	0.565613
CalibrationIntercept	0.162754
CalibrationSlope	1.210590
heldout_pearson_to_teacher	0.820876
heldout_mae_to_teacher	0.097693
heldout_rmse_to_teacher	0.124776

Absolute-error measures are probabilities (multiply by 100 for percentage points). Correlation measures association, not interchangeability. Validation metrics use the average of each patient's five held-out probabilities.

Supplementary Table 7. Time-to-event sensitivity analysis

N=469 with 167 deaths through admission day365 and 35 earlier censored observations. The seven core predictors were used; numeric missingness indicators were retained. Numeric predictors were completed/scaled in training data; categorical predictors used reference coding. NT-proBNP used ln(1+x). Estimates below use repeated 5x5 held-out validation. Cox estimates do not compare the same cohort or full predictor set as the main machine-learning workflow.

Metric	Mean repeat	SD
Harrell concordance	0.724	0.008
IPCW AUROC365	0.755	0.008

Risk stratum	N	Deaths	Early censored	KM risk % (95% CI)
Low	157	20	14	12.9 (8.5 to 19.2)
Intermediate	156	57	12	36.8 (29.7 to 44.9)
High	156	90	9	57.7 (50.1 to 65.6)

Strata use tertiles of averaged held-out Cox probabilities, not the SGLT2i strata. Log-rank P<0.001. IPCW uses the censoring Kaplan-Meier estimate immediately before the event/evaluation time, excluding unknown binary labels while retaining their censoring information.

Final cohort fit for clinical coefficients

Predictor	HR (95% CI)	P
Age	1.35 (1.17-1.55)	<0.001
Albumin	0.73 (0.62-0.85)	<0.001
LVEF	0.93 (0.79-1.10)	0.399
ln(1+NT-proBNP)	1.76 (1.41-2.21)	<0.001
eGFR	0.96 (0.81-1.14)	0.646
Male versus female	0.91 (0.66-1.25)	0.550
Previous MI, yes versus no	1.70 (1.23-2.33)	0.001

Continuous-variable HRs are per one training-cohort SD in the final fit. Four standardized missingness indicators were included but are not listed as clinical predictors. Coefficient intervals refer to the final cohort fit, not external validation. Rank-based proportional-hazards tests had minimum P=0.067; the full diagnostic output accompanies the code/source data.

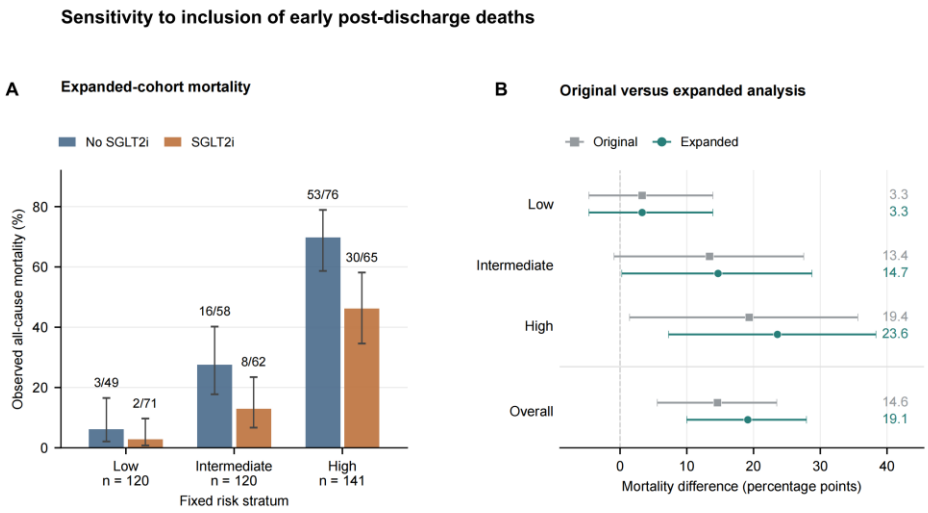
Supplementary Table 8. Multiple-imputation prediction sensitivity

This is a fixed gradient-boosting, 51-variable sensitivity analysis, not a 30-candidate nested MI selection analysis. Twenty stochastic numeric imputations were generated per outer training fold using Bayesian ridge posterior sampling; validation outcomes were not used. Complete categorical covariates served as encoded auxiliary variables. Imputation models and support bounds were fitted on training data. Observed values, outcomes and treatment fields were not imputed or changed. Held-out probabilities were averaged across imputations, then evaluated within each complete validation repeat.

Workflow	AUROC	AUPRC	Brier	Calibration slope
Single	0.753	0.632	0.195	0.885
MI5	0.755	0.630	0.194	0.916
MI10	0.757	0.633	0.194	0.930
MI20	0.758	0.634	0.193	0.935

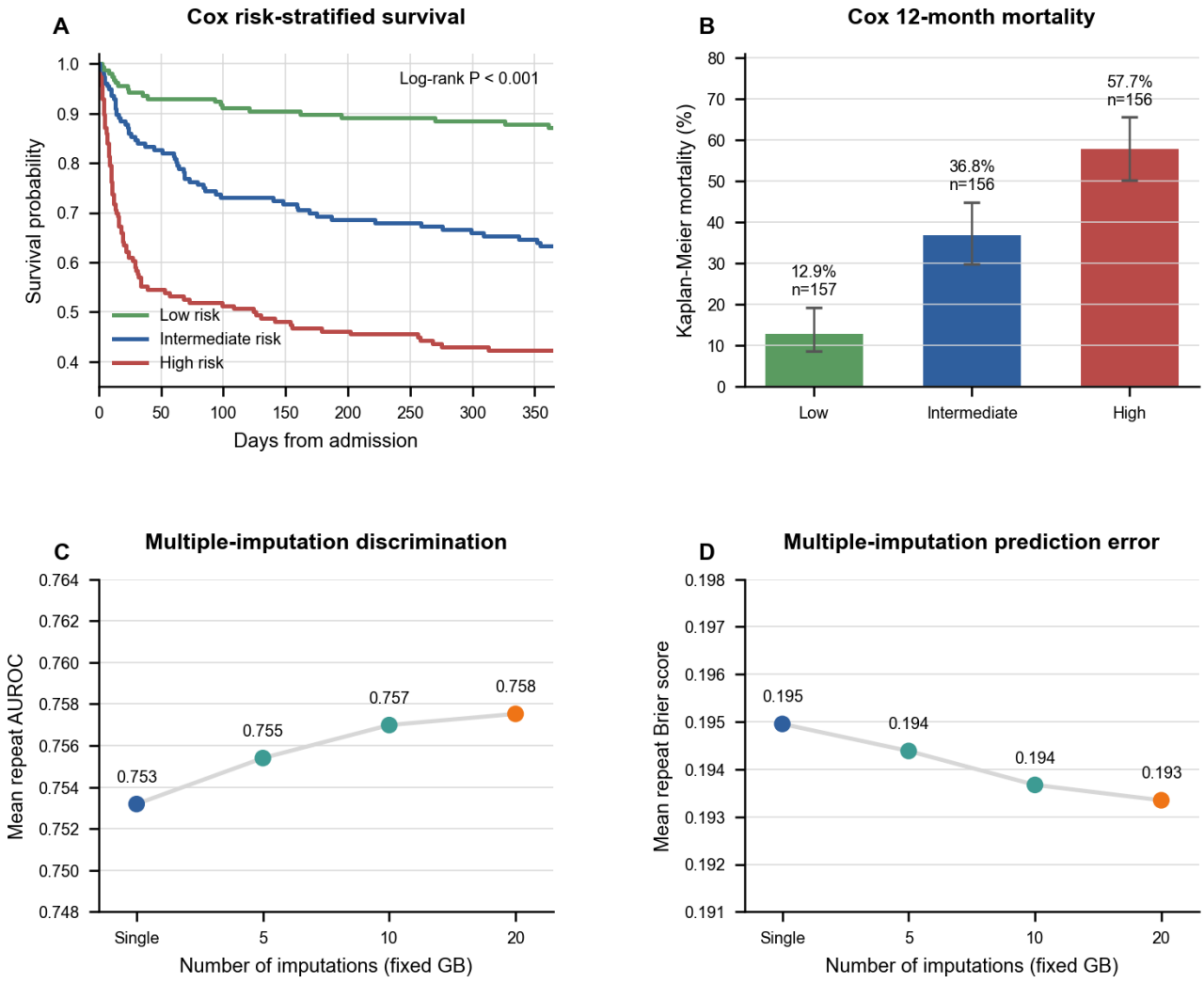
Values are mean repeat estimates. The mean absolute difference between MI10 and MI20 averaged probabilities was 0.37 pp. The average within-patient, within-repeat SD across imputation-specific probabilities was 4.57 pp. These summaries describe numerical sensitivity, not a formal guarantee of imputation-chain convergence. The single-imputation baseline reproduced the archived predictions with maximum difference 1.7e-16.

Supplementary Figure S1. Sensitivity to inclusion of early post-discharge deaths.



Removing the 30-day survival restriction added 23 early deaths (18 non-users and five SGLT2i users), expanding the cohort from 358 to 381 patients. All other eligibility criteria and the original risk-score boundaries were held fixed; no original patient was removed or reassigned. (A) Observed all-cause mortality proportions in the expanded low-, intermediate- and high-risk strata. Bar labels show deaths/number of patients; error bars indicate Wilson 95% confidence intervals. (B) Observed mortality differences (non-user minus user mortality), in percentage points, in the original and expanded cohorts. Squares denote the original analysis and circles the expanded analysis. Horizontal lines indicate Newcombe 95% confidence intervals conditional on the fixed risk assignments; the dashed line marks zero. The cohorts overlap, and no independent-samples test between cohorts was performed. Outcomes cover the entire available follow-up, not a fixed 12-month endpoint. These are unadjusted descriptive comparisons, not estimates of causal treatment benefit. SGLT2i, sodium-glucose cotransporter 2 inhibitor.

Supplementary Figure S2. Survival and multiple-imputation sensitivity analyses



Supplementary Fig. S2 | Survival and multiple-imputation sensitivity analyses. (A) Kaplan–Meier survival across tertiles of averaged held-out Cox predictions in 469 patients with verified follow-up. Admission is time zero; follow-up ends at death, last confirmed contact or day 365. Thirty-five short event-free observations are retained as right-censored records. Log-rank P<0.001. (B) Kaplan–Meier 12-month mortality estimates with 95% confidence intervals; labels show mortality and original group sizes. These Cox strata are distinct from the SGLT2i treatment strata. (C,D) Mean repeat AUROC and Brier score for the fixed gradient-boosting model with the 51-variable primary analysis set using single training-fold median/mode imputation or 5, 10 and 20 stochastic numeric imputations. The same outer validation splits were used; validation outcomes did not enter imputation. Predictions were averaged across imputations before repeat-level evaluation. These panels assess the fixed model, not nested selection among 30 candidates. AUROC, area under the receiver operating characteristic curve; GB, gradient boosting; MI, multiple imputation.