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.

Temporal Validation of Machine-Learning Models for Predicting Postoperative Atrial Fibrillation After Valve or Coronary Artery Bypass Surgery

Yalu Yu
School of Medicine, University of Electronic Science and Technology of China, 610072 Chengdu, Sichuan, China

Jiaqi Xia
363 HospitalDepartment of Cardiothoracic and Breast Surgery, 363 Hospital

Cheng Yang
Macau University of Science and Technology

Qin Jiang
jq349@163.com

Sichuan Provincial People's Hospital

Research Article

Keywords: postoperative atrial fibrillation, cardiac surgery, left atrial volume index, machine learning, temporal validation, calibration, prognosis

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11295554/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Postoperative atrial fibrillation (POAF) is common after valve or coronary artery bypass grafting (CABG) surgery, but prediction models often lack temporal validation and calibration assessment. We hypothesized that left atrial volume index (LAVI), as a marker of chronic atrial substrate, would be more transportable than perioperative stress variables and that regularized models would generalize better than complex algorithms under temporal shift.

Methods

We used a case-enriched retrospective development cohort of 347 patients (189 POAF events; May 2022–May 2025) and a prospective temporal cohort of 167 consecutive patients (44 events; June–December 2025). Eleven classifiers were assessed by repeated stratified five-fold cross-validation with fold-contained preprocessing, then frozen and applied to the temporal cohort. We evaluated discrimination, calibration, decision-curve net benefit, explainability, independent predictors, and six-month prognosis.

Results

LAVI was the strongest perioperative predictor (adjusted odds ratio per SD, 3.19; 95% CI, 1.66–6.12; P < 0.001). Random forest had the highest internal AUROC (0.827; 95% CI, 0.783–0.870), whereas the frozen preoperative ridge model performed best temporally (AUROC, 0.723; 95% CI, 0.639–0.808). The case-enriched development sample caused absolute-risk overprediction (mean predicted 0.442 vs observed 0.263; integrated calibration index [ICI], 0.173); intercept-only recalibration improved ICI to 0.048. Platt scaling fitted in the temporal cohort yielded an apparent ICI of 0.009 and was treated as exploratory updating. DCA indicates that within the decision threshold range of 0.10–0.50, the clinical benefit is greater than that of the "all-treatment/no-treatment" strategy. POAF was associated with six-month AF recurrence (7/43 vs 1/112; odds ratio, 21.58) and a higher composite-event hazard (adjusted hazard ratio, 3.22; 95% CI, 1.00–10.34; P = 0.050).

Conclusions

LAVI captured a transportable atrial-substrate signal, while the internally best nonlinear model did not retain superiority under temporal shift. The preoperative ridge model requires recalibration and independent multicenter validation before deployment. POAF identified persistent arrhythmic vulnerability through six months.

1. Introduction

Postoperative atrial fibrillation (POAF) is the most frequent sustained arrhythmia after adult cardiac surgery. Its occurrence varies with procedure mix, from approximately 15%–40% after isolated CABG to 30%–50% after valve surgery and still higher after combined procedures [1, 2]. Although this arrhythmia has traditionally been regarded as transient and self-limiting, with spontaneous conversion to sinus rhythm generally achieved before patient discharge, an increasing volume of empirical evidence suggests that it should not be classified as a benign postoperative complication. POAF can lead to severe adverse outcomes, including an increased risk of stroke, a higher incidence of perioperative complications (such as heart failure and renal insufficiency), prolonged mechanical ventilation duration, extended ICU length of stay, increased postoperative hospitalization duration, greater utilization of medical resources, and elevated long-term mortality rates [3–5]. It also predicts later atrial fibrillation (AF), challenging the historical view that it is a benign and self-limited perioperative event [6].

The pathophysiology reflects an interaction between a pre-existing atrial substrate and acute surgical triggers. Chronic pressure or volume loading, fibrosis, and chamber enlargement reduce conduction reserve, while surgical trauma-induced inflammation, oxidative stress, autonomic imbalance, ischemia–reperfusion injury, electrolyte disturbance, and catecholamine exposure promote initiation of AF [7, 8]. Pericardial-fluid biomarker studies further support an important local inflammatory component [9]. Recent mechanistic investigations implicates macrophage-mediated IL-6–STAT3–CaMKII signaling and ryanodine receptor-2 calcium leak, while atrial proteomic studies suggest metabolic and redox remodeling in patients who develop POAF [10, 11]. Clinically, LAVI and IL-6 have emerged as potentially informative markers of this substrate–trigger interaction [12].

Although POAF has imposed a substantial burden on clinical diagnosis and treatment, available risk scores and prediction models have limitations. Traditional risk prediction of POAF is primarily based on established clinical scoring systems, including the CHA_2DS_2-VASc score, the Society of Thoracic Surgeons (STS) POAF score, and the McSPI AF risk index. However, these tools are mainly developed for non-surgical AF or specific surgical subgroups. Their discriminative performance among heterogeneous patient populations undergoing cardiac surgery remains constrained, which imposes obstacles to the implementation of truly individualized prevention strategies [13]. A systematic review found substantial heterogeneity in predictor definition, validation, and reporting, with relatively little prospective evaluation [14]. Machine-learning studies have reported promising internal performance, including in mixed intensive-care cohorts and after mitral valve surgery or off-pump CABG [15–17]. Nevertheless, machine learning does not consistently outperform well-specified logistic regression in clinical prediction, particularly in modest samples or under dataset shift [18]. Contemporary reporting and risk-of-bias frameworks therefore emphasize leakage control, transparent preprocessing, calibration, and evaluation in data separated by place or time [19, 20].

We evaluated three pre-specified hypotheses: (1) chronic atrial substrate, captured principally by LAVI, is a stronger and more stable determinant of POAF than intraoperative stress alone; (2) although nonlinear

algorithms may achieve higher internal discrimination, parsimonious regularized models will transport better to a later clinical cohort; and (3) POAF identifies persistent atrial vulnerability and therefore predicts six-month arrhythmia recurrence and adverse events. The study's innovation lies in connecting etiologic risk-factor analysis, leakage-audited model comparison, prospective temporal validation, calibration under prevalence shift, explainability, decision-curve analysis, and six-month prognosis within one valve/CABG cohort by machine-learning.

2. Methods

2.1 Study design, setting, and reporting framework

This was a single-center retrospective development study followed by prospective temporal validation at the Department of Cardiac Surgery, Sichuan Provincial People's Hospital. Patients treated from May 2022 to May 2025 formed a case-enriched development cohort; consecutive patients treated from June to December 2025 formed a later prospective temporal cohort. Because both cohorts originated from the same center, the latter is described as temporal validation rather than geographic external validation.

The study was conducted in accordance with the Declaration of Helsinki. Reporting was aligned with TRIPOD + AI, and potential sources of bias were considered using PROBAST + AI [19, 20]. The research protocol has been approved by the Institutional Review Board of this hospital. The retrospective component used de-identified routine data; prospective participants provided written informed consent.

2.2 Participants

Adults aged 18 years or older were eligible if they underwent elective valve repair or replacement, on-pump or off-pump CABG, or combined valve–CABG surgery through sternotomy, partial sternotomy, thoracotomy, or thoracoscopic access. Preoperative sinus rhythm was required, with no documented history of AF or atrial flutter. The onset of the outcome event is restricted to the postoperative hospitalization period, and the diagnostic approaches include postoperative bedside continuous real-time electrocardiography monitoring, conventional electrocardiography, Holter monitoring, as well as auxiliary confirmation by echocardiography and laboratory examinations. Patients were excluded for history of previous AF, atrial flutter, or other sustained atrial arrhythmiasa; permanent pacemaker or implantable cardioverter-defibrillator; concomitant maze surgery or AF ablation; heart transplantation, ventricular assist device implantation, or other mechanical circulatory support; cardiac procedures outside the prespecified valve/CABG spectrum; perioperative death before ascertainment of POAF; or more than 20% missingness across core candidate predictors.

2.3 Outcome definitions and follow-up

The primary outcome was new-onset POAF during the index hospitalization, defined as AF or atrial flutter lasting at least 30 seconds on continuous bedside telemetry, a standard 12-lead electrocardiogram, or Holter recording, or an episode requiring antiarrhythmic treatment or electrical cardioversion. This

operational definition was aligned with cardiac-surgery guidance [21] and was applied to patients with no preoperative history of AF or flutter.

Secondary short-term outcomes included 30-day stroke or transient ischemic attack, worsening heart failure, renal failure, reoperation for bleeding or severe infection, and death. The prospective cohort was followed by clinic visit or telephone at approximately 1, 3, and 6 months. Six-month outcomes included documented AF/atrial flutter recurrence, cardiovascular readmission, all-cause death, and a prespecified composite of adverse events. Prevention, anticoagulation, and rhythm-management decisions remained clinician directed and were not assigned by the model, consistent with contemporary AF guidance [22, 23].

2.4 Candidate predictors and data quality

Candidate variables were chosen before model fitting from clinical knowledge, availability, and prior literature. They included age, sex, body mass index (BMI), comorbidities, New York Heart Association class, preoperative rhythm descriptors, hemoglobin, leukocyte count, potassium, C-reactive protein (CRP), troponin, BNP/NT-proBNP, echocardiographic left atrial diameter (LAD), LAVI, left ventricular ejection fraction (LVEF), valve and coronary disease characteristics, preoperative medications, operative approach and procedure type, cardiopulmonary bypass (CPB), aortic cross-clamp time, operative duration, blood loss, left atrial appendage management, mechanical ventilation, and intensive-care/total postoperative hospital stay.

Structural non-applicability was distinguished from random missingness: for example, graft variables are undefined in patients who did not undergo CABG, and CPB time is not applicable to off-pump surgery. Variables with more than 40% non-structural missingness were excluded from prediction modeling. Continuous variables were median-imputed and standardized, and categorical variables were mode-imputed and one-hot encoded. Crucially, all imputation, scaling, coding, feature filtering, and hyperparameter tuning were estimated within each training fold and applied without refitting to its assessment fold. Category labels were harmonized before modeling; duplicate and contradictory records were removed. This leakage audit eliminated spurious categories with 0% or 100% POAF rates.

2.5 Regression analyses

Univariable logistic regression screened clinically plausible associations. Multivariable preoperative and perioperative logistic models were then fitted to estimate adjusted odds ratios (aORs) and 95% confidence intervals. Continuous predictors were standardized so that odds ratios represent a one-SD increase. Multiple imputation was used for regression analyses, and estimates were pooled across 10 imputed datasets. Variance inflation factors were examined; strongly collinear variables were not interpreted as mutually independent effects. A sensitivity analysis redefined the endpoint as POAF occurring within 72 hours.

2.6 Model development and internal validation

The development cohort was not split once into arbitrary 70:30 subsets. Instead, internal performance was estimated from out-of-fold predictions generated by repeated stratified five-fold cross-validation (five repeats). This design used all available development observations for both training and assessment while maintaining fold-level separation and reducing variance relative to a single split.

Two predictor sets were considered: a preoperative model using information available before surgery and a perioperative model that additionally included intraoperative data. The candidate algorithms were standard logistic regression, LASSO logistic regression, ridge logistic regression, elastic-net logistic regression, random forest, XGBoost, decision tree, support-vector machine (radial and linear kernels), neural network, and k-nearest neighbors. The development sample and number of events were acknowledged as constraints on model complexity; model comparison was therefore exploratory rather than a claim of definitive algorithmic superiority [24, 25].

2.7 Temporal validation, calibration, and model updating

After development, fitted pipelines were frozen and applied to the later prospective cohort without refitting. This unupdated evaluation is the primary temporal-validation analysis. Discrimination was quantified by AUROC with 95% confidence intervals and compared with the standard logistic model using DeLong tests. Sensitivity, specificity, positive and negative predictive values, accuracy, F1 score, and Brier score were reported at prespecified or Youden-based thresholds.

Because the retrospective development cohort was deliberately case enriched, its event fraction did not estimate clinical incidence. Calibration-in-the-large, calibration slope, Brier score, integrated calibration index (ICI), maximum calibration error, and observed-versus-predicted plots were evaluated in the temporal cohort [26]. Intercept-only recalibration was prespecified as the most defensible correction for the prevalence shift.

2.8 Clinical utility, explainability, and prognosis

Decision-curve analysis quantified net benefit across plausible risk thresholds [27]. Risk distributions and observed rates across predicted-risk quintiles were examined. SHapley Additive exPlanations (SHAP) summarized global feature contribution and direction for the selected tree-based model [28]. These analyses describe potential decision support; they do not establish the effectiveness or safety of model-guided prophylaxis.

For the prospective cohort, Kaplan–Meier curves and log-rank tests compared event-free survival by POAF status. Cox regression adjusted for age, sex, LAVI, and surgery type; hazard ratios (HRs) and 95% confidence intervals were reported. Because the number of six-month events was small, prognostic modeling was deliberately parsimonious and interpreted as exploratory.

2.9 Statistical analysis

Continuous variables are presented as mean ± SD or median (interquartile range) and compared using the t test or Mann–Whitney U test, as appropriate. Categorical variables are presented as number

(percentage) and compared using the chi-square or Fisher exact test. Standardized mean differences complement P values. Tests were two sided, with $P < 0.05$ considered statistically significant. Analyses were performed in R version 4.3.0 or later using glmnet, randomForest, xgboost, survival, pROC, and shapviz.

3. Results

3.1 Study population and data quality

After removal of 8 contradictory outcome labels, 28 records duplicated across cohorts, and 3 within-cohort duplicates, a total of 514 patients (including 233 POAF events) were included in the final analysis (Fig. 1). The development cohort included 347 case-enriched patients with 189 POAF events (54.5%); this proportion is not an incidence estimate. The prospective temporal cohort included 167 consecutive patients with 44 events (26.3%).

Core predictors (LAVI, LAD, age, sex, and BMI) had less than 1% missingness. The proportion of randomly missing data for laboratory parameters ranges from 0.2% (for hemoglobin and white blood cell count) to 10.7% (for troponin). High missingness for coronary angiography, graft, and bypass-specific variables largely represented structural non-applicability rather than random data loss (Fig. 2). CPB and cross-clamp times were highly correlated (Spearman rho = 0.961). Data harmonization removed category-level separation caused by spelling variants and reduced the internal-to-temporal performance gap (Supplementary Figure S1).

3.2 Baseline and operative characteristics

In the development cohort (Table 1), patients with POAF had higher LAVI (median, 33.8 vs 23.2 mL/m²; $P < 0.001$; SMD = 0.756), larger LAD (45.0 vs 39.0 mm; $P < 0.001$; SMD = 0.748), more mitral disease (74.6% vs 44.9%; $P < 0.001$), more tricuspid disease (57.7% vs 28.5%; $P < 0.001$), and more affected valves ($P < 0.001$; SMD = 0.781). Furthermore, preoperative BNP/NT-proBNP levels were significantly higher in the POAF group compared with the non-POAF group ($P < 0.001$), indicating that patients with compromised preoperative cardiac functional reserve have an elevated likelihood of developing POAF. In the temporal cohort (Table 1), LAVI (26.0 vs 23.3 mL/m²; $P = 0.007$), LAD ($P = 0.017$), and number of affected valves ($P < 0.001$) again differentiated POAF from non-POAF.

Table 1
Selected baseline and operative characteristics in the development and prospective temporal cohorts.

Characteristic	Development POAF (n = 189)	Development no POAF (n = 158)	P	Temporal POAF (n = 44)	Temporal no POAF (n = 123)	P
Age, years	57.0 (52.0–65.0)	56.0 (50.0–62.0)	0.453	59.5 (54.0–70.0)	56.0 (50.0–62.5)	0.028
BMI, kg/m²	23.7 ± 3.3	23.5 ± 3.2	0.564	22.9 ± 3.2	23.8 ± 3.1	0.111
LAD, mm	45.0 (40.0–51.0)	39.0 (35.2–44.8)	< 0.001	41.5 (38.8–49.0)	39.0 (35.0–45.0)	0.017
LAVI, mL/m²	33.8 (26.6–46.8)	23.2 (20.1–27.7)	< 0.001	26.0 (22.4–30.6)	23.3 (20.8–26.8)	0.007
BNP/NT-proBNP	137.9 (55.0–324.5)	66.0 (25.0–179.5)	< 0.001	169.0 (94.0–498.6)	94.0 (29.0–411.0)	0.013
Mitral valve disease, n (%)	141 (74.6)	71 (44.9)	< 0.001	29 (65.9)	62 (50.4)	0.076
Tricuspid valve disease, n (%)	109 (57.7)	45 (28.5)	< 0.001	21 (47.7)	26 (21.1)	< 0.001
Affected valves, median (IQR)	2.0 (1.0–3.0)	1.0 (1.0–2.0)	< 0.001	2.0 (1.0–3.0)	1.0 (1.0–2.0)	< 0.001
Cardiopulmonary bypass, n (%)	91.2%	63.9%	< 0.001	90.0%	64.0%	0.002

Note. Values are median (IQR), mean ± SD, or n (%). BMI, body mass index; BNP, B-type natriuretic peptide; LAD, left atrial diameter; LAVI, left atrial volume index; POAF, postoperative atrial fibrillation.

Comparison of the distribution characteristics of key continuous variables (Fig. 3) confirmed that LAVI and LAD presented the most significant intergroup differences. Use of CPB was more frequent among patients with POAF in both cohorts, whereas CPB duration and aortic cross-clamp duration substantially overlapped. The higher incidence of left atrial appendage ligation observed in the POAF group does not support the conclusion that "left atrial appendage ligation causes atrial fibrillation." Instead, it suggests that this procedure was performed based on clinicians' pre-procedural assessment that the patient was at high risk of atrial fibrillation. This observation implies that the morphology and function of the left atrial appendage are associated with the risk of atrial fibrillation-related embolism. These unadjusted contrasts are not evidence that operative approach, left atrial appendage management, or medication exposure caused or prevented POAF; indication, procedure type, and structural non-applicability are plausible sources of confounding.

3.3 Independent predictors

In the preoperative multivariable model (Table 2), LAVI was the strongest independent predictor (aOR per SD, 2.18; 95% CI, 1.34–3.52; P = 0.002), followed by tricuspid disease (aOR, 2.38; 95% CI, 1.29–4.37; P = 0.005), BMI (aOR per SD, 1.39; 95% CI, 1.04–1.87; P = 0.025), and age (aOR per SD, 1.38; 95% CI, 1.05–1.83; P = 0.023). LAD was not independently associated after LAVI was included (aOR per SD, 1.16; P = 0.452).

Table 2
Multivariable logistic regression for POAF in the development cohort.

Predictor	Preoperative aOR (95% CI)	P	Perioperative aOR (95% CI)	P
Age (per SD)	1.38 (1.05–1.83)	0.023	1.23 (0.90–1.68)	0.187
BMI (per SD)	1.39 (1.04–1.87)	0.025	1.46 (1.06–2.02)	0.022
LAVI (per SD)	2.18 (1.34–3.52)	0.002	3.19 (1.66–6.12)	< 0.001
LAD (per SD)	1.16 (0.79–1.72)	0.452	1.08 (0.70–1.65)	0.725
Mitral valve disease	1.80 (0.95–3.42)	0.073	1.76 (0.86–3.61)	0.123
Tricuspid valve disease	2.38 (1.29–4.37)	0.005	2.30 (1.21–4.39)	0.012
CPB time (per SD)	Not included	—	0.89 (0.59–1.34)	0.576
LAA ligation/clipping	Not included	—	1.06 (0.56–2.00)	0.862
Intraoperative blood loss (per SD)	Not included	—	1.21 (0.82–1.79)	0.331

Note. Continuous variables are expressed per one standard deviation. Estimates were pooled across 10 multiply imputed datasets. aOR, adjusted odds ratio; CPB, cardiopulmonary bypass; LAA, left atrial appendage.

In the perioperative model (Table 2), LAVI remained dominant (aOR per SD, 3.19; 95% CI, 1.66–6.12; P < 0.001); tricuspid disease (aOR, 2.30; 95% CI, 1.21–4.39; P = 0.012) and BMI (aOR per SD, 1.46; 95% CI, 1.06–2.02; P = 0.022) also remained independently associated. CPB time, left atrial appendage ligation/clipping, and intraoperative blood loss were not independently associated. This absence of statistical independence does not prove that modifying bypass duration would have no effect.

When POAF was restricted to the first 72 postoperative hours (Supplementary Table S1), LAVI (aOR per SD, 1.95; 95% CI, 1.14–3.36; P = 0.016) and tricuspid disease (aOR, 2.28; 95% CI, 1.02–5.09; P = 0.045) remained significant. Preoperative potassium chloride was also associated in this sensitivity analysis (aOR, 2.51; P = 0.021), but the direction is compatible with confounding by indication and should not be interpreted causally.

3.4 Internal and temporal model performance

Five-fold cross-validation was adopted to evaluate the performance of 10 machine learning algorithms, and their predictive performance was compared with that of the conventional multivariate logistic regression model in both the development cohort and the prospective temporal validation cohort (Fig. 4). For the preoperative predictor set (Table 3), random forest had the highest internally cross-validated AUROC (0.817; 95% CI, 0.772–0.862), followed by radial SVM (0.797) and XGBoost (0.795). For the perioperative set (Table 3), random forest again ranked highest (AUROC, 0.827; 95% CI, 0.783–0.870), followed by XGBoost (0.813) and radial SVM (0.811). Rankings changed during temporal validation (Table 3). The frozen preoperative ridge model performed best (AUROC, 0.723; 95% CI, 0.639–0.808), followed by linear SVM (0.721) and standard logistic regression (0.717). The random forest and XGBoost AUROCs fell to 0.684 and 0.692, respectively. For perioperative models, elastic-net logistic regression performed best (AUROC, 0.709; 95% CI, 0.623–0.796), closely followed by random forest (0.705) and LASSO (0.704).

Table 3
Discrimination of candidate models during repeated cross-validation and prospective temporal validation.

Model	Preoperative internal AUROC	Preoperative temporal AUROC	Perioperative internal AUROC	Perioperative temporal AUROC
Standard logistic	0.753 (0.701–0.805)	0.717 (0.628–0.805)	0.758 (0.708–0.809)	0.625 (0.531–0.719)
LASSO logistic	0.770 (0.720–0.821)	0.706 (0.619–0.794)	0.773 (0.723–0.824)	0.704 (0.614–0.793)
Ridge logistic	0.769 (0.719–0.819)	0.723 (0.639–0.808)	0.771 (0.721–0.822)	0.693 (0.603–0.782)
Elastic-net logistic	0.772 (0.722–0.822)	0.712 (0.628–0.796)	0.772 (0.721–0.823)	0.709 (0.623–0.796)
Random forest	0.817 (0.772–0.862)	0.684 (0.591–0.776)	0.827 (0.783–0.870)	0.705 (0.616–0.793)
XGBoost	0.795 (0.748–0.843)	0.692 (0.598–0.786)	0.813 (0.768–0.858)	0.680 (0.586–0.773)
Decision tree	0.729 (0.673–0.785)	0.603 (0.521–0.685)	0.748 (0.693–0.803)	0.603 (0.521–0.685)
SVM, radial	0.797 (0.749–0.844)	0.675 (0.579–0.771)	0.811 (0.766–0.856)	0.668 (0.571–0.765)
SVM, linear	0.766 (0.715–0.816)	0.721 (0.636–0.807)	0.764 (0.713–0.815)	0.634 (0.540–0.728)
Neural network	0.746 (0.694–0.798)	0.621 (0.516–0.727)	0.759 (0.708–0.810)	0.643 (0.545–0.741)
K-nearest neighbors	0.778 (0.729–0.828)	0.655 (0.561–0.750)	0.776 (0.726–0.826)	0.681 (0.593–0.769)

Note. Internal estimates are from out-of-fold predictions in repeated stratified five-fold cross-validation. Temporal estimates are from the frozen models applied to 167 later patients with 44 events.

Although the ranking order of all models within the prospective temporal validation cohort changed significantly, removing category-label leakage substantially reduced the implausibly high internal AUROC estimates and narrowed the internal–external performance gap across all models, with the gap decreasing from 0.108–0.285 (with leakage) to 0.036–0.133 (after leakage removal), particularly for complex algorithms such as random forest (0.254→0.133) and LASSO logistic regression (0.285→0.064) (Fig. 5).

DeLong comparisons showed no statistically significant superiority of complex nonlinear models over the corresponding standard logistic benchmark (Supplementary Table S2). This finding suggests that complex models are more susceptible to overfitting when the sample size is limited, while simple linear models exhibit superior generalization performance.

After intercept recalibration, the preoperative ridge model at its selected threshold had 65.9% sensitivity, 67.5% specificity, 42.0% positive predictive value, 84.7% negative predictive value, 67.1% accuracy, and an F1 score of 0.513 in the temporal cohort (Supplementary Figure S2).

3.5 Calibration, clinical utility, and explainability

The unupdated preoperative ridge model preserved a calibration slope near 1.0 but systematically overpredicted absolute risk (Fig. 6): calibration intercept, − 0.913; mean predicted risk, 0.442; observed incidence, 0.263; ICI, 0.173; and Brier score, 0.2047. Intercept-only recalibration (Table 4) improved ICI to 0.048 and Brier score to 0.1754 without changing AUROC. Exploratory Platt scaling fitted on the temporal cohort produced an apparent ICI of 0.009 and Brier score of 0.1732, whereas isotonic regression compressed predictions and reduced discrimination. The Platt result is model-updating performance, not independent validation.

Table 4
Calibration of the preoperative ridge model in the prospective temporal cohort.

Calibration approach	Slope	Intercept	ICI	Emax	Brier	Mean predicted risk	AUROC
Unupdated (primary temporal validation)	0.999	−0.913	0.173	0.230	0.2047	0.442	0.723
Intercept-only recalibration	0.999	0.294	0.048	0.115	0.1754	0.216	0.723
Platt scaling (exploratory/apparent)	0.973	0.003	0.009	0.120	0.1732	0.263	0.718
Isotonic regression (exploratory/apparent)	0.167	−0.017	0.048	0.253	0.1786	0.266	0.687

Note. Observed POAF incidence was 0.263. Platt and isotonic methods were fitted and evaluated in the temporal cohort; their values are apparent model-updating performance and are not an independent validation result.

Decision curves indicated higher net benefit than treat-all or treat-none strategies at thresholds of approximately 0.10–0.35, with the clearest separation near 0.10–0.25 (Fig. 7). In the temporal cohort, observed POAF frequency rose monotonically across predicted-risk quintiles, from approximately 8% in the lowest quintile to 45% in the highest; When dichotomizing based on the median risk score, the incidence of POAF was three times higher in high-risk patients (39.3% vs. 13.3%, $P < 0.001$) (Supplementary Figure S3).

SHAP analysis ranked LAVI first, followed by LAD, tricuspid involvement, BMI, CRP, number of affected valves, and LVEF (Fig. 8). SHAP values explain model behavior but do not establish etiologic effects. The coefficient estimation results of ridge regression are consistent with the variable importance ranking obtained by SHAP analysis, indicating that the variable importance of the model is not affected by the

selection of model type, which confirms that the conclusions of this study are robust (Supplementary Table 3).

3.6 In-hospital and six-month outcomes

In the retrospective cohort, POAF was associated with longer ICU and postoperative stays, but these differences were not reproduced consistently in the temporal cohort. The high-versus-low predicted-risk comparison was descriptive: median ICU stay was 68.25 hours in the high-risk group and 84 hours in the low-risk group, with no formal P value. Accordingly, the data do not support a claim that the high-risk group had a longer ICU stay (Supplementary Table 4). The median time to the first POAF episode was 4 days (IQR: 2–7 days), with 40% of episodes occurring within 72 hours postoperatively. The distribution was right-skewed, with the highest incidence observed between postoperative days 2 and 5, after which it gradually declined. In the development cohort, 50.8% of POAF episodes required pharmacological or electrical cardioversion therapy, and 16.9% of patients experienced recurrence (Supplementary Figure S4).

In the prospective cohort study, all patients completed a 6-month follow-up period; the detection rate of atrial fibrillation/atrial flutter in the POAF group showed a continuous increase spanning 1months, 3 months and 6 months. Patients in the POAF group exhibited a significantly elevated risk of atrial fibrillation recurrence at 3 months (11.6% vs. 0.9%, OR 15.13, P = 0.014) and at 6 months (16.3% vs. 0.9%, OR 21.58, P = 0.005) (P < 0.001) (Supplementary Figure S4). In the prospective cohort (Table 5), AF recurrence by six months occurred in 7 of 43 patients with in-hospital POAF (16.3%) and 1 of 112 without POAF (0.9%; OR, 21.58; 95% CI, 2.57–181.39; P = 0.005). The six-month composite endpoint occurred in 9 of 44 (20.5%) and 6 of 123 (4.9%), respectively (OR, 5.01; 95% CI, 1.67–15.06; P = 0.004). In Cox analysis adjusted for age, sex, LAVI, and surgery type, POAF was associated with a higher hazard of the composite endpoint (adjusted HR, 3.22; 95% CI, 1.00–10.34; P = 0.050), while LAVI remained independently associated (HR per mL/m², 1.16; 95% CI, 1.03–1.30; P = 0.012) (Table 6). Wide confidence intervals and the borderline P value indicate imprecision. Kaplan–Meier survival analysis demonstrated that patients in the POAF group had a significantly worse event-free survival rate (P = 0.002; unadjusted HR = 4.90; 95% CI: 1.64–14.61) (Fig. 9).

Table 5
Six-month outcomes in the prospective temporal cohort.

Outcome	POAF, n/N (%)	No POAF, n/N (%)	Odds ratio (95% CI)	P
AF recurrence by 3 months	5/43 (11.6)	1/116 (0.9)	15.13 (1.71–133.61)	0.014
AF recurrence by 6 months	7/43 (16.3)	1/112 (0.9)	21.58 (2.57–181.39)	0.005
AF/atrial flutter on 6-month ECG	7/35 (20.0)	1/82 (1.2)	20.25 (2.39–171.93)	0.006
Cardiovascular readmission	1/43 (2.3)	4/118 (3.4)	0.68 (0.07–6.25)	0.732
30-day readmission	1/43 (2.3)	6/118 (5.1)	0.44 (0.05–3.80)	0.459
Six-month composite endpoint	9/44 (20.5)	6/123 (4.9)	5.01 (1.67–15.06)	0.004

Table 6
Cox regression for the six-month composite endpoint.

Variable	Model	Hazard ratio (95% CI)	P
POAF	Univariable	4.90 (1.64–14.61)	0.004
POAF	Adjusted	3.22 (1.00–10.34)	0.050
Age, per year	Adjusted	1.02 (0.97–1.08)	0.383
Male sex	Adjusted	1.16 (0.38–3.60)	0.792
LAVI, per mL/m²	Adjusted	1.16 (1.03–1.30)	0.012
Valve surgery vs combined	Adjusted	0.57 (0.12–2.86)	0.499
CABG vs combined	Adjusted	0.41 (0.03–5.21)	0.495

Note. The adjusted model included age, sex, LAVI, and surgery type. Estimates are imprecise because the number of events was small.

4. Discussion

4.1 Principal findings

This study yielded four principal findings. First, LAVI was the most stable independent predictor across preoperative, perioperative, and 72-hour sensitivity analyses. Second, the internally best nonlinear algorithm did not remain best in a later cohort: ridge regression achieved the strongest preoperative temporal discrimination. Third, prevalence shift from a case-enriched development sample to a consecutive prospective cohort primarily impaired calibration-in-the-large, making recalibration essential. Fourth, POAF marked persistent atrial vulnerability, with substantially higher six-month AF recurrence

and adverse-event risk.Together, these findings support the prespecified substrate hypothesis and the anticipated advantage of parsimony under temporal shift.

4.2 Atrial substrate and biological plausibility

LAVI integrates cumulative atrial exposure to pressure and volume load more comprehensively than a single linear diameter. Enlargement and dysfunction promote heterogeneous conduction and provide a substrate on which postoperative inflammation and autonomic or metabolic stress can trigger AF. Recent work showing that left atrial strain adds information after CABG is directionally consistent with this substrate interpretation [29]. Prospective risk-factor data also emphasize the combined contributions of atrial dimensions, procedure type, and systemic vulnerability [30]. The normal reference range of LAVI for adults is generally defined as less than 34 mL/m². In this study, the median LAVI of the POAF group was 33.8, which is close to the upper limit of the normal reference range, indicating that preoperative atrial structural remodeling is the core pathological basis for POAF onset. Traditional POAF risk assessment systems, including the POAF score, CHA_2DS_2-VASc score, and Cleveland Clinic score, all include common risk factors such as age, hypertension and valvular heart disease, but do not incorporate the quantitative assessment indicator of left atrial volume. The findings of this study support the implementation of routine preoperative LAVI measurement in clinical practice.

Tricuspid disease may be an under-recognized marker of advanced biatrial remodeling, right-sided pressure/volume loading, or pulmonary vascular disease. Its independent association after adjustment warrants replication; it should not be interpreted as proof of a direct causal pathway. Likewise, CRP and other inflammatory features plausibly reflect susceptibility to the postoperative inflammatory surge. Cardiac surgery can trigger systemic inflammatory responses via multiple pathological processes including tissue injury, extracorporeal circulation, ischemia-reperfusion injury and oxidative stress. This inflammatory response can subsequently modify atrial conduction function, induce abnormal calcium ion metabolism, and ultimately increase the incidence of atrial ectopic beats. Meta-analytic evidence links perioperative inflammatory markers with POAF [31], and inflammatory composite indices have shown predictive associations after isolated CABG [32]. These finding further confirms the core position of the inflammatory theory in the pathogenesis of POAF.

Electrolyte signals require particular caution. Population studies associate lower potassium and magnesium with incident AF [33, 34], yet time-matched cardiac-surgery analyses have not shown routine potassium or magnesium supplementation to be protective and have raised the possibility of confounding [35]. In our data, medication associations varied across analyses and likely reflect indication and calendar-time effects. The appropriate conclusion is to correct documented electrolyte abnormalities according to clinical care pathways, not to infer benefit from indiscriminate supplementation.

4.3 Why simple models transported better

Random forest captured nonlinearities and interactions that improved out-of-fold discrimination within the retrospective sample, but these gains were not stable across time. Changes in case mix, surgical

access, perioperative practice, and the development sampling fraction can alter both predictor distributions and predictor–outcome relationships. In a moderate sample, flexible algorithms also have more opportunity to learn unstable patterns. The temporal results therefore reinforce a general lesson: internal ranking is not evidence of clinical transportability, and algorithmic complexity is not itself innovation. This result is consistent with the conclusion of the systematic review completed by Christodoulou et al.: In the field of clinical prediction modeling, machine learning models rarely outperform traditional logistic regression models by a clinically significant margin. In the majority of comparative studies included in the review, the difference in AUROC between the two types of models was less than 0.05 [36].

The more distinctive contribution of this analysis is the explicit separation of three concepts: discrimination of the frozen model, calibration correction for a known sampling-fraction shift, and post-hoc model updating. Reporting Platt scaling as if it were independent validation would overstate performance because the same temporal cohort was used to estimate and evaluate the recalibration mapping. A future study should estimate any recalibration parameters in a dedicated updating sample and evaluate the updated model in another untouched cohort.

4.4 Prognostic implications

The marked recurrence gradient supports treating POAF as a marker of atrial vulnerability rather than an event that necessarily ends at discharge. Clinicians ought to discard the view that "POAF is merely a transient complication that only requires intra-hospital acute management before discharge", and should instead recognize it as a biomarker indicating long-term cardiovascular risk. The American Heart Association similarly frames AF detected during acute hospitalization as carrying a meaningful risk of recurrence and requiring structured follow-up [37]. A randomized trial showed that continuous patch monitoring after cardiac surgery detects otherwise unrecognized AF after discharge, achieving early risk stratification and hourly prediction for POAF[38]. Wearable-device algorithms and smartphone photoplethysmography may further extend surveillance, although their clinical pathways and cost-effectiveness remain unsettled [39, 40].

Risk estimates from this study should be used to support clinical attention, not to mandate therapy. Beta-blocker or amiodarone prophylaxis, rhythm monitoring, and anticoagulation should follow individualized guideline-based assessment of contraindications, bleeding risk, AF burden, and thromboembolic risk. Moreover, the model-generated threshold of 30% is not a validated treatment threshold, and left atrial appendage closure cannot be recommended from these observational data.

4.5 Strengths and limitations

Strengths include use of a later prospective cohort, harmonization of category labels, explicit leakage auditing, fold-contained preprocessing, comparison of 11 algorithms, evaluation beyond AUROC, and integration of prediction with six-month prognosis. Data cleaning reduced the discrepancy between internal and temporal performance, highlighting the importance of preprocessing procedures in clinical machine-learning studies.

Several limitations materially constrain interpretation. First, development and temporal validation were performed at one center; the later cohort is independent in time but not an external geographic population. Second, the development cohort was case enriched, so its outcome proportion is not incidence and its absolute probabilities are not transportable without adjustment. Third, the sample size and event count limit reliable estimation of high-dimensional interactions and contribute to uncertainty in subgroup, DeLong, and prognosis analyses. Fourth, missingness was partly structural and monitoring intensity may have differed over time. Fifth, medication and operative associations are observational and susceptible to confounding by indication, procedure type, and secular practice changes. Sixth, Platt and isotonic recalibration evaluated in the same temporal data are exploratory/apparent. Seventh, follow-up was six months, event counts were small, and rhythm ascertainment was not based on continuous monitoring for every patient. Finally, all model coefficients, preprocessing pipelines, and analysis scripts are publicly accessible to support methodological reproducibility.

4.6 Future research

The next step is a prospectively registered, multicenter geographic validation using a frozen preoperative ridge model and a prespecified calibration strategy. Site-specific calibration should be evaluated separately from discrimination. A subsequent impact trial could test whether model-informed monitoring or prophylaxis (e.g. perioperative anti-inflammatory therapy, intensive β-blocker regimens, or early rhythm control) improves patient outcomes without increasing treatment-related harm. Multimodal extensions could incorporate atrial strain, continuous ECG/photoplethysmography, inflammatory profiles, AF-related polymorphic genes and cardiac imaging features; radiomics may offer additional quantitative substrate information, but it requires standardized acquisition and careful external validation [41].

5. Conclusions

In adults undergoing valve surgery, CABG, or combined procedures, LAVI was the most stable predictor of POAF and supported a substrate-dominant hypothesis. Complex models achieved higher internal discrimination, whereas a preoperative ridge model was more robust in prospective temporal validation. The model requires recalibration and independent multicenter evaluation before deployment. POAF was associated with substantially greater six-month AF recurrence and adverse-event risk, supporting structured post-discharge surveillance. Final optimization of diagnostic and therapeutic guidance, resource allocation, and preventive strategies for patients with POAF following valve or CABG surgery.

Declarations

Author Agreement

Ethics approval and consent to participate

The Medical Ethical Review Board of Sichuan Provincial People's Hospital approved this study (Approval No. 2026523).

Data Availability Statement

The data that support the findings of this study are available on request from the first author or corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Author Contributions

YY, Xia J and QJ contributed equally to the manuscript. YY and QJ designed and performed the research. YY formulated the original draft. YY and Xia J acquired the data and analyzed the data. CY conducted the verification and contributed new analytical tools. QJ contributed to the revision of the manuscript. All authors reviewed the manuscript critically and commented on manuscript drafts.

Acknowledgment

Not applicable.

Funding

This research was funded by Huanhua Talent for Discipline Backbone of Sichuan Provincial People's Hospital, grant number SY2022017, Science Fund for Distinguished Young Scholars of Sichuan Province, grant number 2021JDJQ0041, National Natural Science and Technology Foundation of China, grant number 81800274.

Conflict of Interest

The authors declare no conflict of interest.

Declaration of AI and AI-assisted technologies

During the preparation of this work the authors used R version 4.3.0 or later using glmnet in order to complete machine learning. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article, including analysis of statistical data using AI tools.

References

1. Suero OR, Ali AK, Barron LR, Segar MW, Moon MR, Chatterjee S. Postoperative atrial fibrillation (POAF) after cardiac surgery: clinical practice review. J Thorac Dis. 2024;16(2):1503–20. 10.21037/jtd-23-1626.
2. Caldonazo T, Kirov H, Rahouma M, et al. Atrial fibrillation after cardiac surgery: a systematic review and meta-analysis. J Thorac Cardiovasc Surg. 2023;165(1):94–e10324. 10.1016/j.jtcvs.2021.03.077.
3. Lin MH, Kamel H, Singer DE, Wu YL, Lee M, Ovbiagele B. Perioperative/postoperative atrial fibrillation and risk of subsequent stroke and/or mortality. Stroke. 2019;50(6):1364–71.

10.1161/STROKEAHA.118.023921.

4. Taha A, Martinsson A, Nielsen SJ, et al. New-onset atrial fibrillation after coronary surgery and stroke risk: a nationwide cohort study. Heart. 2024;111(1):18–26. 10.1136/heartjnl-2024-324573. Published 2024 Dec 13.
5. Ahlsson A, Fengsrud E, Bodin L, Englund A. Postoperative atrial fibrillation in patients undergoing aortocoronary bypass surgery carries an eightfold risk of future atrial fibrillation and a doubled cardiovascular mortality. Eur J Cardiothorac Surg. 2010;37(6):1353–9. 10.1016/j.ejcts.2009.12.033.
6. William J, Rowe K, Hogarty J, et al. Predictors of late atrial fibrillation recurrence after cardiac surgery. JACC Clin Electrophysiol. 2024;10(7 Pt 2):1711–9. 10.1016/j.jacep.2024.05.030.
7. Nattel S, Burstein B, Dobrev D. Atrial remodeling and atrial fibrillation: mechanisms and implications. Circ Arrhythm Electrophysiol. 2008;1(1):62–73. 10.1161/CIRCEP.107.754564.
8. Haysley J, Soliman-Aboumarie H, Huang J, Kalra DK. Perioperative atrial fibrillation. BJA Educ. 2025;25(3):99–106. 10.1016/j.bjae.2024.10.006.
9. Liblik K, Zucker J, Baranchuk A, Fernandez AL, Zhang S, Diasty ME. The role of pericardial fluid biomarkers in predicting post-operative atrial fibrillation: a comprehensive review of current literature. Trends Cardiovasc Med. 2024;34(4):244–7. 10.1016/j.tcm.2023.02.009.
10. Keefe JA, Aguilar-Sanchez Y, Navarro-Garcia JA, et al. Macrophage-mediated IL-6 signaling drives ryanodine receptor-2 calcium leak in postoperative atrial fibrillation. J Clin Invest. 2025;135(9):e187711. 10.1172/JCI187711.
11. Zhang Y, Chen HX, Hou HT, et al. Atlas of lysine propionylation of human right atrium and ALDH6A1-NADH pathway in new-onset atrial fibrillation after coronary surgery. Commun Biol. 2025;8:858. 10.1038/s42003-025-08264-9.
12. Tao H, Shen X, Zou L, Zhang C, Hong L. Left atrial volume index and interleukin-6 as predictors for postoperative atrial fibrillation. J Cardiothorac Surg. 2024;19:325. 10.1186/s13019-024-02813-9.
13. Mariscalco G, Biancari F, Zanobini M, et al. Bedside tool for predicting the risk of postoperative atrial fibrillation after cardiac surgery: the POAF score. J Am Heart Assoc. 2014;3(2):e000752. 10.1161/JAHA.113.000752.
14. Fleet H, Pilcher D, Bellomo R, et al. Predicting atrial fibrillation after cardiac surgery: a scoping review and systematic review of existing prediction models. Perfusion. 2023;38(1):92–108. 10.1177/02676591211037025.
15. Karri R, Mutluer FO, Amir-Khalili A, et al. Machine learning outperforms clinical scoring tools for prediction of postoperative atrial fibrillation in patients admitted to a cardiac surgical intensive care unit. Heart Lung Circ. 2021;30(12):1929–37. 10.1016/j.hlc.2021.05.101.
16. Andrei AC, Cox JL, Shah S, et al. Machine learning-based prediction of new onset of atrial fibrillation after mitral valve surgery. Int J Arrhythm. 2024;25:18. 10.1186/s42444-024-00127-4.
17. Zheng Y, Zhou M, Lin Y, Zhang G. Development and internal validation of an OPCABG-specific prediction model for postoperative atrial fibrillation in Chinese patients: a retrospective cohort study. BMC Cardiovasc Disord. 2025;25:316. 10.1186/s12872-025-04780-y.

18. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. 10.1016/j.jclinepi.2019.02.004.
19. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. 10.1136/bmj-2023-078378.
20. Moons KGM, Damen JAA, Kaul T, et al. PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. 10.1136/bmj-2024-082505.
21. Frendl G, Sodickson AC, Chung MK, et al. 2014 AATS guidelines for the prevention and management of perioperative atrial fibrillation and flutter for thoracic surgical procedures. J Thorac Cardiovasc Surg. 2014;148(3):e153–93. 10.1016/j.jtcvs.2014.06.036.
22. Van Gelder IC, Rienstra M, Bunting KV, et al. 2024 ESC Guidelines for the management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). Eur Heart J. 2024;45(36):3314–414. 10.1093/eurheartj/ehae176.
23. Chinese Society of Pacing and Electrophysiology; Chinese Medical Doctor Association Arrhythmology Committee. Chinese guideline for the management of atrial fibrillation (2025). Chin J Card Pacing Electrophysiol. 2025;39(4):273–343. 10.13333/j.cnki.cjcpe.2025.04.001.
24. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441. 10.1136/bmj.m441.
25. Efthimiou O, Seo M, Chalkou K, et al. Developing clinical prediction models: a step-by-step guide. BMJ. 2024;386:e078276. 10.1136/bmj-2023-078276.
26. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. 10.1186/s12916-019-1466-7.
27. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak. 2006;26(6):565–74. 10.1177/0272989X06295361.
28. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–74.
29. Grasso Granchietti A, Ciardetti N, Mazzoni C, et al. Left atrial strain and risk of atrial fibrillation after coronary artery bypass-grafting. Int J Cardiol. 2025;422:132981. 10.1016/j.ijcard.2025.132981.
30. Bowdish ME, et al. A prospective study of risk factors for postoperative atrial fibrillation after cardiac surgery. J Surg Res. 2024;294:262–8. 10.1016/j.jss.2023.09.060.
31. Weymann A, Popov AF, Sabashnikov A, et al. Baseline and postoperative levels of C-reactive protein and interleukins as inflammatory predictors of atrial fibrillation following cardiac surgery: a systematic review and meta-analysis. Kardiol Pol. 2018;76(2):440–51. 10.5603/KP.a2017.0242.
32. Uğuz B, Topal D, Badem S, Kahraman N, Uğuz İ. Systemic Immune-Inflammation Index: a novel predictor for risk of postoperative atrial fibrillation in patients undergoing isolated coronary artery bypass grafting. Heart Surg Forum. 2022;25(5):E665–73. 10.1532/hsf.4785.

33. Krijthe BP, Heeringa J, Kors JA, et al. Serum potassium levels and the risk of atrial fibrillation: the Rotterdam Study. Int J Cardiol. 2013;168(6):5411–5. 10.1016/j.ijcard.2013.08.048.
34. Khan AM, Lubitz SA, Sullivan LM, et al. Low serum magnesium and the development of atrial fibrillation in the community: the Framingham Heart Study. Circulation. 2013;127(1):33–8. 10.1161/CIRCULATIONAHA.111.082511.
35. Lancaster TS, Schill MR, Greenberg JW, et al. Potassium and magnesium supplementation do not protect against atrial fibrillation after cardiac operation: a time-matched analysis. Ann Thorac Surg. 2016;102(4):1181–8. 10.1016/j.athoracsur.2016.06.066.
36. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. 10.1016/j.jclinepi.2019.02.004.
37. Chyou JY, Barkoudah E, Dukes JW, et al. Atrial fibrillation occurring during acute hospitalization: a scientific statement from the American Heart Association. Circulation. 2023;147(15):e676–98. 10.1161/CIR.0000000000001133.
38. Ha ACT, Verma S, Mazer CD, et al. Effect of continuous electrocardiogram monitoring on detection of undiagnosed atrial fibrillation after hospitalization for cardiac surgery: a randomized clinical trial. JAMA Netw Open. 2021;4(8):e2121867. 10.1001/jamanetworkopen.2021.21867.
39. Hiraoka D, Inui T, Kawakami E, et al. Diagnosis of atrial fibrillation using machine learning with wearable devices after cardiac surgery: algorithm development study. JMIR Form Res. 2022;6(8):e35396. 10.2196/35396.
40. Gruwez H, De Melio N, Vermunicht P, et al. Improving atrial fibrillation or flutter detection and management by smartphone-based photoplethysmography rhythm monitoring following cardiac surgery: a pragmatic randomized trial. Europace. 2025;27(2):euaf015. 10.1093/europace/euaf015.
41. Baeßler B, Engelhardt S, Hekalo A, et al. Perfect match: radiomics and artificial intelligence in cardiac imaging. Circ Cardiovasc Imaging. 2024;17(6):e015490. 10.1161/CIRCIMAGING.123.015490.

Figures

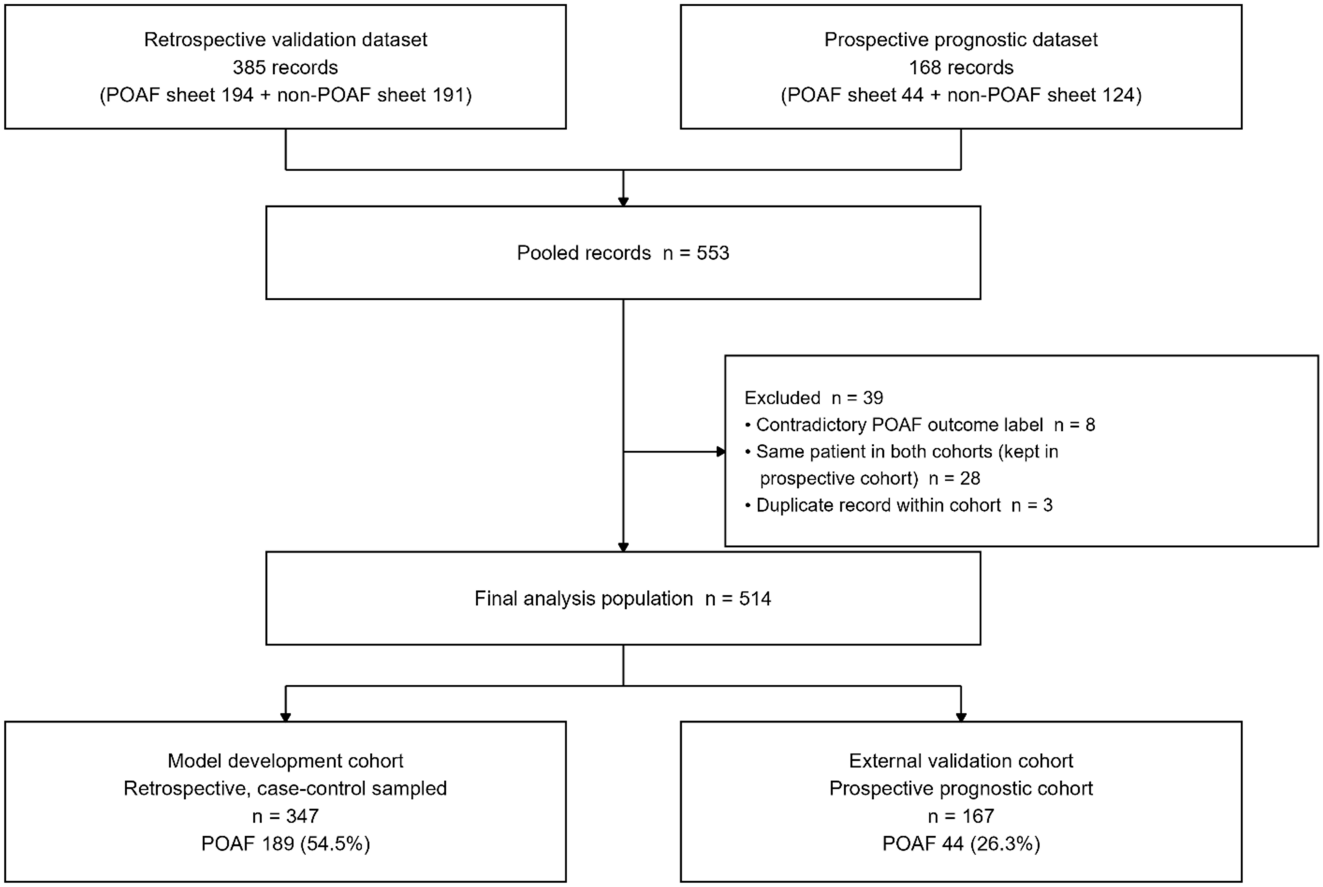


Figure 1

Study flow and analytic cohorts. The final analysis comprised 347 patients in the retrospective case-enriched development cohort and 167 in the prospective temporal validation cohort.

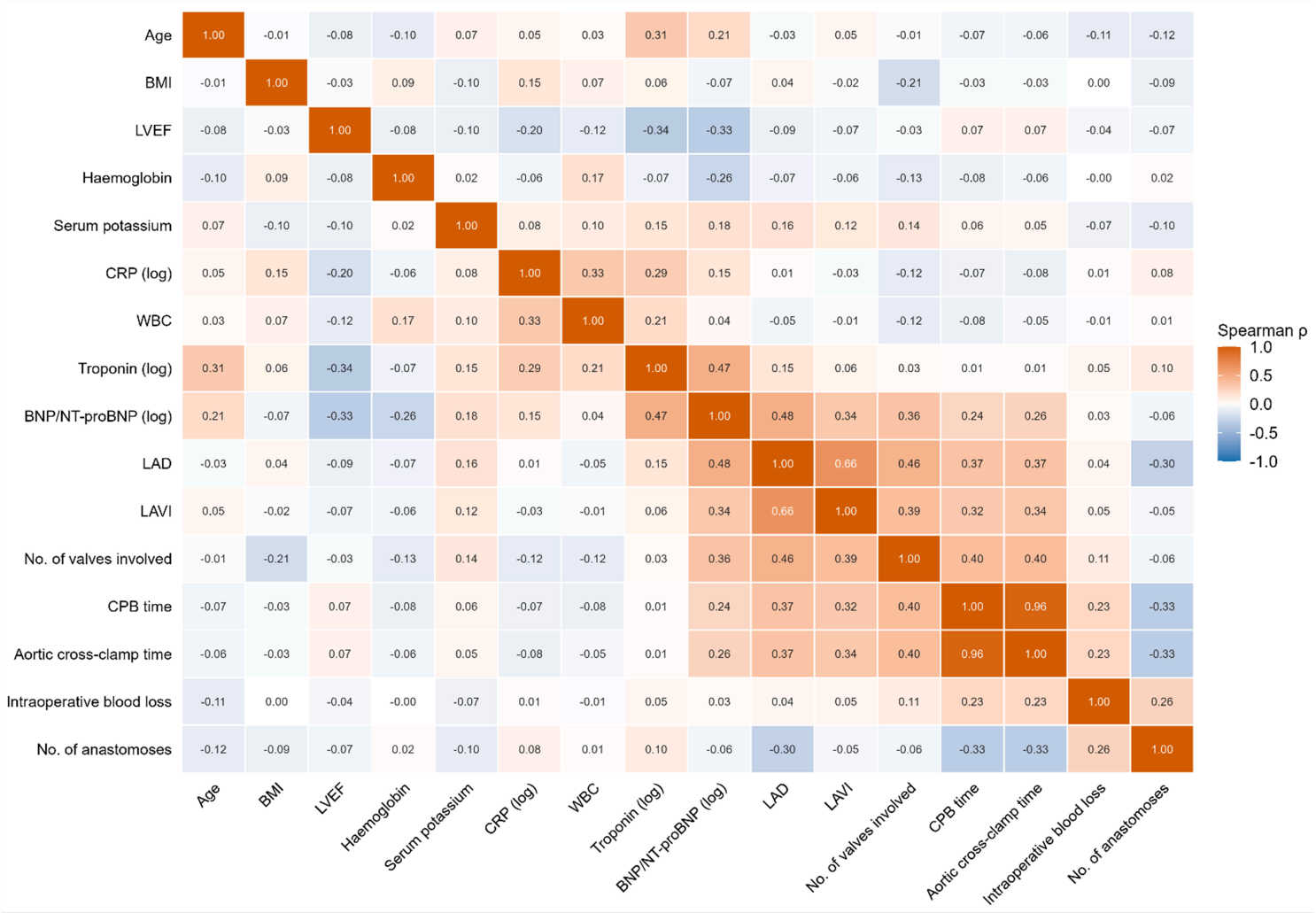


Figure 2

Spearman correlation matrix of continuous candidate predictors in the development cohort. Cardiopulmonary bypass time and aortic cross-clamp time were the most strongly correlated pair (rho = 0.961); entering both into a multivariable model produced variance inflation factors of 14.8 and 14.6. Neither variable was independently associated with POAF under any of four strategies for handling this collinearity, indicating that the association of bypass time with POAF is confounded by type of surgery rather than masked by collinearity.

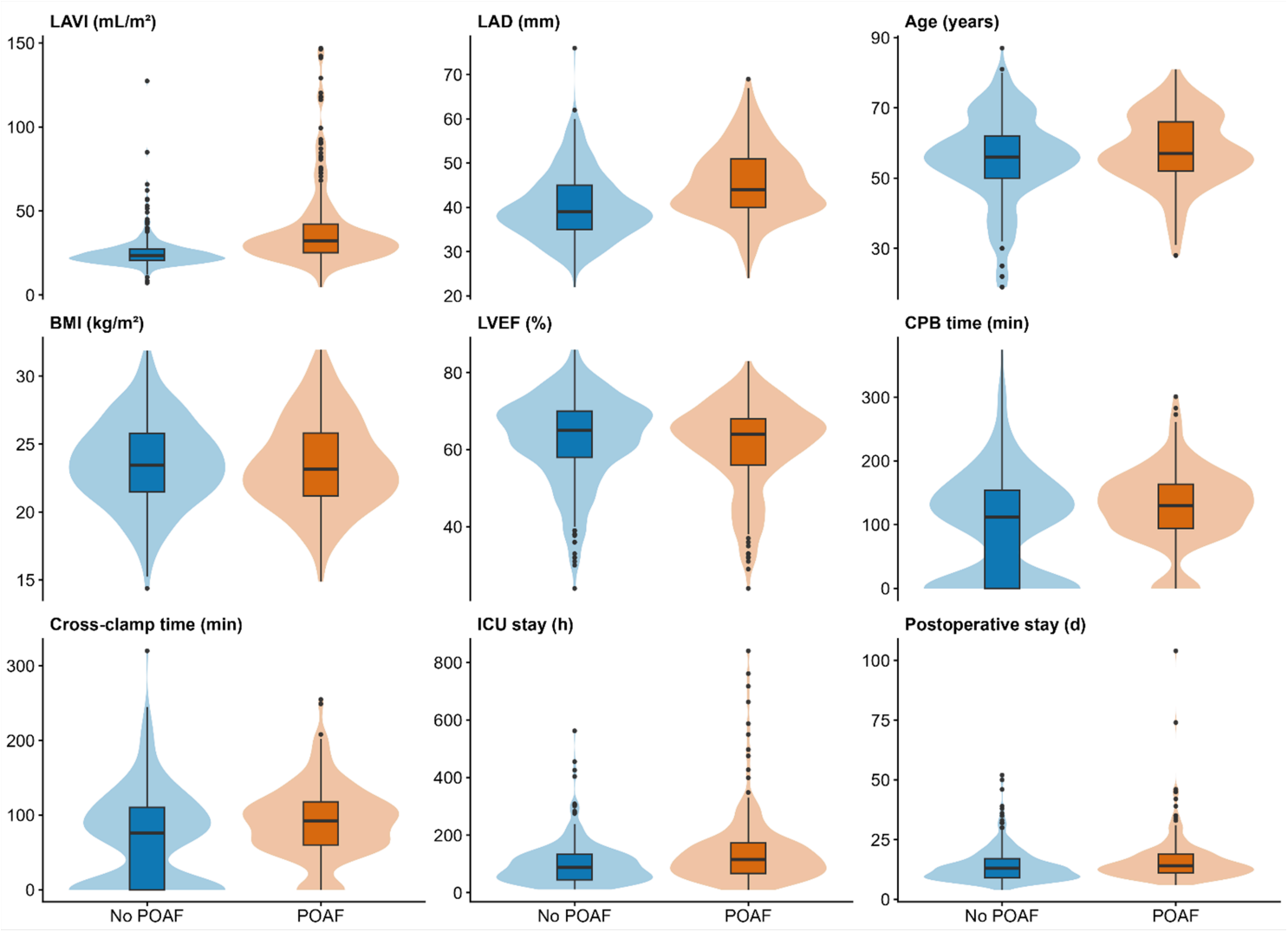


Figure 3

Distribution of key continuous variables by POAF status. Violin plots show density with median and interquartile range. Bypass and cross-clamp times recorded as zero for off-pump procedures represent structural non-applicability and should not be interpreted as observed zero-duration bypass.

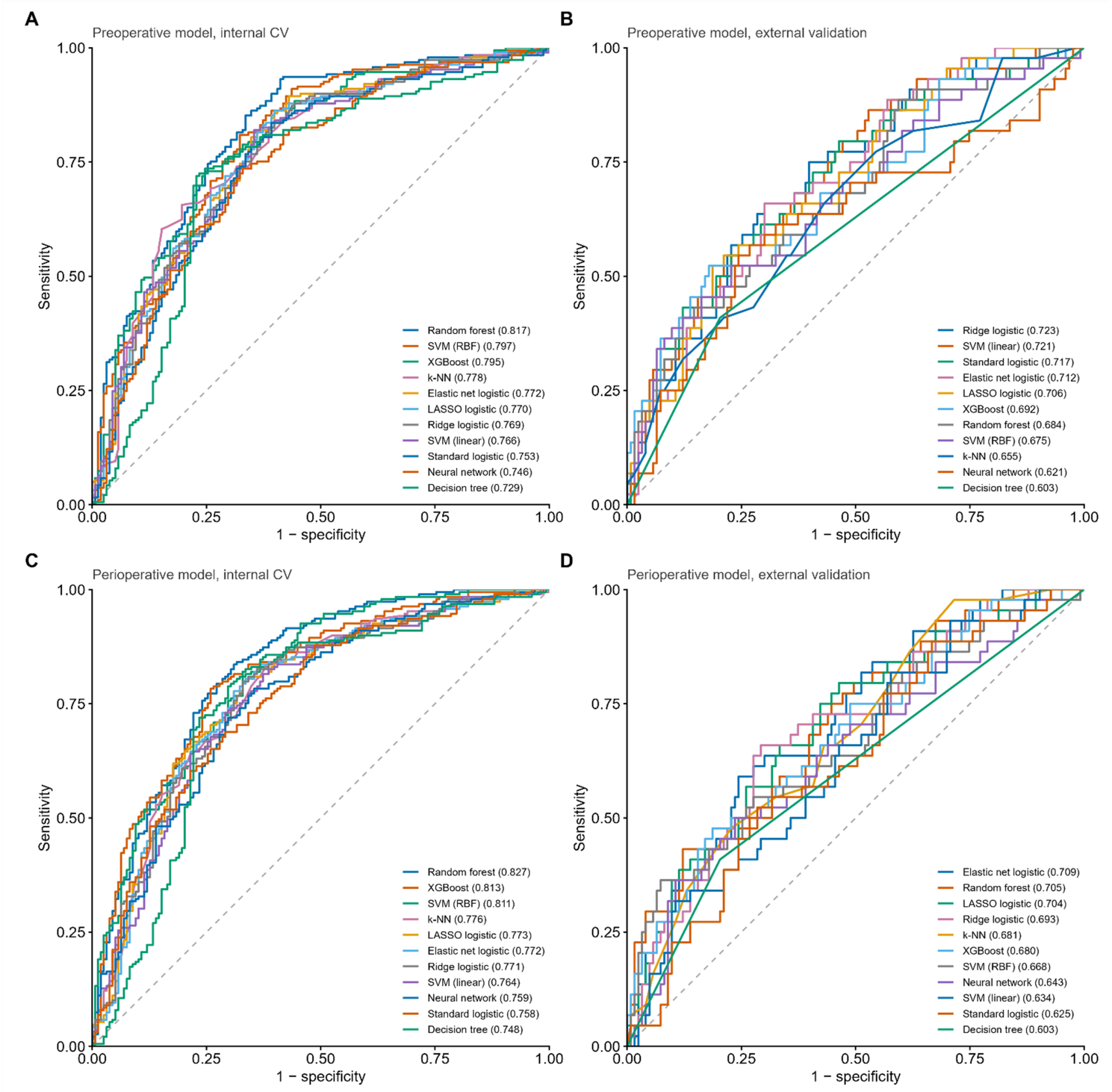


Figure 4

Receiver operating characteristic curves for 11 candidate models. Panels A and C show repeated cross-validated out-of-fold performance in the development cohort; panels B and D show frozen-model performance in the prospective temporal cohort.

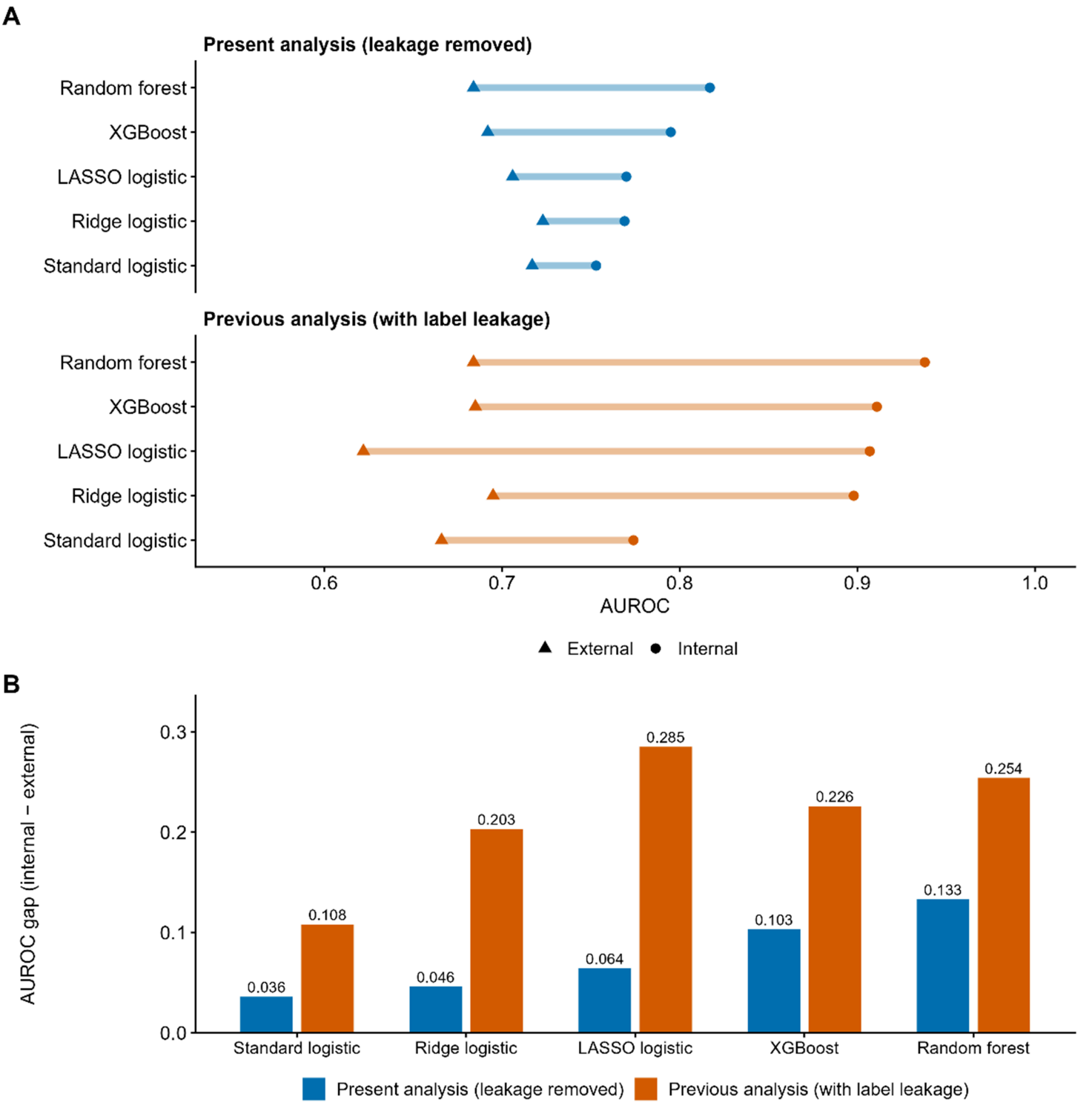


Figure 5

Internal and temporal AUROC before and after removal of category-label leakage. The corrected analysis reduces implausibly high internal performance and narrows the internal–temporal gap.

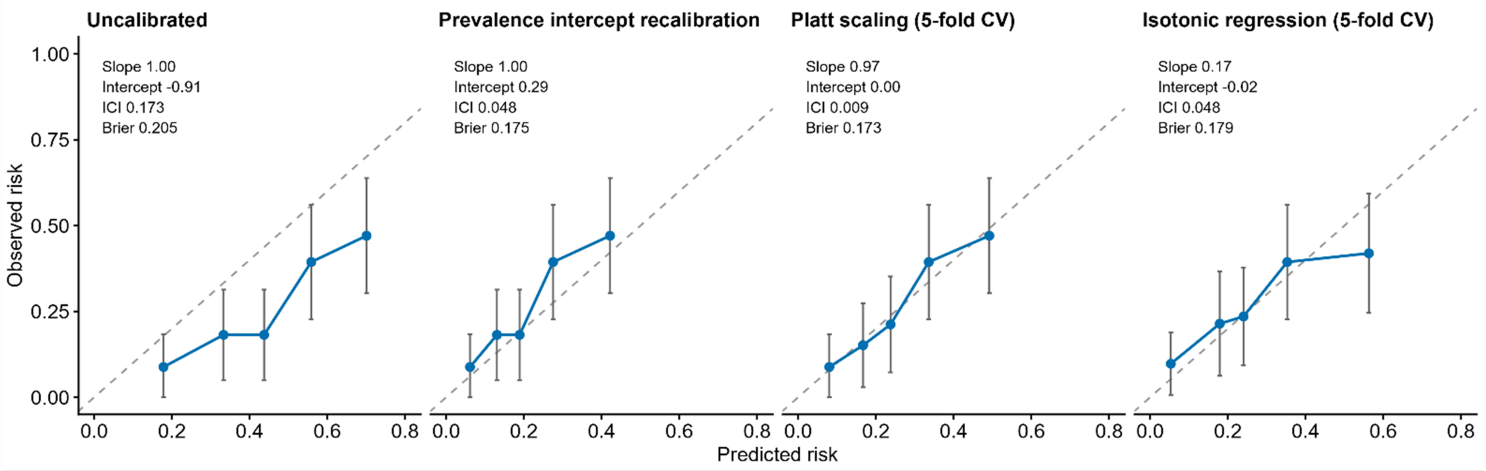


Figure 6

Calibration of the preoperative ridge model in the temporal cohort. Platt and isotonic curves are exploratory apparent model-updating results because their mappings were fitted in the same cohort.

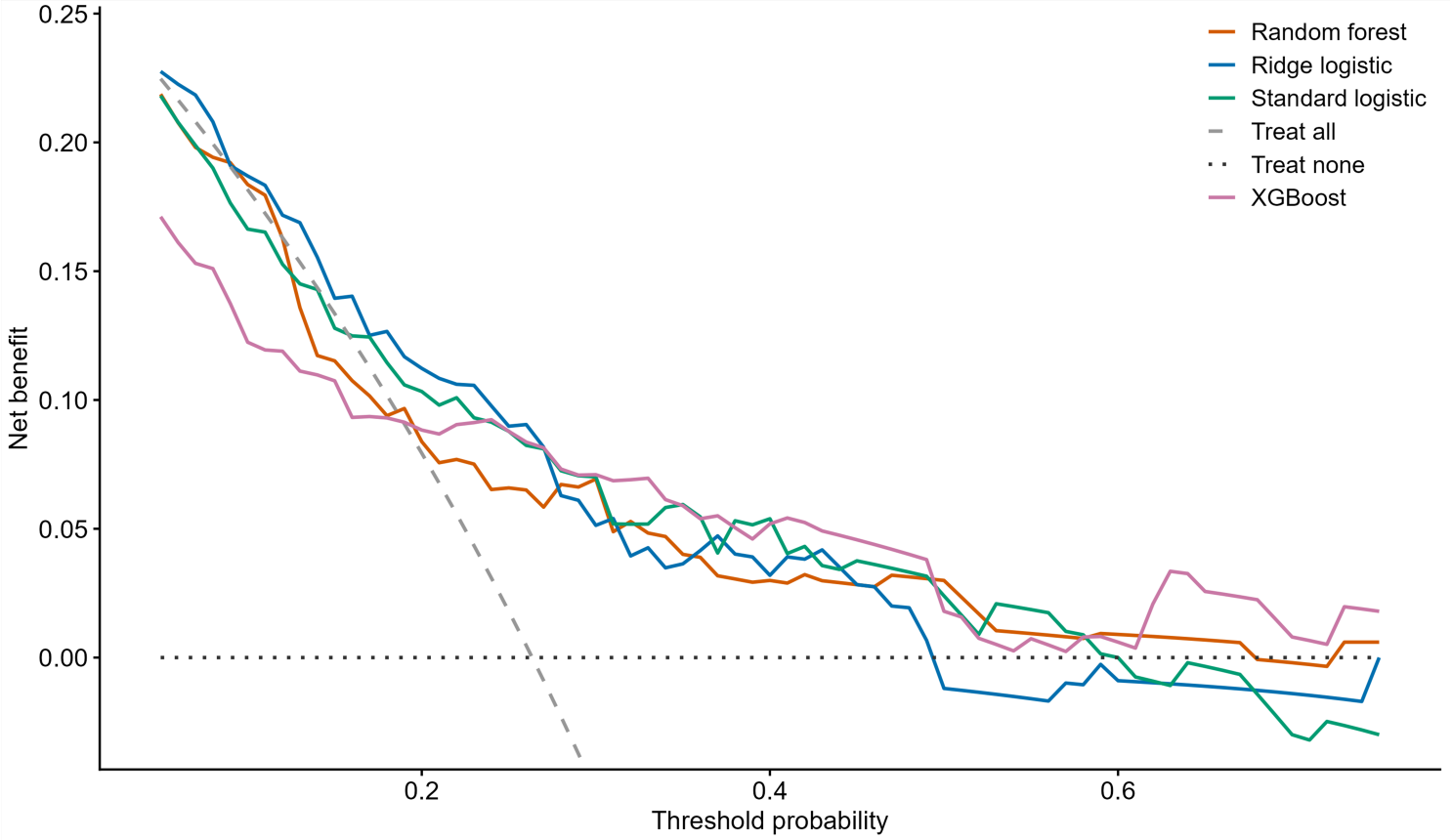


Figure 7

Decision-curve analysis in the prospective temporal cohort. Net benefit was greatest at thresholds of approximately 0.10–0.25; between-model differences were small.

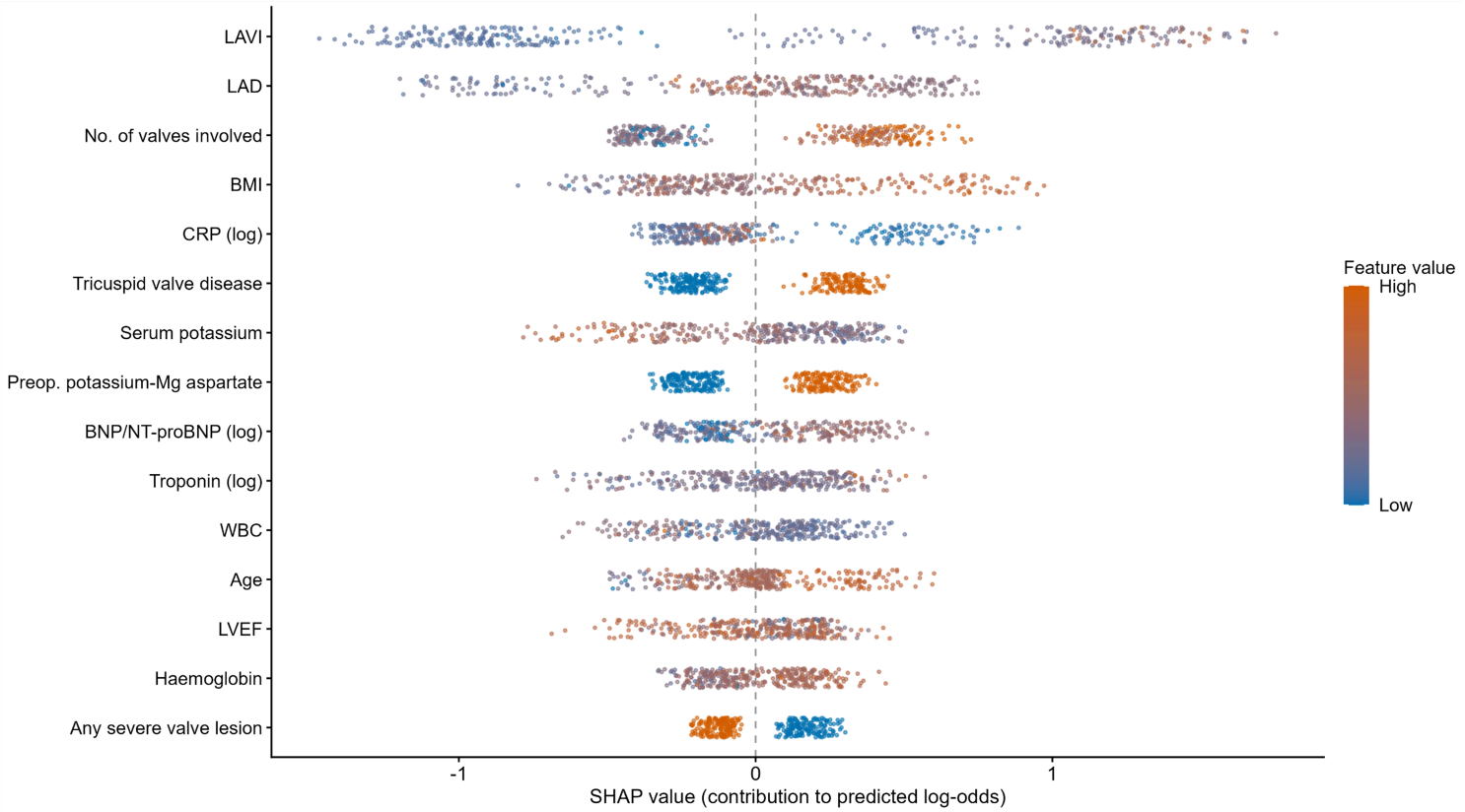


Figure 8

SHAP summary plot for the tree-based model. Feature contributions describe model behavior and are not causal effects.

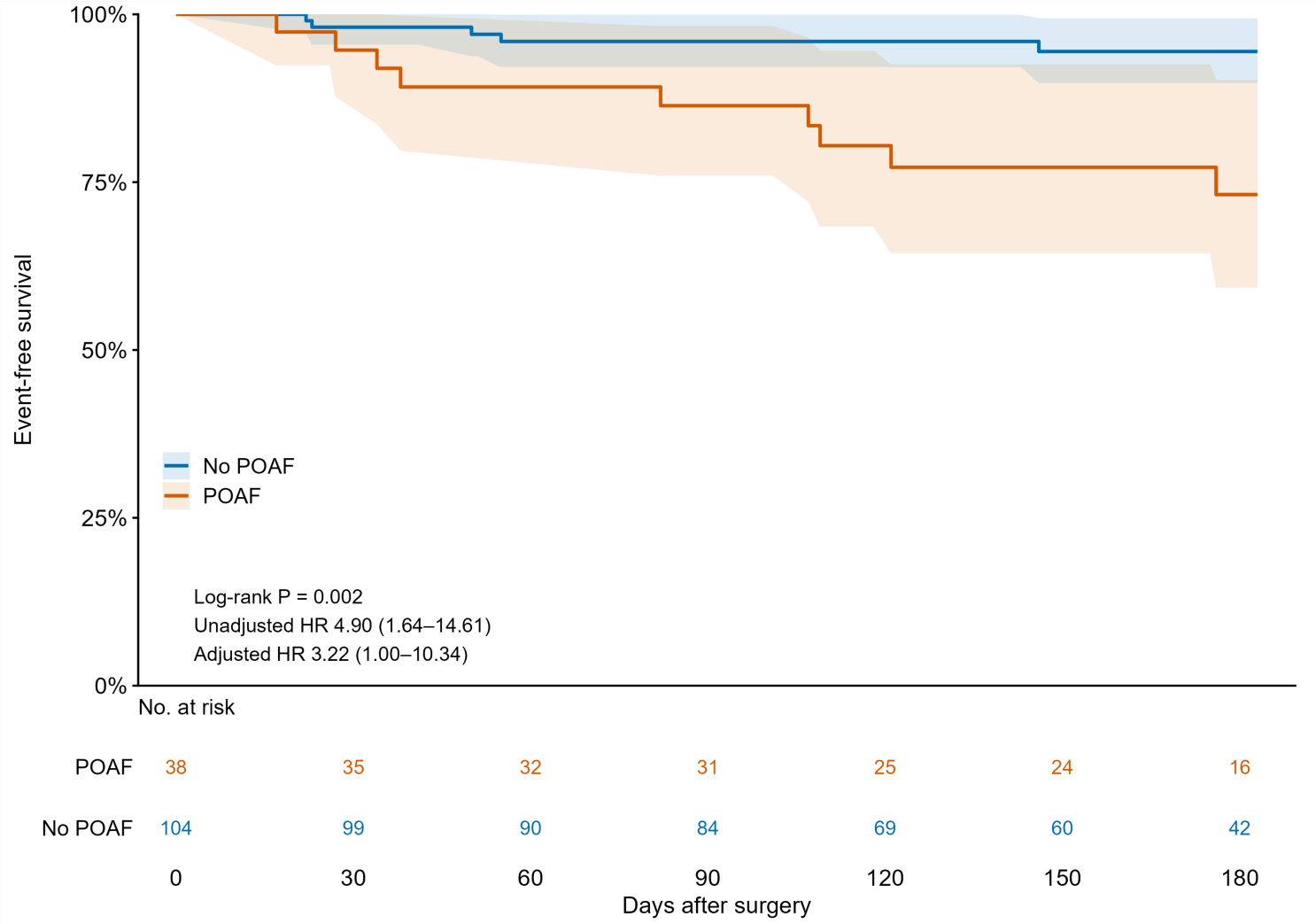


Figure 9

Kaplan–Meier event-free survival by in-hospital POAF status in the prospective cohort. Adjusted HR, 3.22 (95% CI, 1.00–10.34; P=0.050). The limited number of events results in wide uncertainty.

Supplementary Files

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

- SupplementaryFigures.zip
- SupplementaryTables.docx