Supplementary Materials

Pre-ablation substrate phenotypes and recurrence after atrial fibrillation catheter ablation: a retrospective cohort study

Zixi Zhang, Jiaxin Chi, Xunhua Wang, Xiaoyu Zeng, Haoqing Yu, Siyuan Tan, Ke Hu, Yichao Xiao, Tao Tu, Kangrong Li, Qiuzhen Lin, Qiming Liu, Yongguo Dai *, Chan Liu *

Supplementary Methods

Supplementary Method 1. Detailed exclusion criteria.

Patients were excluded from the study if they met any of the following criteria at baseline:

(1) Acute heart failure or hemodynamic instability (n = 8);

(2) Severe valvular heart disease requiring surgical or transcatheter intervention (n = 14);

(3) Acute myocardial infarction (n = 5);

(4) Severe hepatic or renal dysfunction (n = 9);

(5) Clinically significant endocrine disorders potentially affecting cardiometabolic or hemodynamic status, including primary hyperaldosteronism (n = 3), Cushing's syndrome (n = 3), or pheochromocytoma (n = 2);

(6) Active malignancy (n = 5), active infection (n = 4), or systemic autoimmune disease (n = 3);

(7) Previous AF catheter ablation (n = 88);

(8) Insufficient clinical, laboratory, procedural, or follow-up information required for the prespecified analyses (n = 8); or

(9) Pregnancy (n = 0).

Overall, 152 of the 1806 screened patients were excluded, leaving 1654 patients in the final analytical cohort.

Supplementary Method 2. Variable grouping and weighting for domain-balanced

Gower distance.

Seventeen prespecified pre-ablation variables were grouped into six clinically defined domains for calculation of the domain-balanced Gower distance. The cardiometabolic substrate domain included BMI, hypertension, diabetes mellitus, dyslipidemia, HbA1c, and log-transformed triglycerides. The inflammatory-nutritional status domain comprised log-transformed CRP and serum albumin. The renal-metabolic status domain included eGFR and uric acid. The atrial structural substrate domain comprised LAD. The myocardial stress-functional status domain included LVEF, heart failure, and log-transformed NT-proBNP. The systemic comorbidity domain comprised age, coronary heart disease, and obstructive sleep apnea.

To prevent domains containing more variables from disproportionately influencing phenotype assignment, each of the six domains was assigned equal total weight, with equal weighting of variables within each domain. Specifically, the weight assigned to variable j belonging to domain d was defined as $W_j = 1/(6P_d)$, where P_d denotes the number of variables in domain d. Accordingly, each domain contributed one-sixth of the total variable weight, irrespective of the number of constituent variables. Weighted Gower distances were calculated to accommodate both continuous and categorical variables, and the resulting dissimilarity matrix was used for PAM clustering. Recurrence outcomes were not incorporated into variable weighting, cluster selection, or phenotype assignment.

Supplementary Method 3. Implementation of internal temporal validation

Internal temporal validation was conducted using an earlier derivation cohort comprising 631 patients treated during 2009–2020 and a subsequent validation cohort comprising 1023 patients treated during 2021–2024. The temporal partition was defined independently of recurrence outcomes.

The domain-balanced Gower–PAM framework was established using the derivation cohort. The same 17 pre-ablation variables and six prespecified clinical domains were used as in the primary full-cohort analysis. Variable transformations, domain weights, continuous-variable scaling parameters, and phenotype medoids were determined in the derivation cohort and fixed before application to the validation cohort. Each validation patient was assigned to the nearest derivation-defined medoid according to the frozen weighted Gower distance specification, without refitting or updating the derivation-defined clustering solution.

To assess temporal reproducibility independently of the frozen assignments, a separate $k = 2$ PAM clustering solution was derived using the validation cohort alone. A weighted Gower distance matrix was calculated from the validation-cohort observations using the prespecified variable definitions and domain-weighting framework. PAM clustering was then performed independently on this matrix, without incorporating derivation-cohort medoids, frozen phenotype assignments, predefined cluster labels, or recurrence outcomes. Cluster labels were aligned after clustering to account for arbitrary label permutations before comparing the two assignment procedures.

Agreement between frozen phenotype assignments and independently derived

validation clusters was quantified using the ARI, overall assignment agreement, and phenotype-specific Jaccard indices. Overall agreement was calculated as the proportion of validation patients assigned to the corresponding phenotype by both procedures after label alignment. For each phenotype, the Jaccard index was defined as the number of patients assigned to that phenotype by both procedures divided by the number assigned to that phenotype by either procedure. The complete cross-classification matrix was reported to permit independent verification of these agreement measures.

Average silhouette width was calculated to characterize cluster separation. Because silhouette width and assignment agreement assess distinct properties of a clustering solution, both measures were considered when evaluating temporal reproducibility. Silhouette width was interpreted as a measure of separation between phenotypic partitions, whereas assignment agreement characterized the consistency of classifications obtained using the two procedures.

Associations between frozen phenotype assignments and recurrence outcomes were subsequently evaluated in the validation cohort using the same outcome definitions, 3-month landmark framework, and Model 4 adjustment strategy as in the primary analyses. The lower systemic-burden phenotype served as the reference. Recurrence outcomes were not incorporated into either phenotype-assignment procedure.

The primary full-cohort phenotyping analysis and the derivation-defined frozen assignment procedure were conducted separately. Consequently, phenotype assignments obtained using the frozen temporal-validation framework were not

required to reproduce those from the primary full-cohort clustering solution. The temporal-validation analysis was intended to assess the reproducibility of the phenotyping framework across treatment periods rather than establish independent external validity.

Supplementary Method 4. Incremental prognostic performance and OOB bootstrap evaluation

Incremental prognostic performance was evaluated in the 3-month landmark cohort comprising 1552 patients. The target outcome was the first documented post-blanking recurrence occurring by 12 months after catheter ablation. The prediction horizon was 9 months after the landmark, corresponding to 12 months after the index procedure.

A clinical benchmark Cox regression model included age, sex, AF type, BMI, LAD, LVEF, hypertension, diabetes mellitus, heart failure, and ablation calendar period. Two augmented models were constructed by separately adding binary phenotype membership or the standardized continuous substrate-position score to the same benchmark model. All three models were fitted and evaluated in the same analytical population.

Apparent prognostic performance was assessed using the time-dependent AUC, IPCW Brier score, and decision-curve net benefit. Incremental discrimination, prediction error, and net benefit were calculated relative to the clinical benchmark model. Nested models were compared using likelihood-ratio tests fitted to the full

analytical sample. Calibration was evaluated by comparing predicted recurrence probabilities with observed recurrence risks estimated with adjustment for censoring. Decision-curve analysis was performed across recurrence-risk thresholds of 5%–35%, with incremental net benefit summarized at thresholds of 10%, 15%, and 20%.

To assess the consistency of incremental prognostic performance, 500 OOB bootstrap resamples were performed. In each iteration, a bootstrap training sample was generated by sampling patients from the landmark cohort with replacement. Patients not selected in that iteration constituted the corresponding OOB evaluation sample. The clinical benchmark model and both augmented Cox models were independently refitted using the bootstrap training sample. Model-based recurrence-risk predictions at the prespecified horizon were then evaluated in the same OOB sample.

Within each OOB evaluation, incremental AUC, Brier score, and net benefit were calculated by comparing each augmented model with the clinical benchmark model. These paired comparisons ensured that performance differences were evaluated among the same patients within each resample. Positive ΔAUC and Δnet benefit indicated improvements relative to the benchmark model, whereas negative ΔBrier scores indicated reductions in prediction error.

The distributions of incremental performance measures across the 500 OOB evaluations were summarized using medians and empirical 2.5th–97.5th percentile ranges. These ranges characterized variability across resampled evaluations and were not interpreted as formal CIs for optimism-corrected incremental performance.

Binary phenotype assignments and continuous substrate-position scores were held

fixed throughout bootstrap resampling. Consequently, the OOB analyses evaluated the incremental prognostic performance of the Cox models conditional on the previously derived phenotypic representations and did not incorporate uncertainty arising from the initial clustering procedure or continuous score derivation.

The OOB analyses were intended to complement the apparent model comparisons by assessing the consistency of incremental discrimination, prediction error, and decision-analytic performance. The resulting estimates were interpreted separately from the apparent performance measures and were not considered independent external validation or complete internal validation of the combined phenotyping and prediction pipeline.

Supplementary Method 5. Research-use phenotype assignment interface.

For transparency and reproducibility, a web-based research interface was developed to illustrate the implementation of the phenotype-assignment procedure based on the primary full-cohort k = 2 clustering solution. The interface applies the fixed distance-based phenotyping framework to the same pre-ablation variables used for phenotype derivation and assigns a submitted patient profile to the nearest phenotype medoid according to the domain-balanced Gower distance framework.

The interface displays the assigned phenotype, relative distances to the lower systemic-burden and higher cardiometabolic-inflammatory phenotype medoids, and descriptive cohort-level information. The interface is available at https://zhangzixi237.shinyapps.io/af-phenotype-tool/.

The interface was developed for research demonstration and does not provide individualized recurrence-risk predictions or treatment recommendations. Its clinical applicability requires independent external validation.

Supplementary Figure Legends

Figure S1. RMST differences in recurrence-free survival after the 3-month landmark.

Differences in RMST between the higher cardiometabolic-inflammatory and lower systemic-burden phenotypes are shown at 36, 60, 90, and 120 months after the 3-month landmark. Differences were calculated as the higher cardiometabolic-inflammatory phenotype minus the lower systemic-burden phenotype, with negative values indicating shorter recurrence-free survival in the higher cardiometabolic-inflammatory phenotype. Error bars represent 95% bootstrap CIs based on 500 resamples.

Abbreviations: CI, confidence interval; RMST, restricted mean survival time.

Figure S2. Sensitivity analyses for follow-up post-blanking recurrence.

Adjusted HRs and 95% CIs are presented for the primary Model 4 analysis and sensitivity analyses incorporating 24-month administrative censoring, restriction to thermal ablation modalities, and additional adjustment for age, LAD, hypertension, and diabetes mellitus.

Abbreviations: CBA, cryoballoon ablation; CI, confidence interval; HR, hazard ratio; LAD, left atrial diameter; RFA, radiofrequency ablation.

Figure S3. Internal temporal validation of the frozen substrate-phenotyping framework.

The phenotyping framework was derived in patients treated during 2009–2020 and

applied without refitting to patients treated during 2021–2024. (A) Cross-classification of phenotype assignments obtained using the frozen derivation-defined framework and independently performed de novo clustering in the temporal-validation cohort (n = 1023). Agreement was summarized using the ARI and overall assignment agreement. (B) Adjusted associations between frozen phenotype assignments and early recurrence, post-blanking recurrence by 12 months after ablation, and follow-up post-blanking recurrence. Cox regression models used the Model 4 adjustment framework, with the lower systemic-burden phenotype as the reference. Early recurrence was assessed in the full temporal-validation cohort, whereas post-blanking outcomes were assessed in the corresponding 3-month landmark cohort.

Abbreviations: ARI, adjusted Rand index; CI, confidence interval; HR, hazard ratio; PAM, partitioning around medoids.

Figure S4. Continuous substrate-position score and follow-up post-blanking recurrence.

The outcome-independent substrate-position score was calculated as $S = d_{lower}/(d_{lower} + d_{higher})$, where x_i denotes an individual patient's pre-ablation profile, m_L and m_H represent the medoids of the lower systemic-burden and higher cardiometabolic-inflammatory phenotypes, respectively, and d denotes the domain-balanced Gower distance. Higher scores indicate greater proximity to the higher cardiometabolic-inflammatory phenotype medoid. Scores were standardized before analysis. (A) Distribution of the standardized substrate-position score according to

binary phenotype membership. (B) Association between the standardized score and follow-up post-blanking recurrence, modeled using restricted cubic splines with four knots at the 5th, 35th, 65th, and 95th percentiles. The solid line represents the estimated HR, and the shaded area indicates the 95% CI. (C) Comparison of model fit and discrimination among Cox models containing the Model 4 covariates alone, with binary phenotype membership added, or with the continuous substrate-position score added. Model fit was assessed using the AIC, and discrimination was evaluated using Harrell's C-index.

Abbreviations: AIC, Akaike information criterion; CI, confidence interval; HR, hazard ratio; SD, standard deviation.

Figure S5. Calibration of prognostic models for post-blanking recurrence by 12 months after ablation.

Calibration plots compare predicted recurrence risks with observed risks estimated using inverse probability of censoring weighting for the clinical benchmark model and models additionally incorporating binary phenotype membership or the continuous substrate-position score. The diagonal reference line represents perfect calibration.

Abbreviations: IPCW, inverse probability of censoring weighting.

Figure S6. Decision curve analysis for post-blanking recurrence by 12 months after ablation.

Decision curves show the net benefit of the clinical benchmark model and models

additionally incorporating binary phenotype membership or the continuous substrate-position score across recurrence-risk thresholds of 5%–35%. Treat-all and treat-none strategies are shown for comparison.

Figure S7. Research-use interface illustrating nearest-medoid phenotype assignment.

Abbreviations: AF, atrial fibrillation.

Table S1. Cluster stability assessment for the primary k = 2 phenotype solution.

Metric	Mean ± SD	Median (IQR)	Range
ARI	0.892 ± 0.071	0.907 (0.861–0.945)	0.602–0.998
Overall assignment agreement	0.972 ± 0.019	0.976 (0.962–0.986)	0.873–1.000
Jaccard index, lower phenotype	0.943 ± 0.038	0.950 (0.930–0.967)	0.746–0.999
Jaccard index, higher phenotype	0.946 ± 0.035	0.953 (0.935–0.970)	0.762–0.999
Average silhouette width	0.205 ± 0.006	0.205 (0.201–0.209)	0.179–0.227

Cluster stability was evaluated using 500 repeated 80% subsamples without replacement. The domain-balanced Gower–PAM clustering procedure was refitted within each subsample. Agreement with the reference clustering solution was evaluated using the ARI, overall assignment agreement, and phenotype-specific Jaccard indices. Abbreviations: ARI, adjusted Rand index; IQR, interquartile range; PAM, partitioning around medoids; SD, standard deviation.

Table S2. Proportional hazards diagnostics for the primary Cox regression models and supporting stratified sensitivity analysis.

Outcome / model	Term	Schoenfeld χ^2	Proportional hazards *P* value	Phenotype HR (95% CI)	Association *P* value
Early recurrence within 3 months — Model 4	Substrate phenotype	0.08	0.777	1.19 (0.80–1.79)	0.388
	Sex	0.12	0.729	—	—
	AF type	1.40	0.237	—	—
	Calendar period	0.30	0.861	—	—
	Smoking status	0.07	0.791	—	—
	Alcohol consumption	1.60	0.206	—	—
	Statin use	0.88	0.348	—	—
	β-blocker use	0.79	0.374	—	—
	ACEI/ARB use	0.82	0.365	—	—
	ARNI use	0.42	0.517	—	—
	VKA use	0.02	0.888	—	—
	NOAC use	0.25	0.617	—	—
	Amiodarone use	0.15	0.699	—	—
	Dronedarone use	0.54	0.462	—	—
	Global test	15.04	0.449	—	—
Post-blanking recurrence by 12 months — Model 4	Substrate phenotype	1.55	0.213	3.47 (2.54–4.75)	< 0.001
	Sex	0.17	0.680	—	—
	AF type	2.64	0.104	—	—
	Calendar period	1.20	0.549	—	—
	Smoking status	0.66	0.417	—	—

	Alcohol consumption	3.61	0.057	—	—
	Statin use	1.09	0.296	—	—
	β-blocker use	7.17	0.007	—	—
	ACEI/ARB use	0.92	0.337	—	—
	ARNI use	0.10	0.752	—	—
	VKA use	0.60	0.439	—	—
	NOAC use	0.01	0.920	—	—
	Amiodarone use	1.60	0.206	—	—
	Dronedarone use	4.96	0.026	—	—
	Global test	29.83	0.013	—	—
Post-blanking recurrence by 12 months — Stratified sensitivity model	Substrate phenotype	1.46	0.226	3.49 (2.55–4.77)	< 0.001
	β-blocker use	Stratified	—	—	—
	Dronedarone use	Stratified	—	—	—
	Global test	18.12	0.153	—	—
Follow-up post-blanking recurrence — Model 4	Substrate phenotype	0.04	0.841	2.85 (2.28–3.56)	< 0.001
	Sex	0.62	0.431	—	—
	AF type	0.41	0.522	—	—
	Calendar period	1.20	0.549	—	—
	Smoking status	0.84	0.359	—	—
	Alcohol consumption	1.85	0.174	—	—
	Statin use	4.36	0.037	—	—
	β-blocker use	2.94	0.086	—	—
	ACEI/ARB use	0.03	0.862	—	—
	ARNI use	0.29	0.590	—	—

	VKA use	1.82	0.177	—	—
	NOAC use	0.18	0.671	—	—
	Amiodarone use	0.37	0.543	—	—
	Dronedarone use	0.26	0.610	—	—
	Global test	17.68	0.280	—	—

The proportional hazards assumption was evaluated using scaled Schoenfeld residuals, with covariate-specific and global tests. Model 4 adjusted for sex, AF type, ablation calendar period, smoking status, alcohol consumption, and pre-ablation use of statins, β-blockers, ACEIs or ARBs, ARNIs, VKAs, NOACs, amiodarone, and dronedarone. Phenotype HRs compare the higher cardiometabolic-inflammatory phenotype with the lower systemic-burden phenotype.

For post-blanking recurrence by 12 months after ablation, the global test indicated a departure from the proportional hazards assumption, with covariate-specific evidence involving β-blocker and dronedarone use. A supporting Cox model was fitted with baseline hazards stratified by these medication variables. The phenotype association remained similar, and the global test was no longer statistically significant. For follow-up post-blanking recurrence, a covariate-specific signal was observed for statin use, whereas the global test was not statistically significant. Association *P* values and proportional hazards diagnostic *P* values refer to distinct statistical tests.

Abbreviations: ACEI, angiotensin-converting enzyme inhibitor; AF, atrial fibrillation; ARB, angiotensin receptor blocker; ARNI, angiotensin receptor–neprilysin inhibitor; CI, confidence interval; HR, hazard ratio; NOAC, non-vitamin K antagonist oral anticoagulant; VKA, vitamin K antagonist.

Table S3. Calendar-period sensitivity analyses for follow-up post-blanking recurrence.

Analysis	N	Events	HR (95% CI)	*P* value
Overall primary Model 4	1552	408	2.85 (2.28–3.56)	< 0.001
Phenotype × calendar-period interaction	1552	408	Not applicable	0.490
Period-specific analysis: 2009–2017	291	76	3.39 (2.20–5.20)	< 0.001
Period-specific analysis: 2018–2020	311	94	2.80 (1.80–4.35)	< 0.001
Period-specific analysis: 2021–2024	950	238	2.70 (2.01–3.62)	< 0.001

Calendar-period analyses used the primary full-cohort k = 2 phenotype assignments. The primary model included calendar period as an adjustment variable. Period-specific Cox models retained the other Model 4 covariates while omitting the calendar-period indicator. Effect modification was evaluated using phenotype-by-calendar-period interaction terms.

Abbreviations: CI, confidence interval; HR, hazard ratio.

Table S4. Cluster composition and phenotype labeling in the alternative k = 3 clustering solution.

k = 3 phenotype	Raw cluster ID	Number	Primary lower phenotype, n	Primary higher phenotype, n	Originally assigned to higher phenotype, %
Lower substrate-burden phenotype	3	520	500	20	3.8
Intermediate substrate-burden phenotype	1	590	270	320	54.2
Higher substrate-burden phenotype	2	544	2	542	99.6
Total	—	1654	772	882	53.3

The alternative k = 3 clustering solution was derived using the same domain-balanced Gower distance and PAM framework as the primary k = 2 analysis. Phenotype labels were assigned after examining the distributions of prespecified baseline clinical characteristics and their correspondence with the primary k = 2 phenotype assignments. Cross-classification with the primary k = 2 solution was used to characterize the composition of each k = 3 cluster. The proportion assigned to the primary higher cardiometabolic-inflammatory phenotype was calculated using the corresponding primary k = 2 assignments. Recurrence outcomes were not used for cluster derivation or phenotype labeling.

Abbreviations: PAM, partitioning around medoids.

Table S5. Recurrence outcomes according to the alternative k = 3 substrate phenotype solution.

Outcome	Lower phenotype	Intermediate phenotype	Higher phenotype
Early recurrence within 3 months	22/520 (4.2)	34/590 (5.8)	46/544 (8.5)
12-month post-blanking recurrence	39/498 (7.8)	84/556 (15.1)	110/498 (22.1)
Follow-up post-blanking recurrence	72/498 (14.5)	137/556 (24.6)	199/498 (40.0)

Values are presented as events/number of patients (percentage). Early recurrence was evaluated in the full cohort, whereas post-blanking outcomes were assessed in the 3-month landmark cohort. Post-blanking recurrence by 12 months referred to recurrence occurring after the landmark and within 12 months of the index ablation. Percentages represent observed event proportions and were not adjusted for censoring.

Table S6. Adjusted Cox associations for recurrence outcomes using the alternative k = 3 substrate phenotype solution.

Outcome	Contrast	Reference phenotype	Analysis N	Events	Adjusted HR (95% CI)	*P* value
Early recurrence within 3 months	Intermediate substrate-burden phenotype	Lower substrate-burden phenotype	1654	102	1.26 (0.74–2.17)	0.400
Early recurrence within 3 months	Higher substrate-burden phenotype	Lower substrate-burden phenotype	1654	102	1.83 (1.08–3.11)	0.025
12-month post-blanking recurrence	Intermediate substrate-burden phenotype	Lower substrate-burden phenotype	1552	233	1.77 (1.20–2.62)	0.004
12-month post-blanking recurrence	Higher substrate-burden phenotype	Lower substrate-burden phenotype	1552	233	2.86 (2.00–4.08)	< 0.001
Follow-up post-blanking recurrence	Intermediate substrate-burden phenotype	Lower substrate-burden phenotype	1552	408	1.59 (1.18–2.15)	0.002
Follow-up post-blanking recurrence	Higher substrate-burden phenotype	Lower substrate-burden phenotype	1552	408	2.65 (2.02–3.47)	< 0.001

Cox models used the Model 4 adjustment framework, including sex, AF type, calendar period, smoking status, alcohol consumption, and prespecified medication use. The lower substrate-burden phenotype in the alternative k = 3 solution served as the reference category. These analyses evaluated the sensitivity of phenotype–recurrence associations to the number of clusters.

Abbreviations: AF, atrial fibrillation; CI, confidence interval; HR, hazard ratio.

Table S7. Sensitivity analysis after re-deriving substrate phenotypes without age.

Analysis component	Measure / group	Full cohort, n	Landmark cohort, n	Events, n	Estimate / result	*P* value
No-age phenotype derivation	Lower no-age substrate-burden phenotype	843	798	129	129/798 (16.2%)	—
	Higher no-age substrate-burden phenotype	811	754	279	279/754 (37.0%)	—
Agreement with primary phenotype	Overall assignment agreement	1654	—	—	0.707	—
	ARI	1654	—	—	0.171	—
Cross-classification	Primary lower → No-age lower	565	—	—	—	—
	Primary lower → No-age higher	207	—	—	—	—
	Primary higher → No-age lower	278	—	—	—	—
	Primary higher → No-age higher	604	—	—	—	—
Cox regression	Model 4, no-age phenotype	—	1552	408	2.71 (2.19–3.35)	< 0.001
	Model 4 additionally adjusted for age	—	1552	408	2.06 (1.61–2.64)	< 0.001

Age was excluded from the clustering variables, and substrate phenotypes were re-derived using the domain-balanced Gower–PAM approach with k = 2. Agreement with the primary phenotype assignments was assessed using overall assignment agreement, the ARI, and cross-classification of phenotype membership. Associations with follow-up post-blanking recurrence were evaluated in the 3-month landmark cohort. The primary Cox model used the Model 4 adjustment framework, including sex, AF type, ablation calendar period, smoking status, alcohol consumption, and prespecified medication use. An additional model further adjusted for age. HRs compare the higher no-age substrate-burden phenotype with the lower no-age substrate-burden phenotype. Abbreviations: AF, atrial fibrillation; CI, confidence interval; HR, hazard ratio; PAM,

partitioning around medoids.

Table S8. Competing-risk analysis for follow-up post-blanking recurrence.

Analysis	Number	AF recurrence events	Competing deaths	Subdistribution HR (95% CI)	*P* value
Fine–Gray competing-risk model	1552	408	15	2.68 (2.16–3.32)	< 0.001

All-cause death before documented AF recurrence was treated as a competing event. Subdistribution HRs were estimated using Fine–Gray regression, incorporating the adjustment variables specified in Model 4. The higher cardiometabolic-inflammatory phenotype was compared with the lower systemic-burden phenotype.

Abbreviations: AF, atrial fibrillation; CI, confidence interval; HR, hazard ratio.

Table S9. Internal temporal validation of the frozen Gower–PAM phenotyping framework.

Component	Measure	Derivation cohort (2009–2020)	Temporal-validation cohort (2021–2024)	Adjusted HR (95% CI)	*P* value
Cohort size	Patients, n	631	1023	—	—
	Lower systemic-burden phenotype, n	297	398	—	—
	Higher cardiometabolic-inflammatory phenotype, n	334	625	—	—
Early recurrence	Events/number at risk	29/631	73/1023	1.29 (0.78–2.12)	0.318
12-month post-blanking recurrence	Events/number at risk	93/602	140/950	3.74 (2.36–5.91)	< 0.001
Follow-up post-blanking recurrence	Events/number at risk	170/602	238/950	2.79 (2.04–3.82)	< 0.001
Cluster separation	Average silhouette width, derivation	0.205	—	—	—
	Average silhouette width, frozen assignment	—	0.201	—	—
	Average silhouette width, independent de novo clustering	—	0.207	—	—
Assignment agreement	ARI	—	0.775	—	—
	Overall assignment agreement	—	94.0%	—	—
	Jaccard index, lower	—	0.866	—	—

	phenotype			
	Jaccard index, higher phenotype	—	0.903	—	—
Frozen vs independent clustering	Frozen lower / Independent lower	—	395	—	—
	Frozen lower / Independent higher	—	3	—	—
	Frozen higher / Independent lower	—	58	—	—
	Frozen higher / Independent higher	—	567	—	—

The domain-balanced Gower–PAM phenotyping framework was established in the earlier derivation cohort (2009–2020). Prespecified variable transformations, domain weights, scaling parameters, and phenotype medoids were fixed before application to the temporal-validation cohort (2021–2024). Each validation patient was assigned to the nearest derivation-defined medoid without refitting the clustering model. To assess temporal reproducibility, a separate k = 2 PAM solution was independently derived within the temporal-validation cohort without using the frozen phenotype assignments, derivation medoids, or recurrence outcomes. Cluster labels were aligned before comparing the frozen and independently derived assignments. Agreement was assessed using the ARI, overall assignment agreement, and phenotype-specific Jaccard indices. The frozen assignment and independent de novo clustering procedures were distinct from the primary full-cohort clustering analysis; consequently, the

phenotype distributions obtained from these procedures were not required to match the primary full-cohort phenotype distribution. Adjusted HRs refer to the associations between frozen phenotype assignments and recurrence outcomes in the temporal-validation cohort, using the Model 4 adjustment framework. The lower systemic-burden phenotype served as the reference.

Abbreviations: AF, atrial fibrillation; ARI, adjusted Rand index; CI, confidence interval; HR, hazard ratio; PAM, partitioning around medoids.

Table S10. Continuous substrate-position score and follow-up post-blanking recurrence.

Analysis	Number	Events	AIC	Harrell C-index	HR (95% CI)	*P* value
Covariate adjustment model	1552	408	5811.2	0.561	—	—
Covariate adjustment model + binary phenotype	1552	408	5719.5	0.652	2.85 (2.28–3.56)	< 0.001
Covariate adjustment model + continuous substrate-position score	1552	408	5727.4	0.655	1.80 (1.61–2.01)	< 0.001
Restricted cubic spline: Overall association	1552	408	—	—	—	< 0.001
Restricted cubic spline: Nonlinearity	1552	408	—	—	—	< 0.001

The continuous substrate-position score was defined as $S = d_{lower}/(d_{lower} + d_{higher})$, where d_{lower} and d_{higher} denote weighted Gower distances to the corresponding phenotype medoids. Higher scores indicate greater proximity to the higher cardiometabolic-inflammatory medoid. The score was standardized and evaluated per 1-SD increment. Nonlinearity was assessed using restricted cubic splines with four knots at the 5th, 35th, 65th, and 95th percentiles. The covariate adjustment model included the same covariates as Model 4. Lower AIC values indicate better relative model fit.

Abbreviations: AIC, Akaike information criterion; CI, confidence interval; HR, hazard ratio; SD, standard deviation.

Table S11. Apparent and OOB bootstrap incremental prognostic performance of phenotype-based representations beyond established clinical predictors.

Measure	Clinical benchmark	+ Binary phenotype	+ Continuous score
Apparent model performance			
Patients, n	1552	1552	1552
Recurrence events by 12 months, n	233	233	233
Time-dependent AUC	0.6941	0.7068	0.7063
ΔAUC	Reference	+0.0127	+0.0122
IPCW Brier score	0.1204	0.1189	0.1195
ΔBrier score	Reference	−0.0015	−0.0009
Net benefit at 10% threshold	0.0690	0.0731	0.0706
Net benefit at 15% threshold	0.0428	0.0507	0.0477
Net benefit at 20% threshold	0.0209	0.0280	0.0284
Likelihood-ratio test *P* value	—	< 0.001	< 0.001
OOB bootstrap incremental performance			
ΔAUC, median (2.5th–97.5th percentile)	Reference	+0.0155 (−0.0162 to 0.0330)	+0.0138 (−0.0123 to 0.0269)
ΔBrier score, median (2.5th–97.5th percentile)	Reference	−0.00134 (−0.00276 to 0.00142)	−0.00085 (−0.00212 to 0.00201)
ΔNet benefit at 10% threshold	Reference	+0.00620 (−0.00342 to 0.01642)	+0.00353 (−0.00503 to 0.01139)
ΔNet benefit at 15% threshold	Reference	+0.00711 (−0.00420 to 0.01965)	+0.00524 (−0.00424 to 0.01744)
ΔNet benefit at 20% threshold	Reference	+0.00234 (−0.01203 to 0.01407)	+0.00272 (−0.00982 to 0.01510)

The clinical benchmark model included age, sex, AF type, BMI, LAD, LVEF, hypertension, diabetes mellitus, heart failure, and ablation calendar period. Binary phenotype membership and the continuous substrate-position score were added separately to the clinical benchmark model. All

three models were evaluated in the same 3-month landmark cohort. The prediction horizon was 9 months after the landmark, corresponding to 12 months after catheter ablation. Apparent prognostic performance was evaluated using the time-dependent AUC, IPCW Brier score, and decision-curve net benefit. Incremental performance measures were calculated relative to the clinical benchmark model. Nested models were compared using likelihood-ratio tests. Exploratory internal validation was performed using 500 OOB bootstrap resamples. In each iteration, the three Cox regression models were refitted in the bootstrap training sample and evaluated in observations not included in that sample. Incremental discrimination, prediction error, and net benefit were calculated within the same OOB evaluation sample. Results are presented as medians and empirical 2.5th–97.5th percentile ranges across resamples. Phenotype assignments and substrate-position scores were held fixed during bootstrap resampling. Therefore, the OOB analysis assessed the incremental performance of the prediction models conditional on the previously derived phenotypic representations, rather than validating the complete phenotype-derivation pipeline. The empirical percentile ranges describe variability across OOB evaluations and were not interpreted as formal CIs for optimism-corrected performance.

Abbreviations: AF, atrial fibrillation; AUC, area under the receiver operating characteristic curve; BMI, body mass index; IPCW, inverse probability of censoring weighting; LAD, left atrial diameter; LVEF, left ventricular ejection fraction; OOB, out-of-bag.