SUPPLEMENTAL MATERIALS FOR:

Phenotyping of Atrial Fibrillation with Cluster Analysis and External Validation

Yuki Saito, MD, PhD[1]; Yuto Omae, PhD[2]; Koichi Nagashima MD, PhD[1]; Katsumi Miyauchi, MD, PhD[3]; Yuji Nishizaki, MD, MPH, PhD[3];
Sakiko Miyazaki, MD, MPH, PhD[3]; Hidemori Hayashi, MD, PhD[3]; Shuko Nojiri, MSc, PhD[4];
Hiroyuki Daida, MD, PhD[3]; Tohru Minamino, MD, PhD[3];
Yasuo Okumura, MD, PhD[1]

[1] Division of Cardiology, Department of Medicine, Nihon University School of Medicine, Tokyo, Japan; [2] Department of Industrial Engineering and Management, College of Industrial Technology, Nihon University, Chiba, Japan; [3] Department of Cardiovascular Biology and Medicine, Juntendo University Graduate School of Medicine, Tokyo, Japan; [4] Medical Technology Innovation Center, Juntendo University, Tokyo, Japan

Running title: Phenotyping of atrial fibrillation

*Corresponding author: Yuki Saito, MD
Division of Cardiology, Department of Medicine, Nihon University School of Medicine,
30-1 Ohyaguchi-kamicho, Itabashi-ku, Tokyo 173-8610, Japan
Tel: +81-3-3972-8111 Fax: +81-3-3972-1098
E-mail: saito.yuki@nihon-u.ac.jp

Supplemental Methods:

Study Population

SAKURA AF registry (derivation cohort)

Patients aged ≥20 years who were diagnosed with AF by 12-lead electrocardiograms (ECGs), 24-hour Holter ECG, or event-activated ECG and who were receiving warfarin or DOACs for stroke prophylaxis were included in the registry. Patients with rheumatic mitral valve disease, a history of prosthetic valve replacement, active infective endocarditis, or who did not provide written informed consent were not included. The patients had at least 2 years of follow-up examinations that ended on December 2017.

RAFFINE registry (external validation cohort)

Patients aged 20 years or older with AF documented by 12-lead electrocardiograms (ECGs) or 24-hour Holter ECGs were enrolled. Patients with a life expectancy of less than 1 year or those who did not provide written informed consent were excluded. All patients were followed up annually for at least 3 years and up to 5 years.

Definitions of variables

The CHA_2DS_2-VASc and HAS-BLED bleeding risk scores were calculated. CHA_2DS_2-VASc comprises congestive heart failure, hypertension, age ≥ 75 years (2 points), diabetes, stroke (2 points), vascular disease, age 65 to 74 years, and female sex;[1] and HAS-BLED bleeding risk score comprises hypertension, abnormal renal/liver function (1 point each), stroke, a bleeding history or predisposition, a labile international normalized ratio (INR; therapeutic time in a range [TTR] < 60%), age > 65 years, and use of drugs (antiplatelet agents and nonsteroidal anti-inflammatory drugs) or alcohol (> 8 U/week) (1 point each).[2] TTR was determined according to the method of Rosendaal et al.[3] In accordance with

Japanese recommendations, the target INR level was set at 1.6 to 2.6 for patients aged 70 years or older and 2.0 to 3.0 for patients younger than 70 years.[4]

Cluster analysis by machine learning

The k-means method which is one of the general clustering algorithms can only adopt numerical variables. In the present study, since we use the dataset containing a mixture of numerical and categorical variables, the k-prototype method (package “kmodes (ver 0.12.2)” on Python 3.8.9) which can adopt these variables has been adopted by this dataset.[5] The distance function for the k-prototype is defined by equation (9) of Huang et al. [6], and the first term represents the Euclidian distance for numerical variables and the second term represents simple matching dissimilarity measure (Hamming distance). This distance function is used to determine the final centroids to be used, iteratively changing the centroids for each cluster. We calculated the sum of squared errors (SSE) at K = 1 to 15 to determine the cluster size K. The SSE is the sum of the squared value of the distance between each sample and the centroid of each cluster, which is interpreted as the smaller the better. The SSE of each cluster size is shown in Supplemental Figure 1. The improved SSEs from K = 1 to 5 are large values. However, the improved SSEs at over K = 6 are small. Therefore, we adopted K = 5 as the number of clusters. This approach to determining K is called the “elbow method”.

Follow-up data and outcome assessment

In the SAKURA AF registry, a web-based registration system was established and follow-up data were collected twice a year (in March and September) by a central registry office for up to 4 years after enrolment. In the RAFFINE registry, follow-up data were collected annually after enrolment. The primary source of data is each patient's medical

Saito Y, *et al. Heart* 2023; 109:1751–1758. doi: 10.1136/heartjnl-2023-322447

record. Sites submit data via an electronic case report form or paper case report for. All patients were assigned a unique identifier and personally identifiable information was removed.

1. Lip GY, Nieuwlaat R, Pisters R, *et al*. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest* 2010;137(2):263-72.
2. Pisters R, Lane DA, Nieuwlaat R, *et al*. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey. *Chest* 2010;138(5):1093-100.
3. Rosendaal FR, Cannegieter SC, van der Meer FJ, *et al*. A method to determine the optimal intensity of oral anticoagulant therapy. *Thromb Haemost* 1993;69(3):236-9.
4. JCS Joint Working Group. Guidelines for Pharmacotherapy of Atrial Fibrillation (JCS 2013). *Circ J* 2014;78(8):1997-2021.
5. ANACONDA ORG, kmodes 0.12.2, https://anaconda.org/conda-forge/kmodes
6. Huang, Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 2, 283–304 (1998). https://doi.org/10.1023/A:1009769707641

Supplemental Results:

Sensitivity analysis

Development of the novel risk score based on clustering analysis

Firstly, we performed multiple logistic analyses to determine the variables to contribute to composing clusters 4 and 5 (vs. clusters 1, 2, and 3) in the derivation cohort. Male sex, CHA2DS2-VASCc score, HAS-BLED score, diabetes, haemoglobin, hypertension, age,

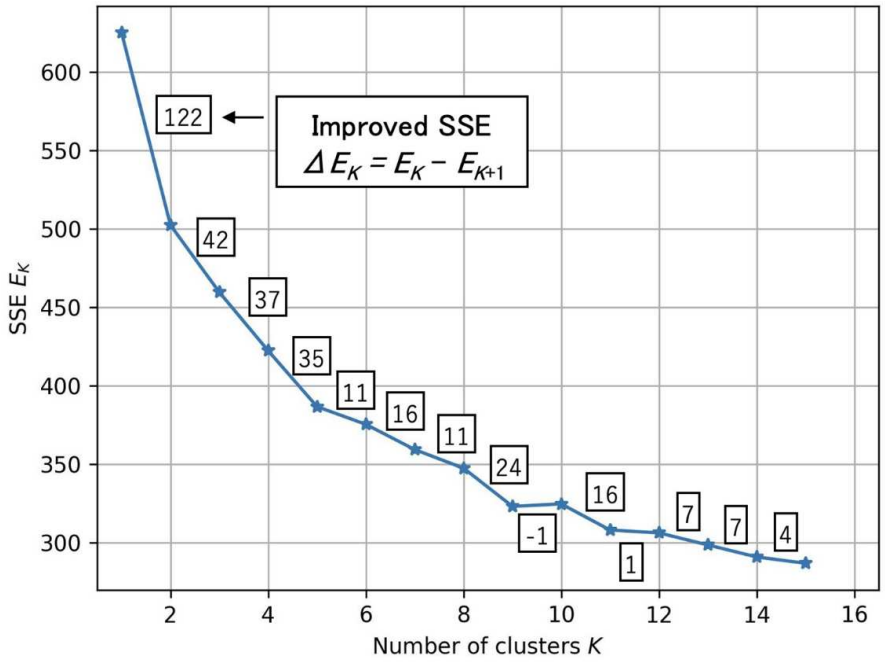
BMI, and CrCl were founded to be contributors to clusters 4 and 5. Secondly, a multivariable Cox regression analysis was performed to determine the weight of each variable for predicting the composite outcomes, which also delivered an estimate of the coefficients. We determined that haemoglobin, male sex, age, and HAS-BLED score were significantly and independently associated with the risk of composite outcomes in the derivation cohort. Finally, we constructed a novel risk score by adding the product of each predictor variable and the estimate of its coefficients derived from the multivariate Cox regression analysis, expressed as the equation below, which would predict the composite clinical events among AF patients:

Novel risk score = {(age) × 0.04 + (Male sex = 1) × 0.31 + (HAS-BLED score) × 0.21 - (haemoglobin) × 0.19} × 10

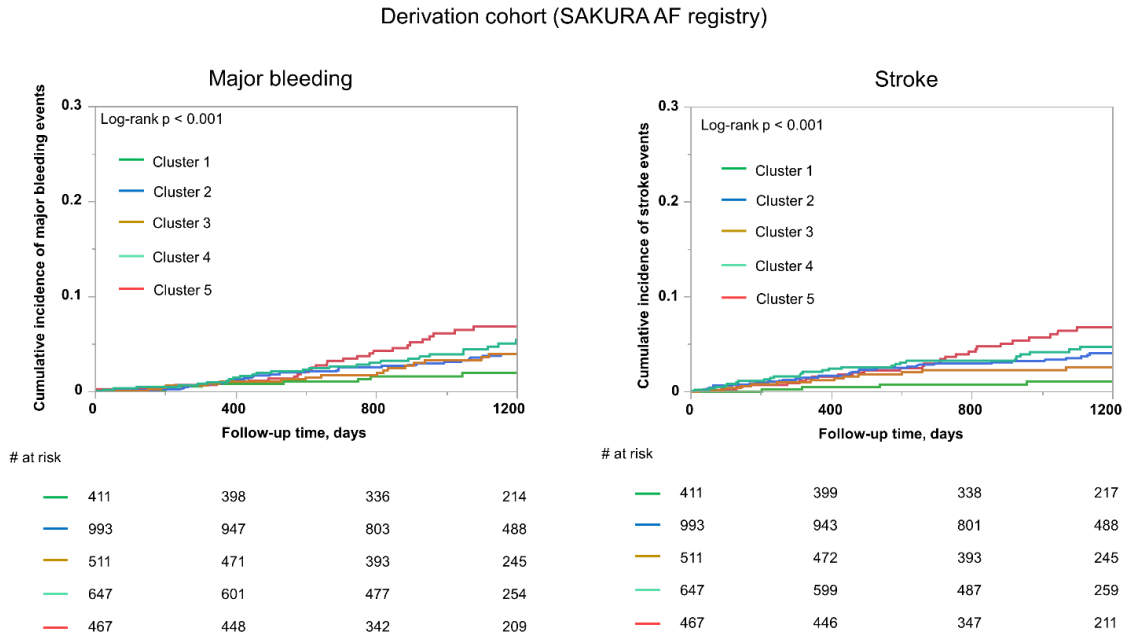
Furthermore, we evaluated the prediction performance of the novel risk score in the external validation cohort. In the external validation cohort, the median value of this score was 11.0 (IQR: 10.5, 11.4). In the receiver operating characteristic (ROC) curve analysis to predict the composite events within 3 years, the cut-off value of 11.2 present presented the area under the curve (AUC) :0.75, sensitivity: 0.70, specificity:0.70. In ROC analysis to predict all-cause mortality within 3 years, the cut-off value of 11.2 present presented the area under the curve (AUC) :0.75, sensitivity: 0.69, specificity:0.72. When we stratified the patients in the external validation cohort into four groups, which were Q1 (lowest risk group) to Q4 (highest risk group), according to the quartile of the novel risk score, Q4 had a significantly higher risk of all-cause mortality and the composite events (log-rank p<0.001, p<0.001, Supplemental Figure 6). In the multivariate Cox regression analysis, the novel risk score was significantly and strongly associated with the risk of all-cause mortality and the

risk of composite events, even after being adjusted for various risk factors (Supplemental Table 8).

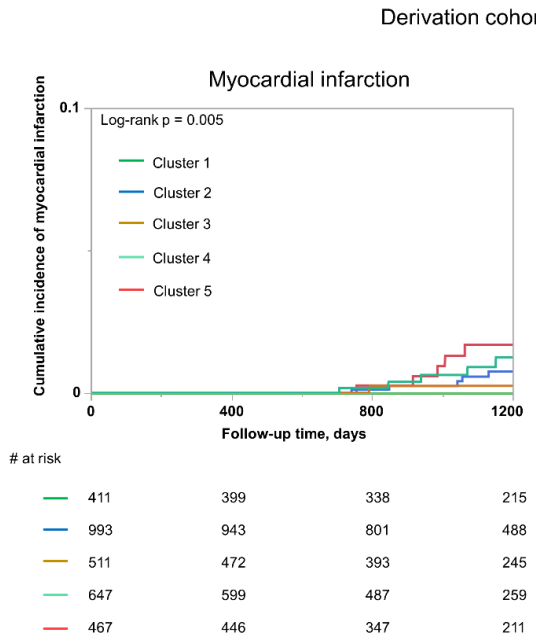
Supplemental Figures:



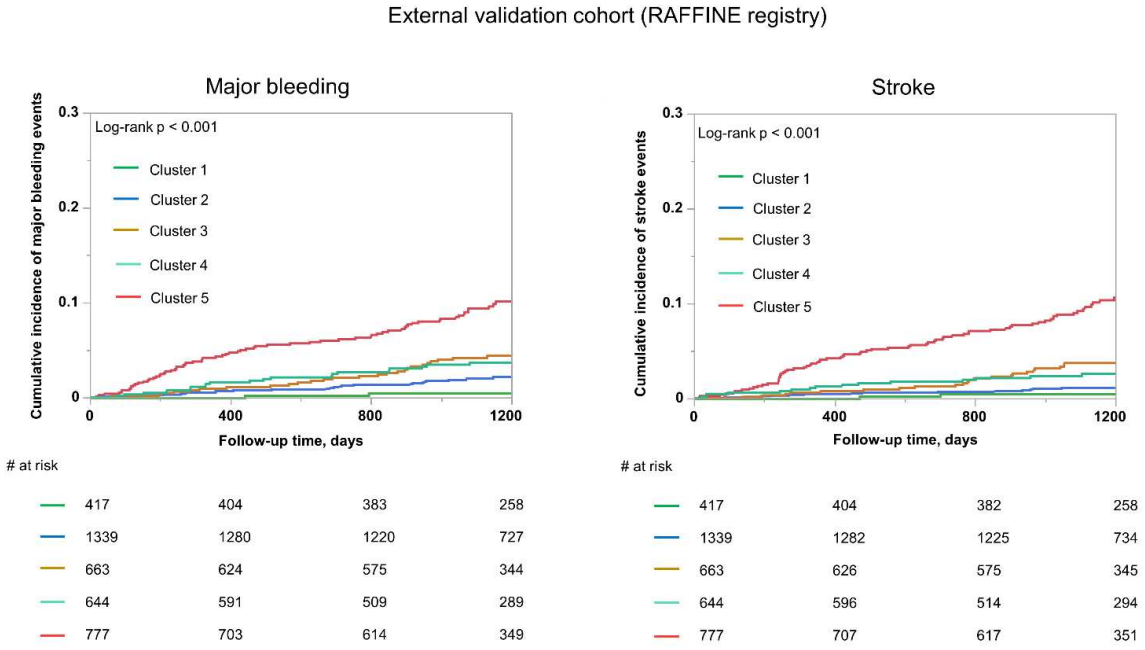
Supplemental Figure 1: Relationship between the sum of squared errors (SSE) and the number of clusters. The numbers in the boxes indicate the improved SSE.



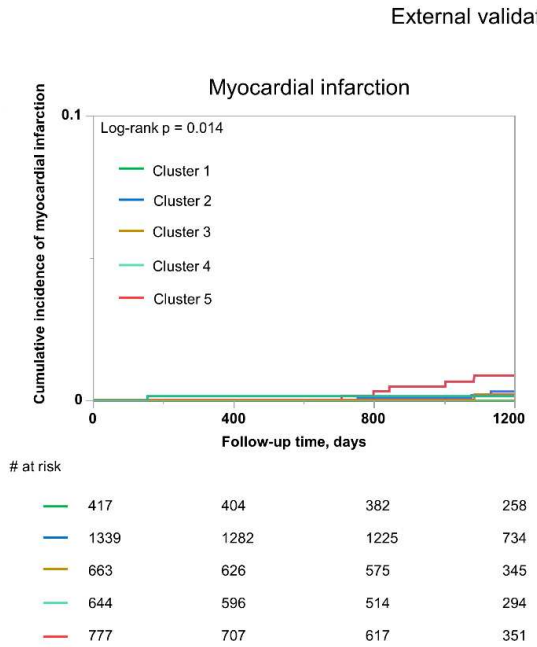
Supplemental Figure 2: Kaplan-Meier curves for the incidence of major bleeding (A) and stroke events (B) during the follow-up period according to the clusters in the derivation cohort (SAKURA AF registry)



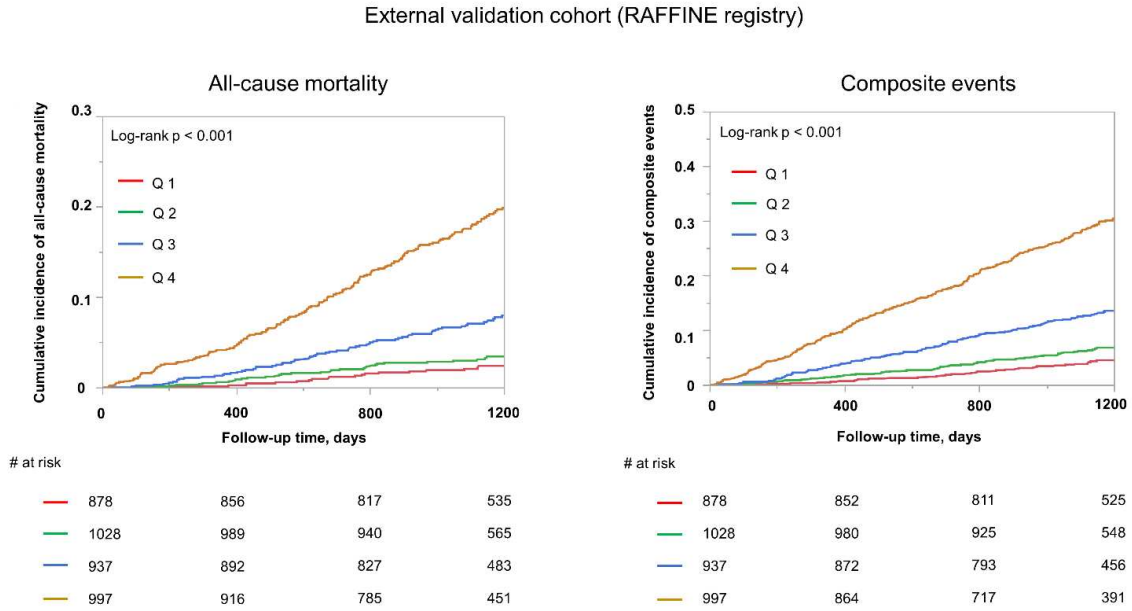
Supplemental Figure 3: Kaplan-Meier curves for the incidence of myocardial infarction during the follow-up period according to the clusters in the derivation cohort (SAKURA AF registry)



Supplemental Figure 4: Kaplan-Meier curves for the incidence of major bleeding (A) and stroke events (B) during the follow-up period according to the clusters in the external validation cohort (RAFFINE registry)



Supplemental Figure 5: Kaplan-Meier curves for the incidence of myocardial infarction during the follow-up period according to the clusters in the external validation cohort (RAFFINE registry)



Supplemental Figure 6: Kaplan-Meier curves for the incidence of composite events during the follow-up period according to the novel risk score (Q1-Q4) in the external validation cohort (RAFFINE registry)

Supplemental Tables:

Supplemental Table 1. The differences in baseline patient characteristics between the derivation and external validation cohorts

Item	Derivation cohort (n = 3055)	External validation cohort (n = 3852)	*P* value
Baseline clinical data			
Age, y	72.0 ± 9.4	72.1 ± 9.6	0.17
Male, n (%)	2250 (73.7)	2644 (68.6)	<0.001
Body mass index, kg/m^2	24.0 ± 3.7	23.7 ± 3.7	0.009
Systolic blood pressure, mm Hg	127 ± 16	125 ± 16	<0.001
AF type			<0.001
Paroxysmal AF, n (%)	1110 (36.7)	1457 (38.2)	-
Persistent AF, n (%)	683 (22.6)	355 (9.3)	-
Permanent AF, n (%)	1236 (40.8)	1998 (52.4)	-
Comorbidities			
Diabetes, n (%)	693 (22.7)	1165 (30.2)	<0.001
Hypertension, n (%)	2180 (71.4)	2806 (72.9)	0.17
History of heart failure, n (%)	679 (22.2)	917 (23.8)	0.12
Ischaemic heart disease, n (%)	291 (9.5)	531 (13.8)	<0.001
CHA_2DS_2-VASc score	3.0 ± 1.5	3.2 ± 1.6	<0.001
HAS-BLED score	1.4 ± 0.9	2.1 ± 0.8	<0.001
Medications			
Antiplatelet use, n (%)	482 (15.8)	994 (27.8)	<0.001
DOAC use, n (%)	1599 (52.3)	1659 (43.1)	<0.001
Warfarin use, n (%)	1456 (47.7)	1722 (44.7)	0.01
Laboratory data			
Haemoglobin, g/dL	13.8 ± 1.7	13.6 ± 1.7	<0.001
Platelet count, ×10^3 /μL	200 ± 59	193 ± 57	<0.001
Total cholesterol, mg/dL	185 ± 32	180 ± 32	<0.001

Triglycerides, mg/dL	137 ± 97	130 ± 85	0.007
Uric acid, mg/dL	5.9 ± 1.5	5.9 ± 1.5	0.88
AST, U/L	26 ± 13	25 ± 17	<0.001
ALT, U/L	22 ± 14	21 ± 15	0.02
BUN, mg/dL	18 ± 7	18 ± 7	0.95
Creatinine, mg/dL	0.9 ± 0.4	1.0 ± 0.7	<0.001
CrCl, mL/min	68 ± 27	68 ± 29	0.68
BNP, pg/mL	96 (51, 186)	110 (58, 197)	<0.001

Values are shown as mean ± SD or median (interquartile range) unless otherwise indicated. Abbreviations: AF, atrial fibrillation; ALT, alanine aminotransferase; AST, aspartate aminotransferase; BNP, B-type natriuretic peptide; BUN, blood urea nitrogen; CHA_2DS_2-VASc, congestive heart failure, hypertension, age ≥ 75 years, diabetes, stroke, vascular disease, age 65-74 years, and male; CrCl, creatinine clearance; DOAC, direct oral anticoagulant; HAS-BLED, hypertension, abnormal renal/liver function, stroke, bleeding, labile international normalized ratio, age > 65 years, and use of drugs/alcohol.

Supplemental Table 2. Baseline patient characteristics (included patients vs. excluded patients) in the derivation cohort (SAKURA AF registry)

Item	Included patients (n = 3055)	Excluded patients (n = 212)	*P* value
Baseline clinical data			
Age, y	72.0 ± 9.4	72.4 ± 9.2	0.23
Male, n (%)	2250 (73.7)	164 (77.4)	0.23
Body mass index, kg/m^2	24.0 ± 3.7	23.9 ± 3.9	0.45
Systolic blood pressure, mm Hg	127 ± 16	126 ± 15	0.30
AF type			0.003
Paroxysmal AF, n (%)	1110 (36.7)	101 (48.6)	-
Persistent AF, n (%)	683 (22.6)	40 (19.2)	-
Permanent AF, n (%)	1236 (40.8)	67 (32.2)	-
Comorbidities			
Diabetes, n (%)	693 (22.7)	52 (24.5)	0.54
Hypertension, n (%)	2180 (71.4)	151 (71.2)	0.96
History of heart failure, n (%)	679 (22.2)	43 (20.3)	0.51

Ischaemic heart disease, n (%)	291 (9.5)	22 (10.4)	0.69
CHA_2DS_2-VASc score	3.0 ± 1.5	3.0 ± 1.5	0.53
HAS-BLED score	1.4 ± 0.9	1.4 ± 0.7	0.54
Medications			
Antiplatelet use, n (%)	482 (15.8)	37 (17.5)	0.52
DOAC use, n (%)	1599 (52.3)	91 (42.9)	0.008
Warfarin use, n (%)	1456 (47.7)	121 (57.1)	0.008

Values are shown as mean ± SD or median (interquartile range) unless otherwise indicated. Abbreviations as Supplemental Table 1.

Supplemental Table 3. Baseline patient characteristics (included patients vs. excluded patients) in the external validation cohort (RAFFINE registry)

Item	Included patients (n = 3852)	Excluded patients (n = 37)	*P* value
Baseline clinical data			
Age, y	72.1 ± 9.6	70.4 ± 9.6	0.23
Male, n (%)	1208 (31.4)	14 (37.8)	0.41
Body mass index, kg/m^2	23.8 ± 3.7	24.1 ± 3.6	0.52
Systolic blood pressure, mm Hg	125 ± 16	129 ± 14	0.14
AF type			0.21
Paroxysmal AF, n (%)	1457 (38.2)	15 (42.9)	-
Persistent AF, n (%)	355 (9.3)	6 (17.1)	-
Permanent AF, n (%)	1998 (52.4)	14 (40.0)	-
Comorbidities			
Diabetes, n (%)	1165 (30.2)	14 (37.8)	0.33
Hypertension, n (%)	2806 (72.9)	21 (56.8)	0.037
History of heart failure, n (%)	917 (23.8)	8 (21.6)	0.75
Ischaemic heart disease, n (%)	531 (13.8)	2 (5.4)	0.10
CHA_2DS_2-VASc score	3.2 ± 1.6	3.1 ± 1.5	0.79
HAS-BLED score	2.1 ± 0.8	1.9 ± 0.9	0.08
Medications			
Antiplatelet use, n (%)	994 (27.8)	6 (20.0)	0.32
DOAC use, n (%)	1659 (43.1)	13 (35.1)	0.33
Warfarin use, n (%)	1722 (44.7)	13 (35.1)	0.24

Values are shown as mean ± SD or median (interquartile range) unless otherwise indicated.

Abbreviations as Supplemental Table 1.

Supplemental Table 4. Univariate Cox regression analysis for major bleeding and stroke risks in the derivation cohort (SAKURA AF registry)

Variable	Univariate analysis	
	HR (95% CI)	*P* value
Major bleeding		
Cluster 1 (ref)	-	-
Cluster 2	2.3 (1.0-5.1)	0.043
Cluster 3	1.9 (0.8-4.7)	0.14
Cluster 4	2.6 (1.1-6.1)	0.02
Cluster 5	3.7 (1.6-8.6)	0.001
Stroke		
Cluster 1 (ref)	-	-
Cluster 2	3.2 (1.3-8.3)	0.01
Cluster 3	2.2 (0.8-6.2)	0.12
Cluster 4	3.8 (1.5-9.9)	0.005
Cluster 5	5.6 (2.2-14.6)	<0.001

CI, confidence interval; HR, hazard ratio; ref, reference

Supplemental Table 5. Univariate Cox regression analysis for major bleeding and strok risks in the external validation cohort (RAFFINE registry)

Variable	Univariate analysis	
	HR (95% CI)	*P* value
Major bleeding		
Cluster 1 (ref)	-	-
Cluster 2	10.6 (1.5-77.8)	0.02
Cluster 3	10.4 (1.4-78.8)	0.02
Cluster 4	12.4 (1.7-93.9)	0.01
Cluster 5	37.4 (51.9-270.3)	<0.001
Stroke		
Cluster 1 (ref)	-	-
Cluster 2	3.8 (0.9-16.2)	0.07
Cluster 3	8.4 (2.0-35.4)	0.002
Cluster 4	9.3 (2.2-39.6)	0.002
Cluster 5	30.0 (7.4-121.9)	<0.001

CI, confidence interval; HR, hazard ratio; ref, reference.

Supplemental Table 6. Univariate and multivariate Cox regression analysis for all-cause mortality and composite event in the derivation cohort (SAKURA AF registry)

Variable	Univariate analysis		Multivariate analysis (Model 1)		Multivariate analysis (Model 2)	
	HR (95% CI)	*P* value	HR (95% CI)	*P* value	HR (95% CI)	*P* value
All-cause mortality						
Cluster 1 (ref)	-	-	-	-	-	-
Cluster 2	8.4 (2.0-35.0)	0.003	2.6 (0.6-11.7)	0.21	2.1 (0.5-9.1)	0.31
Cluster 3	17.4 (4.2-72.0)	<0.001	3.1 (0.7-13.7)	0.14	3.8 (0.9-16.4)	0.08
Cluster 4	22.0 (5.1-90.2)	<0.001	5.5 (1.1-27.1)	0.04	2.7 (0.6-12.2)	0.19
Cluster 5	18.2 (4.4-75.7)	<0.001	2.5 (0.5-12.4)	0.25	3.1 (0.7-13.9)	0.13
Composite events						
Cluster 1 (ref)	-	-	-	-	-	-
Cluster 2	4.2 (2.2-7.8)	<0.001	2.2 (1.1-4.4)	0.03	1.9 (1.0-3.8)	0.04
Cluster 3	5.2 (2.7-9.9)	<0.001	2.2 (1.1-4.4)	0.03	2.1 (1.1-4.3)	0.03
Cluster 4	6.3 (3.4-11.8)	<0.001	3.4 (1.5-7.8)	0.003	1.9 (0.9-3.9)	0.07
Cluster 5	7.9 (4.2-14.7)	<0.001	2.8 (1.2-6.2)	0.01	2.9 (1.5-5.9)	0.002

CI, confidence interval; HR, hazard ratio; ref, reference.
Model 1: adjusted for age (continuous variable), sex, CHA2DS2-VASCc (congestive heart failure, hypertension, age ≥ 75 years, diabetes, stroke, vascular disease, age 65-74 years, and male) score, comorbidities (hypertension and diabetes); Model 2: adjusted for age (continuous variable), sex, and type of atrial fibrillation.

Supplemental Table 7. Univariate and multivariate Cox regression analysis for all-cause mortality and composite event in the external validation cohort (RAFFINE registry)

Variable	Univariate analysis		Multivariate analysis (Model 1)		Multivariate analysis (Model 2)	
	HR (95% CI)	*P* value	HR (95% CI)	*P* value	HR (95% CI)	*P* value
All-cause mortality						
Cluster 1 (ref)	-	-	-	-	-	-
Cluster 2	3.0 (1.5-5.8)	<0.001	0.9 (0.4-2.0)	0.89	0.8 (0.4-1.8)	0.64
Cluster 3	3.7 (1.9-7.3)	<0.001	0.9 (0.4-1.9)	0.82	0.9 (0.4-2.0)	0.88
Cluster 4	7.2 (3.7-13.9)	<0.001	2.7 (1.0-5.3)	0.06	1.1 (0.5-2.5)	0.74
Cluster 5	8.1 (4.2-15.4)	<0.001	1.7 (0.7-4.0)	0.19	1.4 (0.7-3.1)	0.31
Composite events						
Cluster 1 (ref)	-	-	-	-	-	-
Cluster 2	3.4 (2.0-6.0)	<0.001	1.8 (1.0-3.4)	0.07	1.4 (0.8-2.5)	0.28
Cluster 3	4.8 (2.7-8.5)	<0.001	1.8 (1.0-3.4)	0.04	1.8 (0.9-3.3)	0.07
Cluster 4	7.2 (4.1-12.5)	<0.001	3.4 (1.6-6.3)	0.001	1.8 (0.9-3.5)	0.06
Cluster 5	12.2 (7.1-21.0)	<0.001	6.6 (3.3-13.2)	<0.001	3.5 (1.9-6.5)	<0.001

CI, confidence interval; HR, hazard ratio; ref, reference.
Model 1: adjusted for age (continuous variable), sex, CHA2DS2-VASCc (congestive heart failure, hypertension, age ≥ 75 years, diabetes, stroke, vascular disease, age 65-74 years, and male) score, comorbidities (hypertension and diabetes); Model 2: adjusted for age (continuous variable), sex, and type of atrial fibrillation.

Supplemental Table 8. Univariate and multivariate Cox regression analysis for all-cause mortality and composite event in the external validation cohort (RAFFINE registry)

Novel risk score (per 0.1 increase)	***All-cause mortality***		***Composite events***	
Models	**HR (95% CI)**	**P value**	**HR (95% CI)**	**P value**
Univariable analysis	1.16 (1.14-1.19)	<0.001	1.16 (1.14-1.18)	<0.001
Model 1	1.14 (1.11-1.17)	<0.001	1.18 (1.15-1.20)	<0.001
Model 2	1.15 (1.12-1.18)	<0.001	1.18 (1.15-1.20)	<0.001

CI, confidence interval; HR, hazard ratio; ref, reference.

Model 1: adjusted for age (continuous variable), sex, CHA2DS2-VASCc (congestive heart failure, hypertension, age ≥ 75 years, diabetes, stroke, vascular disease, age 65-74 years, and male) score, comorbidities (hypertension and diabetes); Model 2: adjusted for age (continuous variable), sex, and type of atrial fibrillation.