Predicting cerebral infarction in patients with atrial fibrillation using machine learning: The Fushimi AF registry

Hidehisa Nishi[1,2] ⓘ, Naoya Oishi[3] ⓘ, Hisashi Ogawa[4],
Kishida Natsue[1], Kento Doi[1], Osamu Kawakami[1],
Tomokazu Aoki[1], Shunichi Fukuda[1], Masaharu Akao[4] and
Tetsuya Tsukahara[1];
on behalf of the Fushimi AF Registry investigators.

Abstract

The $CHADS_2$ and CHA_2DS_2-VASc scores are widely used to assess ischemic risk in the patients with atrial fibrillation (AF). However, the discrimination performance of these scores is limited. Using the data from a community-based prospective cohort study, we sought to construct a machine learning-based prediction model for cerebral infarction in patients with AF, and to compare its performance with the existing scores. All consecutive patients with AF treated at 81 study institutions from March 2011 to May 2017 were enrolled (n = 4396). The whole dataset was divided into a derivation cohort (n = 1005) and validation cohort (n = 752) after excluding the patients with valvular AF and anti-coagulation therapy. Using the derivation cohort dataset, a machine learning model based on gradient boosting tree algorithm (ML) was built to predict cerebral infarction. In the validation cohort, the receiver operating characteristic area under the curve of the ML model was higher than those of the existing models according to the Hanley and McNeil method: ML, 0.72 (95%CI, 0.66–0.79); $CHADS_2$, 0.61 (95%CI, 0.53–0.69); CHA_2DS_2-VASc, 0.62 (95%CI, 0.54–0.70). As a conclusion, machine learning algorithm have the potential to perform better than the $CHADS_2$ and CHA_2DS_2-VASc scores for predicting cerebral infarction in patients with non-valvular AF.

Introduction

Patients with atrial fibrillation (AF) are at high risk for cardiogenic stroke.[1] The $CHADS_2$ and CHA_2DS_2-VASc scoring systems are often used to assess the risk of future cerebral infarction and to make decisions about anticoagulation.[2,3] However, the $CHADS_2$ and CHA_2DS_2-VASc scores are not highly discriminative for ischemic events.[4]

Many other risk factors for ischemic event not included in the $CHADS_2$ and CHA_2DS_2-VASc scores have been reported in the literature. For example, the type of AF, renal disease, and left atrial size on echo-cardiography.[4–6] Theoretically, a model that effectively utilizes a wide range of risk factors has the potential to outperform the $CHADS_2$ and CHA_2DS_2-VASc scores, but in the standard statistical approaches, which commonly utilize logistic regression, the model tends to become unstable as the number of variables increases.[7]

[1]Department of Neurosurgery, National Hospital Organization Kyoto Medical Center, Kyoto, Japan
[2]Department of Neurosurgery, Kyoto University Graduate School of Medicine, Kyoto, Japan
[3]Medical Innovation Center, Kyoto University Graduate School of Medicine, Kyoto, Japan
[4]Department of Cardiology, National Hospital Organization Kyoto Medical Center, Kyoto, Japan

Corresponding author:
Naoya Oishi, Medical Innovation Center, Kyoto University Graduate School of Medicine, 53 Shogoin Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan. Email: noishi@kuhp.kyoto-u.ac.jp

Machine learning could be an alternative approach to deal with the problem above. Machine learning is a subset of artificial intelligence in which algorithms learn from data without explicit programming. Unlike a traditional logistic regression modelling, machine learning can control the collinearity of variables with regularization, and is effective in modeling multifactorial events in various fields.[8] Therefore, we aimed to construct a model to predict future cerebral infarction using machine learning in order to effectively utilize many risk factors. More accurate risk stratification of patients with AF will lead to more appropriate decision-making with respect to anticoagulation therapy.

Using data from a large-scale community-based prospective registry of AF patients, we constructed a machine learning model for predicting future cerebral infarction in patients with AF who are not receiving anticoagulation therapy, validated its performance, and compared its performance with those of the $CHADS_2$ and CHA_2DS_2-VASc scores.[9]

Materials and methods

Patient population

The detailed study design, patient enrollment, definitions of measurements, and baseline clinical characteristics of subjects of the Fushimi AF Registry were previously described (UMIN Clinical Trials Registry: UMIN000005834). The inclusion criterion was the documentation of AF by 12-lead electrocardiogram or Holter monitoring at any time. There were no exclusion criteria. Eighty-one institutions participated in the registry: 2 cardiovascular centers (National Hospital Organization Kyoto Medical Center and Ijinkai Takeda Hospital), 9 small- and medium-sized hospitals, and 70 primary care clinics. Patient enrollment began in March 2011 and the participating institutions attempted to enroll all consecutive patients with AF under regular outpatient care or under admission. Clinical data were registered in the Internet database system (http://edmsweb16.eps.co.jp/edmsweb/002001/FAF/top.html) at each institution. Data were automatically checked for missing or contradictory entries and values outside the normal range. Additional editing checks were performed by clinical research coordinators at the general office of the registry. Follow-up data were collected at least annually by the doctors in charge or telephone follow-up.

In this analysis, the data of the patients from one cardiovascular center (n = 958) and the 19 hospitals and primary care clinics (n = 1426) were randomly assigned as the derivation cohort (n = 2382), and the data of the patients from another cardiovascular center (n = 609) and an additional 20 hospitals and primary care clinics (n = 1403) were assigned as the validation cohort (n = 2012) (Figure 2). We excluded the patients with valvular AF and the patients under anticoagulation therapy, and also the patients with an observation period of less than 1 year without events in order to reduce potential bias due to the short observation period.

The study protocol conforms to the ethical guidelines of the 1975 Declaration of Helsinki, and was approved by the ethical committees of the National Hospital Organization Kyoto Medical Center and Ijinkai Takeda General Hospital. Since the present research was part of an observational study not using human biological specimens, consent to participate in this study was obtained with an opt-out approach and written informed consent was not obtained from each patient; this design conformed to the ethical guidelines for epidemiological research issued by the Ministry of Education, Culture, Sports, Science and Technology and Ministry of Health, Labour and Welfare, Japan. However, we have published all relevant details of this study to allow its replication, and we provided each patient an opportunity to refuse inclusion in this research by posting the details at each participating clinic and at the homepages of our institutions.

Prediction target: cerebral infarction during follow-up

The prediction target in the analysis was the incidence of cerebral infarction during follow-up. Stroke was defined as the sudden onset of a focal neurologic deficit in a location consistent with the territory of a major cerebral artery during follow-up, and the diagnosis of cerebral infarction was confirmed by computed tomography or magnetic resonance imaging (MRI). As described above, follow-up data were collected annually through review of the inpatient and outpatient medical records, and through contact with patients, relatives and/or referring physicians by mail or telephone.

Standard prediction models (scoring methods): $CHADS_2$, CHA_2DS_2-VASc score

The $CHADS_2$ scoring system assigns 1 point each for congestive heart failure, hypertension, age ≥ 75 years, and diabetes mellitus, and 2 points for prior stroke or transient ischemic attack (TIA).[2] The CHA_2DS_2-VASc scoring system is identical to the $CHADS_2$ scoring system except that it adds an extra point each for female sex and vascular disease (myocardial infarction and peripheral vascular disease), and divides age into 3 categories (<60, 60—74, ≥ 75 years) instead of the 2 categories used in the $CHADS_2$ score.[3]

Overview of the machine learning model development

The entire data was divided into training data (derivation cohort) and validation data (validation cohort), and only the training data was used to build the machine learning model. Since the number of baseline variables in the database was large, we first preprocessed the variables and assessed the importance of variables in predicting target. Then, only a relatively small number of important variables were used to build the final prediction model (Figure 1). The performance of the constructed model was evaluated using the validation data and compared to the existing scores.

Variables and data preprocessing for machine learning model

The data include baseline demographic information, such as the patient's age, sex, comorbidity, past medical history, social history, and the results of blood tests, chest X-ray, and transthoracic echocardiography.

A total of 168 baseline variables were included in the dataset.

For data preprocessing, variables that were not considered clinically meaningful were deleted, such as patient ID, the date of inclusion, and the name of the study institution. Also, specific drug names were replaced with the appropriate category name. For example, aspirin, clopidogrel, ticlopidine, prasugrel, and cilostazol were all converted into the categorical variable "antiplatelet drugs". In addition, several variables were created using existing variables. For example, body mass index and body surface area were calculated using patient's height and weight.

Variables for which more than 30% of the values were missing in the derivation cohort were deleted. Then the missing values were imputed using the mean value from the derivation cohort. After all, 68 variables were listed as candidates for constructing model. All the variables after data preprocessing and its numbers of missing data points are listed in supplemental Tables II.

For model evaluation with the validation cohort, the missing values were imputed 20 times with multiple imputation with chained equations (MICE) to address
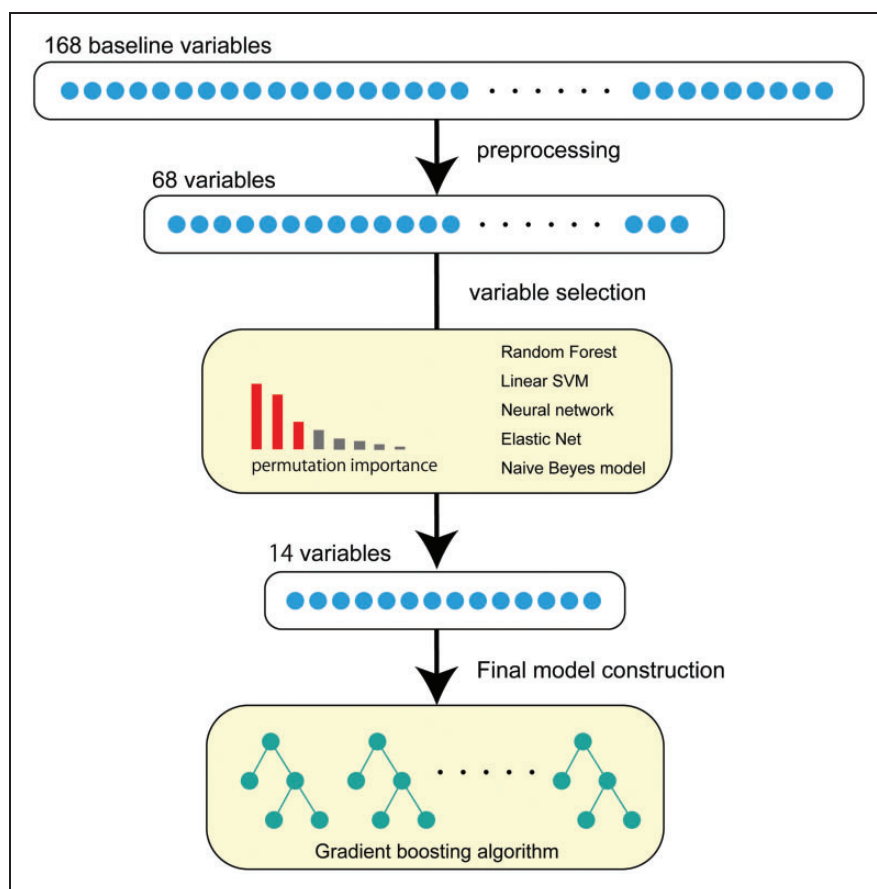

Figure 1. The processing steps of the machine learning model development in this study.

the randomness of the estimation.[10,11] In the imputation process, all other variables in the validation cohort were used to create imputed results; data from the derivation cohort were not used with MICE.

Variable selection for the machine learning model

Although there were as many as 68 baseline variables in the database, several variables were extracted for model construction with a view to future practicality.

In this step, we utilized 5-fold cross-validation in the derivation cohort: the derivation cohort data were divided into training and test data in 5 patterns without duplication. In each step, machine learning models based on five different algorithms (random forest, regularized logistic regression, linear support vector machine, neural network, and naïve Beyes model) were separately trained and the importance of the variables was evaluated by permutation importance. Permutation importance is defined as the decrease in the model score when a single variable value is randomly shuffled.[12] Since this procedure breaks the relationship between the variable and the target, the decrease in the model score indicates how much the model depends on the variable. In this study, the shuffling of the variables was repeated 10 times. After 5-fold cross validation, the mean, variance, and t statistics on the distribution of the importance of each variable were calculated. Variables with significantly higher t statistics were selected for the final model (in this case, $p < 0.0001$ was set as the cut-off). These operations were performed on each algorithm separately.

Model derivation and internal validation

For the final machine learning model, CatBoost was chosen for the model algorithm. CatBoost is an algorithm for gradient boosting on decision trees; it was developed to handle categorical variables and implements a framework to avoid over-fitting.[13] Same as the variable selection described above, we performed 5-fold cross-validation in the derivation cohort for the model training. In the training step, model hyperparameters were optimized with a grid search algorithm. Grid search tunes and optimizes the model hyperparameters in a greedy way (the actual hyperparameters are shown in Supplemental Table I). After choosing the best hyperparameters, the model was trained again with the whole derivation cohort data.

For the internal evaluation, the sensitivity, specificity, accuracy, and area under the receiver operating curve (AUROC) were evaluated. In addition, since there was a strong imbalance between the classes, we also evaluated the F-1 score, the Matthews correlation coefficient and the area under the precision-recall curve (AUPRC).

The model algorithms, cross-validation, and grid search were based on the Python library scikit-learn.

Validation

After the model derivation, the machine learning model was evaluated for its performance and compared with the $CHADS_2$ and CHA_2DS_2-VASc scores on the validation cohort. As described in the preprocessing step above, the missing values were imputed 20 times with MICE.

The Kaplan-Meier curves were also plotted to display the clinical course for the subgroups stratified by each measurement on the validation cohort. The low-, moderate-, and high-risk subgroups were defined as below. For the $CHADS_2$ score, low-, moderate- and high-risk were respectively defined as scores of 0, 1, and 2—6, according to the previous reports.[14,15] For the CHA_2DS_2-VASc score, low-, moderate- and high-risk were respectively defined as scores of 0, 1, and 2—9, according to the previous report.[14] For the machine learning model, the distribution of the predicted probability was divided into quartiles; low-risk was defined as probability in the first quartile, moderate-risk as probability in the second and third quartiles, and high risk as probability in the fourth quartile of risk.

Statistical analysis

Continuous variables were expressed as mean ± standard deviation (SD) or median with interquartile range, depending on the distribution of variables. The normality of the distributions was assessed by Shapiro-Wilk test.

As described above, missing values were imputed with MICE and the averaged prediction probability for each case was used for the evaluation. Model performance metrics were displayed with 95% confidence intervals (CI) by the bootstrap method. Receiver operating curves were compared by the Henley and McNeil method.[16] Kaplan-Meier curves were plotted for each model to display the event incidence across time. For the moderate- and high-risk subgroups in each model, the hazard ratio of events was calculated using the Cox proportional hazards model.

Two-sided p-values <0.05 were considered to indicate statistical significance. The statistical analyses were performed with R statistical software (version 4.0.0).

Results

Data availability

All relevant data are available, upon reasonable request, from the corresponding author.

Baseline characteristics

Figure 2 shows the patient flow chart. The derivation cohort included the data for 2384 patients. After excluding 41 patients with valvular AF, 1222 patients receiving anticoagulation therapy, and 116 patients with less than 1 year of follow-up without any event, 1005 patients were entered into the analysis. The validation cohort included the data for 2012 patients. After excluding 46 patients with valvular AF, 1133 patients receiving anticoagulation therapy, and 81 patients with less than 1 year of follow-up without any event,

752 patients were entered into the analysis. The median follow-up period was 4.5 years in the derivation cohort and 5.1 years in the validation cohort.

The patient backgrounds are shown in Table 1. In brief, the patients in the derivation cohort had a more advanced type of AF and a higher frequency of stroke-related comorbidities. The previous histories of cerebral infarction, TIA and systemic thromboembolism were similar between the groups. There was no significant difference in the distributions of the $CHADS_2$ and CHA_2DS_2-VASc scores (the distribution of the scores is also shown in the supplemental figure IV).

Clinical outcomes

The clinical outcomes are shown in Table 2. Cerebral infarction (the prediction target of this study) occurred in 79 patients (7.9%) in the derivation cohort and 50
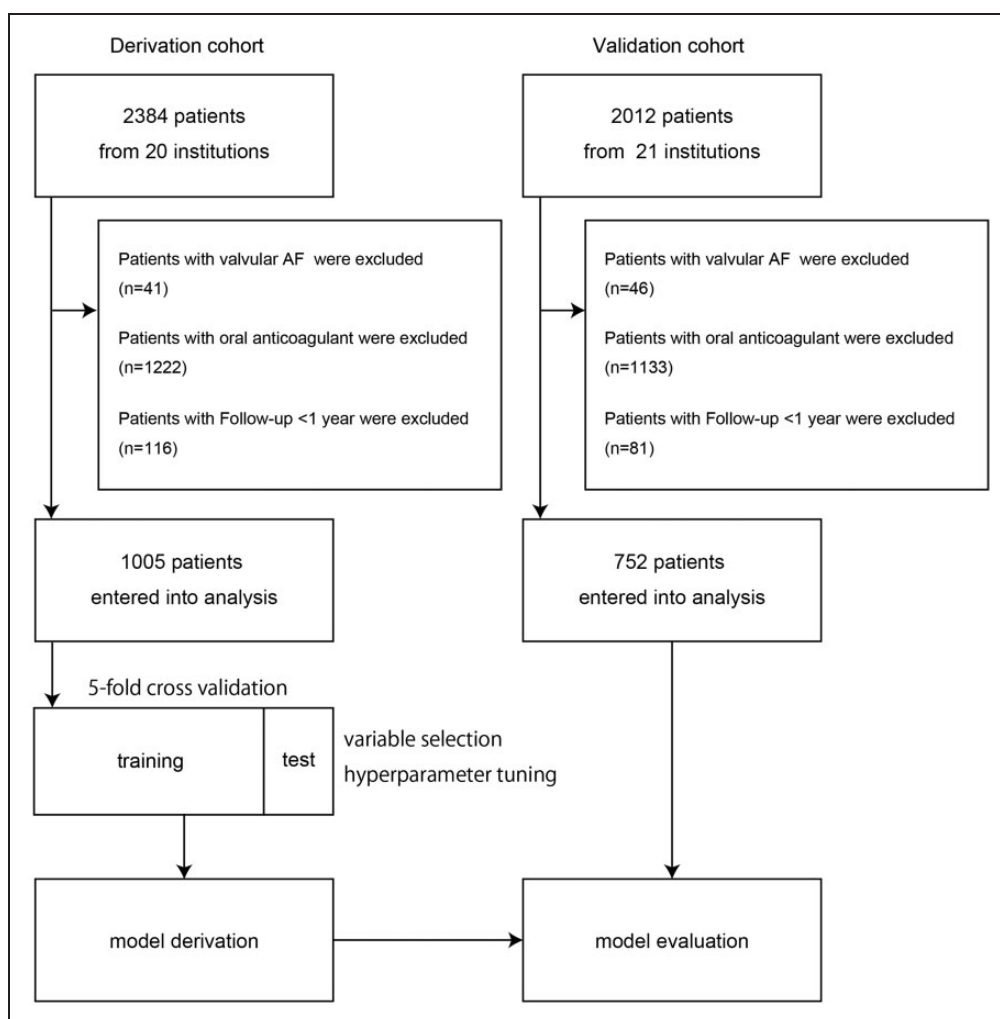

Figure 2. Flow diagram of the patients included in this study.

Table 1. Patient characteristics.

	Derivation cohort (n = 1005)	Validation cohort (n = 752)
Age, median (IQR), y	73 (66–81)	74 (64–81)
Sex		
Male, n (%)	587 (58.3)	413 (54.9)
Female, n (%)	419 (41.7)	339 (45.1)
Body weight, mean ± SD, kg	59.7 ± 13.9	58.8 ± 12.8
Types of atrial fibrillation		
Paroxysmal, n (%)	577 (57.4)	538 (71.5)
Sustained, n (%)	429 (42.6)	214 (28.5)
Comorbidities		
Hypertension, n (%)	650 (64.6)	423 (56.3)
Diabetes mellitus, n (%)	226 (22.4)	152 (20.2)
Dyslipidemia, n (%)	394 (39.2)	358 (47.6)
Chronic kidney disease, n (%)	345 (34.2)	197 (26.1)
Coronary artery disease, n (%)	129 (12.8)	132 (17.5)
History of disease		
Cerebral infarction, n (%)	96 (9.5)	77 (10.2)
Transient ischemic attack, n (%)	12 (1.2)	5 (0.7)
Systemic thromboembolism, n (%)	9 (0.9)	3 (0.4)
The $CHADS_2$ score (IQR)	2 (1–3)	1 (1–2)
The CHA_2DS_2-VASc score (IQR)	3 (2–4)	3 (2–4)
Antithrombotic medication		
Antiplatelet drug, n (%)	322 (32.0)	246 (32.7)

IQR: interquartile range; SD, standard deviation.

Table 2. Clinical outcomes.

	Derivation cohort (n = 1005)	Validation cohort (n = 752)
Cerebral infarction, n (%)	79 (7.9)	50 (6.6)
Transient ischemic attack, n (%)	4 (0.4)	4 (0.5)
Systemic thromboembolism, n (%)	7 (0.7)	1 (0.1)
Mortality, n (%)	209 (20.8)	110 (14.6)

patients (6.6%) in the validation cohort. The incidence rates of TIA and systemic thromboembolism during follow-up were lower than that of cerebral infarction at <1% in both cohorts. The mortality rate was 20.8% (209 patients) in the derivation cohort and 14.6% (110 patients) in the validation cohort.

Variable selection, model derivation, and internal validation

The 14 variables were extracted by a variable selection step (Figure 3(b)) (the important variables for each algorithm in the variable selection step are shown in Supplemental Figure I). There were 3 variables related to biological information (age, height, weight), 2 to

past medical and treatment history (hypertension, history of ablation), 2 to drug information (the intake of antiplatelet drug, the intake of antiarrhythmic drug), 2 to blood test results (creatinine clearance, blood urea nitrogen), and 4 to transthoracic echocardiogram data (mitral regurgitation, aortic regurgitation, relative wall thickness, left ventricular mass).

After training, the performance metrics of the ML model on the derivation cohort were as follows: sensitivity 82%, specificity 62%, accuracy 63%, F-1 score 0.26, Matthews correlation coefficient 0.24, AUROC 0.82, and AUPRC 0.40 (Figure 3(a)).

Comparison of the performance of each model in the validation cohort

The performance metrics on the validation cohort are shown in Table 3. Briefly, the machine learning model (ML model) had the best balance of sensitivity and specificity, and showed a higher accuracy, F-1 score, and higher Matthews correlation coefficient compared to the $CHADS_2$ and CHA_2DS_2-VASc scoring systems.

The AUROC values of each model were as follows: ML model, 0.72 (95%CI, 0.66–0.79); $CHADS_2$, 0.61 (95%CI, 0.52–0.69); CHA_2DS_2-VASc, 0.62 (95%CI, 0.54–0.70) (Figure 3(a)). According to the Hanley and McNeil method, the ML model was superior to the $CHADS_2$ (p = 0.01) and CHA_2DS_2-VASc (p = 0.02) scores.

Kaplan-Meier curves of the ML model, $CHADS_2$ score, and CHA_2DS_2-VASc score

Kaplan-Meier curves of the different risk subgroups classified by the ML model, the $CHADS_2$ score and the CHA_2DS_2-VASc score on the validation cohort are shown in Figure 4. By the ML model, the rates of cerebral infarction over a period of 3 years were 1.3% (95% CI, 0.0–3.0) for the low-risk group, 2.3% (95%CI, 0.6–3.9) for the moderate-risk group, and 6.9% (95%CI, 3.3–10.3) for the high-risk group. By the $CHADS_2$ score, the corresponding rates were 1.5% (95%CI, 0.0–3.4) for the low-risk group (score = 0), 2.8% (95%CI, 0.5–5.0) for the moderate-risk group (score = 1), and 4.5% (95%CI, 2.2–6.7) for the high-risk group (score > 2). And by the CHA_2DS_2-VASc score, the rates were 3.0% (95%CI, 0.0–7.0) for the low-risk group (score = 0), 1.2% (95%CI, 0.0–3.6) for the moderate-risk group (score = 1), and 3.8% (95%CI, 2.2–5.5) for the high-risk group (score > 2).

Discussion

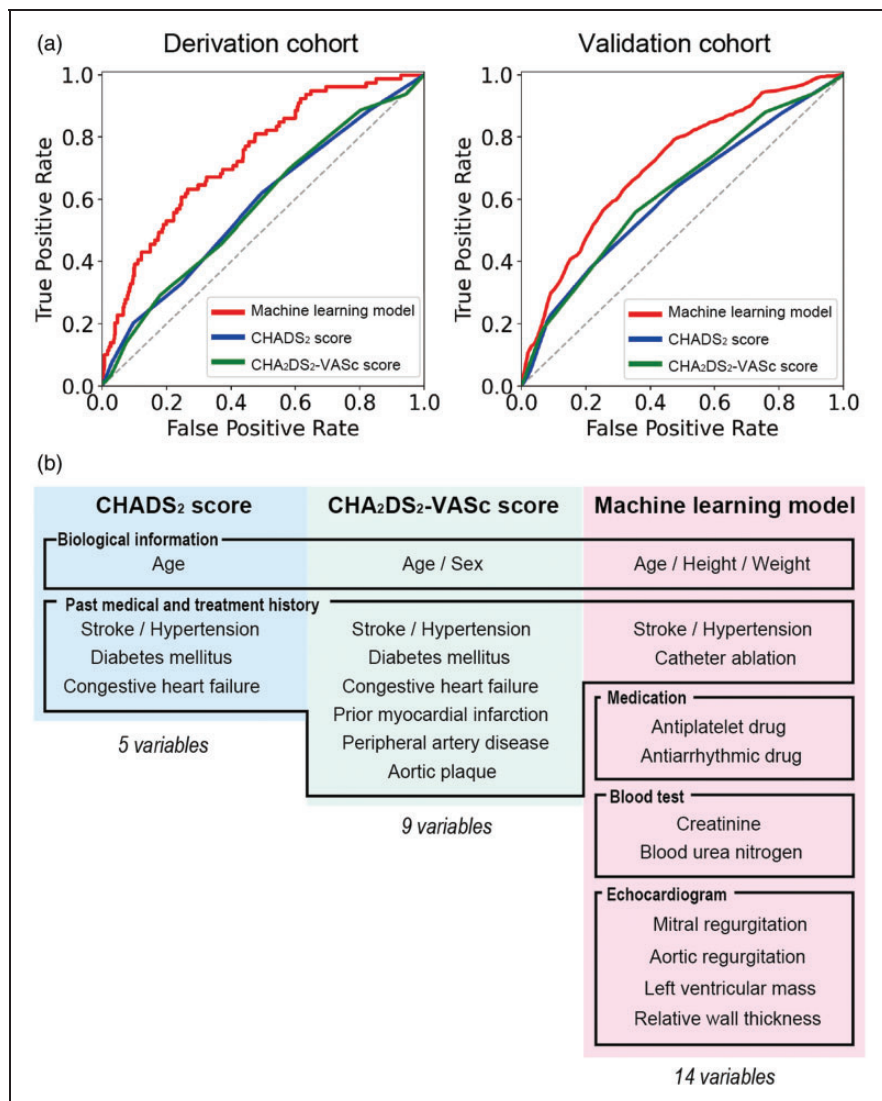In this study, we have shown that our model based on machine learning showed higher discriminative ability

Figure 3. (a) Receiver operating characteristic curves of the machine learning model and the CHADS$_2$ and CHA$_2$DS$_2$-VASc scoring systems in the derivation and validation cohort. (b) The 14 variables included in the machine learning model.

Table 3. Scores for each model on validation cohort.

Model	Sensitivity, %	Specificity, %	Accuracy, %	F-1 score	The Matthews correlation coefficient	AUROC	AUPRC
ML model	74 (59–87)	54 (48–59)	55 (50–60)	0.18 (0.13–0.23)	0.14 (0.06–0.20)	0.72 (0.66–0.79)	0.15 (0.09–0.23)
CHADS$_2$ score	52 (50–77)	52 (48–56)	53 (49–56)	0.15 (0.11–0.20)	0.08 (0.01–0.15)	0.61 (0.52–0.69)	0.10 (0.07–0.15)
CHA$_2$DS$_2$-VASc score	74 (61–86)	40 (37–44)	43 (39–46)	0.15 (0.10–0.19)	0.07 (0.01–0.14)	0.62 (0.54–0.70)	0.11 (0.07–0.16)

AUROC: area under the receiver operating curve; AUPRC: area under the precision-recall curve.

for ischemic events in non-valvular AF patients who were not receiving anticoagulation therapy than either the CHADS$_2$ or the CHA$_2$DS$_2$-VASc scoring system. A machine learning model incorporating only 14 variables (basic patient information, several blood sampling parameters, and echocardiographic data) has the potential to become a more appropriate risk stratification tool for patients with non-valvular AF.

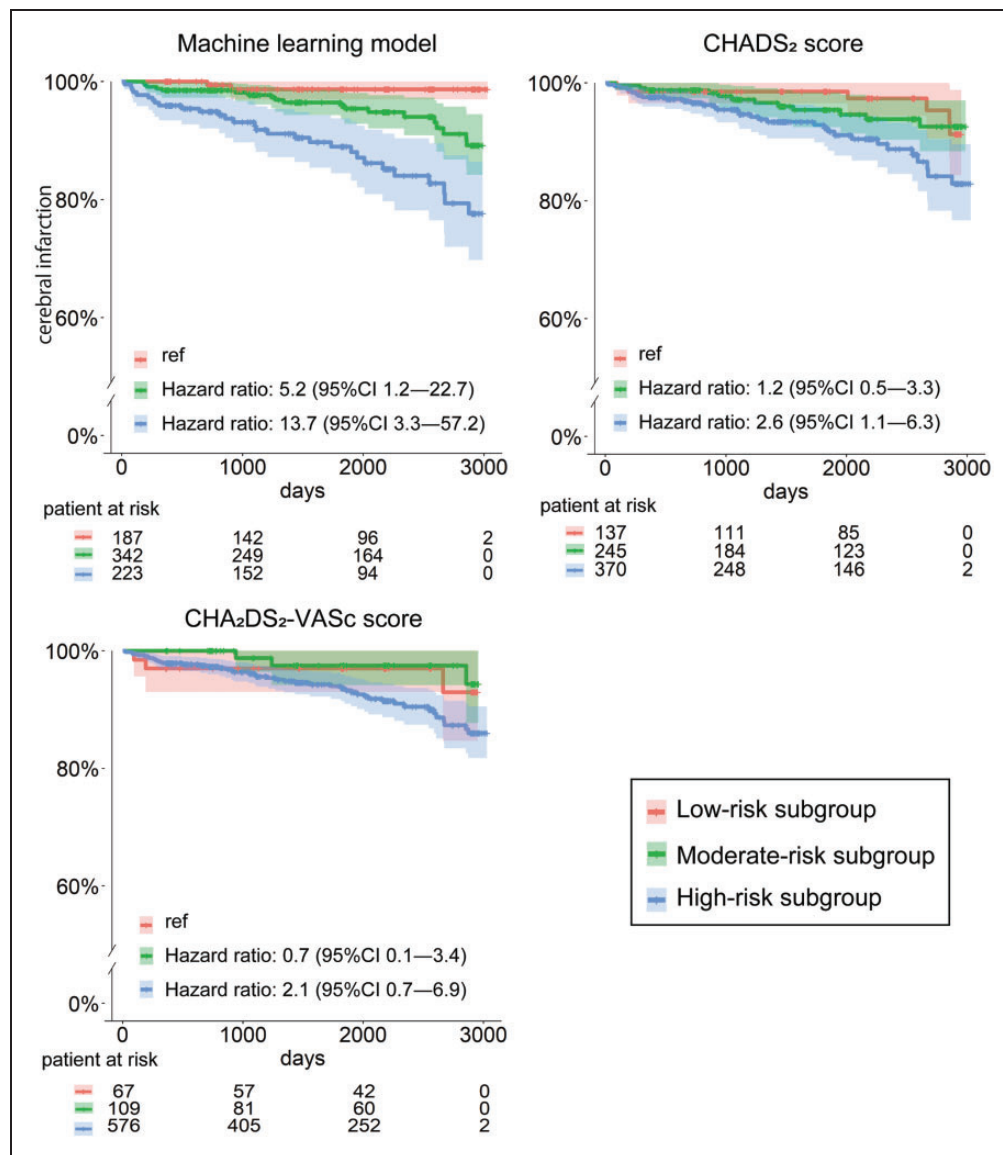The AUROC of the CHADS$_2$ score was within the range of the previous reports: 0.56—0.82.[2,3,14,17–19]

Figure 4. Kaplan-Meier curves of the machine learning model and the CHADS$_2$ and CHA$_2$DS$_2$-VASc scoring systems in the validation cohort. The hazard ratios of the moderate-risk and high-risk groups compared to the low-risk group are displayed for each measure.

The AUROC of CHA$_2$DS$_2$-VASc score was also within the range of the previous reports: 0.58—0.88.[3,14,17–22] In this study, The ML model had a significantly higher AUROC (0.72 vs 0.61–0.62), and a significantly higher AUPRC (0.15 vs 0.10–0.11) compared with the CHADS$_2$ and CHA$_2$DS$_2$-VASc scores. On the other hand, the AUROC of the ML model did not exceed 0.8, which is usually considered the cut-off for excellent performance.[23] In regard to the use of the validation cohort data to simulate clinical application, if we assume that anticoagulation therapy is indicated for patients with moderate or higher risk according to various guidelines, 565 patients in the ML model, 615 patients in the CHADS$_2$ model, and 685 patients in

the CHA$_2$DS$_2$-VASc model would be indicated for anticoagulation therapy (the definition of each risk category is described in the Methods section). Overall, 231 patients were correctly categorized for the ML model, 173 for the CHADS$_2$ score, and 109 for the CHA$_2$DS$_2$-VASc score. It is considered that the ML model may not increase the number of cerebral infarction cases for which anticoagulation is not indicated, and may reduce the overall administration of unnecessary anticoagulants (see also the Supplemental Table V).

Several studies have used machine learning to predict stroke incidence in patients with AF. Li et al. tested logistic regression, Naïve Bayes, and decision tree algorithms using the Chinese Atrial Fibrillation

Registry data. In their 1864 patients with AF who had not been treated with anticoagulation, the machine learning model showed a higher AUROC (0.71—0.74) than the Framingham score (0.65) or the CHA_2DS_2-VASc score (0.69).[24] They did not distinguish between non-valvular and valvular AF, and they did not statistically analyze the superiority or inferiority between models. Han et al. used data from 3185 patients with AF who had not been treated with anticoagulation and developed a machine learning model to predict future ischemic stroke.[25] All the patients were implanted with a cardiac implantable electronic device, and the pattern of the daily AF burden was found to be useful for predicting stroke (AUROC 0.66–0.70). They also did not distinguish between non-valvular and valvular AF, and did not statistically analyze the superiority or inferiority between models. These two studies were based on supervised machine learning. Meanwhile, Inohara et al. performed an unsupervised machine learning analysis in 9749 patients with AF.[26] Using hierarchical clustering, the entire population was divided into 4 distinct clusters, and each cluster displayed different clinical characteristics and risks for the major adverse cardiovascular or neurological events. However, they did not examine whether the clustering-based approach was superior to the existing risk scores. To the best of our knowledge, our present study is the first to show the superiority of a machine learning model over the standard prediction score by means of statistical test methods among the patients with non-valvular AF who were not receiving anticoagulation. Also, our model is unique and has potential for clinical application in that it can be calculated with as few as 14 variables that can be evaluated in daily clinical practice.

In our model, many previously known risk factors for ischemic events appeared as important variables: patient age, hypertension, renal insufficiency, and previous history of stroke are well-known risk factors and were included among the 14 variables in our model.[14,17,18,20,27] Although less frequently than these factors, body weight and some valvular heart diseases other than mitral stenosis have also been reported to be associated with stroke incidence in patients with AF.[28–31]

As for the information related to treatment, the intake of antiplatelet drugs, intake of antiarrhythmic drugs, and history of catheter ablation were all found to be important for the model. Since antiplatelet medication reduces the risk of stroke by 20—30% in AF patients, it is not surprising that the information on antiplatelet medication is an important factor.[32] On the other hand, rhythm control based on antiarrhythmic drugs or catheter ablation has not shown clear efficacy in stroke prevention, but some studies have suggested that rhythm control itself led to reduced stroke incidence. It has been speculated that the machine learning model uses the effect of rhythm control therapy for outcome prediction.[33]

With respect to imaging data, left ventricular relative wall thickness (LVRWT) and LV mass were important in the present model. We had previously reported the importance of left atrial enlargement and LVRWT as independent predictors of ischemic event in patients with AF.[34,35] Although there are few studies indicating a correlation between LVRWT and ischemic events, it is considered that LVRWT might reflect cardiac remodeling and might be related to the pathogenesis of AF and its complications. In this study, we found that echocardiogram-related information (aortic regurgitation, mitral regurgitation, LV mass, LVRWT) is important for predicting cerebral infarction, but on the other hand, the use of echocardiogram data compromises the simplicity of the score. When we constructed a machine learning model using only factors other than this echocardiogram-related information, the model performance decreased slightly to ROAUC 0.70 (95%CI 0.63-0.77), but still tended to be higher than the $CHADS_2$ and CHA_2DS_2-VASc scores (the model performance of the ML model without echocardiographic data is shown in the Supplemental Figure II and Table III).

There were several limitations in this study. First, we used a single community-based prospective registry database which we divided into a derivation and validation cohort, and ultimately, external validation in a completely separate population is desirable. Second, this study was based on a purely observational study of the patients with AF, and more than 50% of the patients were excluded because they were receiving anticoagulation (2355/4396; 53.5%). There were several differences in patient background between the anticoagulated and non-anticoagulated population; in particular, the $CHADS_2$, and CHA_2DS_2-VASc scores were significantly higher in the anticoagulation therapy group (see also Supplemental Table IV and Figure III). The population of the present study is likely to be different from the AF patient population, and likely to be biased. Third, it might be possible to improve the performance of the model using information that is not included in the database, such as the information of AF burden, left atrial function, left atrial appendage anatomy, and more detailed electrocardiographic markers, such as abnormal P-wave axis.[4] Also, we had to exclude some potentially useful variables due to the relatively high percentage of missing values, such as the cardiac biomarkers (troponin, brain natriuretic peptide/NT-pro brain natriuretic peptide). Finally, the model performance needs to be tested for its generalizability to various populations, such as patients

with different ethnicities and social backgrounds or undergoing different treatments.

Conclusions

The use of machine learning algorithms increased the accuracy of ischemic-event prediction in patients with non-valvular AF, compared to the $CHADS_2$ and CHA_2DS_2-VASc scoring systems, in our community-based registry. Along with the machine learning algorithm, the effective utilization of more comprehensive information may improve the prediction of cerebral infarction in patients with AF, and lead to more appropriate indications for anticoagulation therapy.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: M. Akao received lecture fees from Pfizer, Bristol-Myers Squibb, Boehringer Ingelheim, Bayer Healthcare, and Daiichi Sankyo. G.Y.H.L. has served as a consultant for Bayer/Janssen, Bristol-Myers Squibb/Pfizer, Medtronic, Boehringer Ingelheim, Novartis, Verseon,and Daiichi-Sankyo; Speaker for Bayer, Bristol-Myers Squibb/ Pfizer, Medtronic, Boehringer Ingelheim, and Daiichi-Sankyo. All other authors have reported that they have no relationships relevant to the contents of this article to disclose.

Authors' contributions

HN designed and conceptualized the study, analyzed the data, performed the statistical and machine learning analysis, and drafted the manuscript for intellectual content.

NO analyzed the data and revised the manuscript for intellectual content. HO had major role in the acquisition of data, interpreted the data, and revised the manuscript for intellectual content. NK, KD, OK, TA, and TT interpreted the data and revised the manuscript for intellectual content. SK analyzed the data, interpreted the data and revised the manuscript for intellectual content. MA designed and conceptualized study, had major role in the acquisition of data, interpreted the data, and revised the manuscript for intellectual content.

ORCID iDs

Hidehisa Nishi https://orcid.org/0000-0003-1763-2517
Naoya Oishi https://orcid.org/0000-0002-0778-3381

Supplemental material

Supplemental material for this article is available online.

References

1. Wolf PA, Abbott RD and Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham study. *Stroke* 1991; 22: 983–988.
2. Gage BF, Waterman AD, Shannon W, et al. Validation of clinical classification schemes for predicting stroke: results from the national registry of atrial fibrillation. *JAMA* 2001; 285: 2864–2870.
3. 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: 263–273.
4. Alkhouli M and Friedman PA. Ischemic stroke risk in patients with nonvalvular atrial fibrillation. *J Am Coll Cardiol* 2019; 74: 3050–3065.
5. Yaghi S and Kamel H. Stratifying stroke risk in atrial fibrillation. Beyond clinical risk scores. *Stroke* 2017; 48: 2665–2670.
6. Ogawa H, An Y, Ikeda S, et al.; on behalf of the Fushimi AF Registry Investigators. Progression from paroxysmal to sustained atrial fibrillation is associated with increased adverse events. *Stroke* 2018; 49: 2301–2308.
7. Nishi H, Oishi N, Ishii A, et al. Predicting clinical outcomes of large vessel occlusion before mechanical thrombectomy using machine learning. *Stroke* 2019; 50: 2379–2388.
8. Lee SI, Celik S, Logsdon BA, et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat Commun* 2018; 9: 42.
9. Akao M, Chun YH, Wada H, Fushimi AF Registry Investigators, et al. Current status of clinical background of patients with atrial fibrillation in a community-based survey: the Fushimi AF registry. *J Cardiol* 2013; 61: 260–266.
10. Aloisio MK, Swanson AS, Micali N, et al. Analysis of partially observed clustered data using generalized estimating equations and multiple imputation. *Stata J* 2014; 14: 863–883.
11. Steyerberg EW and Vergouwe Y. Towards better clinical prediction models: seven steps for development

and an ABCD for validation. *Eur Heart J* 2014; 35: 1925–1931.

12. Breiman L. Random forests. *Machine Learning* 2001; 45: 5–32.

13. Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. *ArXiv* 1706.09516.

14. Aakre CA, McLeod CJ, Cha SS, et al. Comparison of clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation. *Stroke* 2014; 45: 426–431.

15. Fuster V, Ryden LE, Cannom DS, et al. ACC/AHA/ESC 2006 guidelines for the management of patients with atrial fibrillation-executive summary: a report of the American College of Cardiology/American Heart Association task force on practice guidelines and the European Society of Cardiology Committee for Practice Guidelines (writing committee to revise the 2001 guidelines for the management of patients with atrial fibrillation). *Eur Heart J* 2006; 27: 1979–2030.

16. Hanley JA and McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983; 148: 839–843.

17. Piccini JP, Stevens SR, Chang Y, et al. Renal dysfunction as a predictor of stroke and systemic embolism in patients with nonvalvular atrial fibrillation: validation of the R(2) CHADS(2) index in the ROCKET AF (rivaroxaban once-daily, oral, direct factor Xa inhibition compared with vitamin K antagonism for prevention of stroke and embolism trial in atrial fibrillation) and ATRIA (AnTicoagulation and risk factors in Atrial Fibrillation) Study Cohorts. *Circulation* 2013; 127: 224–232.

18. Van den Ham HA, Olaf HK, Singer DE, et al. Comparative performance of ATRIA, CHADS₂, and CHA₂DS₂-VASc risk scores predicting stroke in patients with atrial fibrillation. Results from a national primary care database. *J Am Coll Cardiol* 2015; 66: 1851–1859.

19. Olesen JB, Lip GYH, Hansen ML, et al. Validation of risk stratification schemes for predicting stroke and thromboembolism in patients with atrial fibrillation: nationwide cohort study. *BMJ* 2011; 342: d124.

20. Singer DE, Chang Y, Borowsky LH, et al. A new risk scheme to predict ischemic stroke and other thromboembolism in atrial fibrillation: the ATRIA study stroke risk score. *J Am Heart Assoc* 2013; 2: e000250.

21. Chao TF, Lip GYH, Liu CJ, et al. Validation of a modified CHA2DS2-VASc score for stroke risk stratification in asian patients with atrial fibrillation: a nationwide cohort study. *Stroke* 2016; 47: 2462–2469.

22. Hijazi Z, Lindback J, Alexander JH, et al. The ABC (age, biomarkers, clinical history) stroke risk score: a biomarker-based risk score for predicting stroke in atrial fibrillation. *Eur Heart J* 2016; 37: 1582–1590.

23. D'Agostino RB, Pencina MJ, Massaro JM, et al. Cardiovascular disease risk assessment: insights from Framingham. *Glob Heart* 2013; 8: 11–23.

24. Li X, Li H, Du X, et al. Integrated machine learning approaches for predicting ischemic stroke and thromboembolism in atrial fibrillation. *AMIA Annu Symp Proc* 2017; 10: 799–807.

25. Han L, Askari M, Altman RB, et al. Atrial fibrillation burden signature and near-term prediction of stroke. A machine learning analysis. *Circ Cardiovasc Qual Outcomes* 2019; 12: e005595.

26. Inohara T, Shrader P, Pieper K, et al. Association of atrial fibrillation clinical phenotypes with treatment patterns and outcomes. A multicenter registry study. *JAMA Cardiol* 2018; 3: 54–63.

27. Go AS, Fang MC, Udaltsova N, et al. Impact of proteinuria and glomerular filtration rate on risk of thromboembolism in atrial fibrillation: the anticoagulation and risk factors in atrial fibrillation (ATRIA) study. *Circulation* 2009; 119: 1363–1369.

28. Wada Y, Mizushige K, Ohmori K, et al. Prevention of cerebral thromboembolism by low-dose anticoagulant therapy in atrial fibrillation with mitral regurgitation. *J Cardiovasc Pharmacol* 2001; 37: 422–426.

29. Bisson A, Bernard A, Bodin A, et al. Stroke and thromboembolism in patients with atrial fibrillation and mitral regurgitation. *Circ Arrhythm Electrophysiol* 2019; 12: e006990.

30. Hamatani Y, Ogawa H, Uozumi R, et al. Low body weight is associated with the incidence of stroke in atrial fibrillation patients – insight from the Fushimi AF registry. *Circ J* 2015; 79: 1009–1017.

31. Lee S-R, Choi E-K, Jung J-H, et al. Body mass index and clinical outcomes in Asian patients with atrial fibrillation receiving oral anticoagulation. *Stroke* 2021; 52: 521–530.

32. Aguilar M and Lart R. Antiplatelet therapy for preventing stroke in patients with non-valvular atrial fibrillation and no previous history of stroke or transient ischemic attacks. *Cochrane Database Syst Rev* 2005; 19: CD001925.

33. Corley SD, Epstein AE, DiMarco JP, et al.; AFFIRM Investigators. The AFFIRM investigators. Relationships between sinus rhythm, treatment, and survival in the atrial fibrillation follow-up investigation of rhythm management (AFFIRM) study. *Circulation* 2004; 109: 1509–1513.

34. Tezuka Y, Iguchi M, Hamatani Y, et al. Association of relative wall thickness of left ventricle with incidence of thromboembolism in patients with non-valvular atrial fibrillation: the Fushimi AF registry. *Eur Heart J Qual Care Clin Outcomes* 2020; 6: 273–283.

35. Hamatani Y, Ogawa H, Takabayashi K, et al. Left atrial enlargement is an independent predictor of stroke and systemic embolism in patients with non-valvular atrial fibrillation. *Sci Rep* 2016; 6: 31042.