TYPE Original Research
PUBLISHED 29 July 2022
DOI 10.3389/fcvm.2022.941237

Automated risk assessment of newly detected atrial fibrillation poststroke from electronic health record data using machine learning and natural language processing

Sheng-Feng Sung[1,2], Kuan-Lin Sung[3], Ru-Chiou Pan[4], Pei-Ju Lee[5]* and Ya-Han Hu[6]*

[1]Division of Neurology, Department of Internal Medicine, Ditmanson Medical Foundation Chiayi Christian Hospital, Chiayi City, Taiwan, [2]Department of Nursing, Min-Hwei Junior College of Health Care Management, Tainan, Taiwan, [3]School of Medicine, National Taiwan University, Taipei, Taiwan, [4]Clinical Data Center, Department of Medical Research, Ditmanson Medical Foundation Chiayi Christian Hospital, Chiayi City, Taiwan, [5]Department of Information Management and Institute of Healthcare Information Management, National Chung Cheng University, Chiayi County, Taiwan, [6]Department of Information Management, National Central University, Taoyuan, Taiwan

Background: Timely detection of atrial fibrillation (AF) after stroke is highly clinically relevant, aiding decisions on the optimal strategies for secondary prevention of stroke. In the context of limited medical resources, it is crucial to set the right priorities of extended heart rhythm monitoring by stratifying patients into different risk groups likely to have newly detected AF (NDAF). This study aimed to develop an electronic health record (EHR)-based machine learning model to assess the risk of NDAF in an early stage after stroke.

Methods: Linked data between a hospital stroke registry and a deidentified research-based database including EHRs and administrative claims data was used. Demographic features, physiological measurements, routine laboratory results, and clinical free text were extracted from EHRs. The extreme gradient boosting algorithm was used to build the prediction model. The prediction performance was evaluated by the C-index and was compared to that of the AS5F and CHASE-LESS scores.

Results: The study population consisted of a training set of 4,064 and a temporal test set of 1,492 patients. During a median follow-up of 10.2 months, the incidence rate of NDAF was 87.0 per 1,000 person-year in the test set. On the test set, the model based on both structured and unstructured data achieved a C-index of 0.840, which was significantly higher than those of the AS5F (0.779, $p = 0.023$) and CHASE-LESS (0.768, $p = 0.005$) scores.

Conclusions: It is feasible to build a machine learning model to assess the risk of NDAF based on EHR data available at the time of hospital admission.

OPEN ACCESS

EDITED BY
Qingpeng Zhang,
City University of Hong Kong,
Hong Kong SAR, China

REVIEWED BY
Pradyumna Agasthi,
Mayo Clinic, United States
Mei Liu,
University of Kansas Medical Center,
United States

*CORRESPONDENCE
Pei-Ju Lee
pjlee@mis.ccu.edu.tw
Ya-Han Hu
yhhu@mgt.ncu.edu.tw

SPECIALTY SECTION
This article was submitted to Heart Failure and Transplantation, a section of the journal Frontiers in Cardiovascular Medicine

RECEIVED 11 May 2022
ACCEPTED 11 July 2022
PUBLISHED 29 July 2022

CITATION
Sung S-F, Sung K-L, Pan R-C, Lee P-J and Hu Y-H (2022) Automated risk assessment of newly detected atrial fibrillation poststroke from electronic health record data using machine learning and natural language processing.
Front. Cardiovasc. Med. 9:941237.
doi: 10.3389/fcvm.2022.941237

Inclusion of information derived from clinical free text can significantly improve the model performance and may outperform risk scores developed using traditional statistical methods. Further studies are needed to assess the clinical usefulness of the prediction model.

KEYWORDS

atrial fibrillation, electronic health records, ischemic stroke, natural language processing, prediction

Introduction

Ischemic stroke is associated with a substantial risk of recurrence with a one-year recurrence rate ranging from 6 to 18% (1–4). The risk of stroke recurrence depends on the subtypes of ischemic stroke. As compared to other stroke subtypes, the recurrence rate of cardioembolic stroke is relatively high (5, 6). Moreover, cardioembolic strokes are often followed by strokes of the same type (6, 7). Atrial fibrillation (AF) is the most common cause of cardioembolic stroke, and even embolic stroke of undetermined source (ESUS) may originate from subclinical AF (8). As the population ages, AF-related strokes have increased and may triple in the next few decades (9, 10). Fortunately, the advancement of non-vitamin K antagonist oral anticoagulant therapy has made great progress in preventing patients with AF from cardioembolic stroke (8). Nonetheless, since AF can be paroxysmal, it may go undetected and therefore undiagnosed in patients undergoing routine electrocardiography (ECG) examinations. In fact, for ischemic stroke patients with undiagnosed AF, delayed use of oral anticoagulants may double the risk of recurrent stroke or transient ischemic attack (TIA) (11). Considering the impact of anticoagulant therapy on the outcome, poststroke screening for AF is thus critical for preventing recurrent stroke in patients with acute ischemic stroke (AIS).

Approximately 30% of all ischemic strokes are without any apparent cause (12). Among these cryptogenic strokes, nearly two-thirds are considered to stem from embolism (12). A study points out that through a series of heart rhythm monitoring, AF can be detected in up to 24% of patients with AIS or TIA (13). In addition to 24-h or even 72-h Holter monitoring (14), numerous studies have established that extended ECG monitoring *via* either implantable or external devices increases the yield of AF detection in patients with AIS (15, 16). However, given the limited medical resources, setting the right priorities of extended ECG monitoring by stratifying patients into different risk groups likely to have newly detected AF (NDAF) is more crucial than implementing population-level screening (17).

To date, more than twenty risk scores have been proposed to assess the risk of poststroke NDAF (18, 19). These risk scores vary in their complexity, target population, outcome definition, predictor variables, and ease of implementation. Most risk scores were derived or validated in patients with AIS while some of them were derived from a specific population with cryptogenic stroke or ESUS (20, 21). The simplest risk score consists of only two predictor variables, that is, age and stroke severity as assessed using the National Institutes of Health Stroke Scale (NIHSS) (22). Nevertheless, many of the risk scores require additional diagnostic work-up or interpretation of examination results to obtain the necessary predictors, such as markers of blood, ECG, echocardiography, as well as brain and vascular imaging (18). Routine use of such risk scores may be impractical in the context of the extra time and cost required.

On the other hand, with the ubiquitous use of electronic health records (EHRs) and the advancement in computational power, it has become feasible to use EHRs for the creation, validation, and implementation of data-driven risk prediction models (23, 24). For example, a previous study developed and validated an EHR-based prediction tool for 5-year AF risk in the general population (25), demonstrating a simple and cost-conscious approach to AF screening. Furthermore, in addition to structured numerical and categorical data, EHRs accommodate a multitude of unstructured textual data such as narrative clinical notes. Combining information extracted from clinical free text through natural language processing with structured data has shown promising results in improving the performance of risk prediction models (26–28).

AF-related strokes tend to be more severe and may manifest with different clinical features than other subtypes of ischemic strokes (29, 30). A higher risk of NDAF has been observed in patients with greater stroke severity (22, 31). Previous studies have shown that information extracted from clinical text can be used to represent patients' stroke severity (28, 32). Furthermore, stroke patients with AF have a higher prevalence of heart diseases and experience more cardiac events than those without AF (29–31). Symptoms, signs, or examinations related to heart diseases are typically documented in clinical notes. However, such information may not be captured or routinely collected as structured data in the EHR system. We thereby hypothesized that clinical text contains information that can discriminate between strokes stemming from AF and those not stemming from AF. In this study, we aimed to develop an EHR-based machine learning (ML) model to assess the risk of NDAF. To this end, we investigated various ML models using structured data, unstructured textual data, or a combination of both. In addition, the prediction performance of the developed ML models was compared to that of two traditional risk scores on a temporal test set of patients hospitalized for AIS.

Materials and methods

Data sources

The study data was obtained from the stroke registry of the Ditmanson Medical Foundation Chia-Yi Christian Hospital and the Ditmanson Research Database (DRD), a deidentified database comprising both EHR data and administrative claims data for research purposes. The DRD currently holds clinical information of over 1.4 million patients. The hospital stroke registry has prospectively enrolled consecutive hospitalized stroke patients since 2007 conforming to the design of the nationwide Taiwan Stroke Registry (33). To create the dataset for this study, we linked the stroke registry to the DRD using a unique encrypted patient identifier. Information regarding risk factors and stroke severity as assessed using the NIHSS was obtained from the stroke registry. Billing information and medical records from 2 years before to 1 year after the index stroke were extracted from the DRD.

The study protocol was approved by the Ditmanson Medical Foundation Chia-Yi Christian Hospital Institutional Review Board (IRB2020135). The requirement for informed consent was waived because of the retrospective design. The study protocol conforms to the ethical guidelines of the 1975 Declaration of Helsinki.

Study population

The study population selection is shown in Supplementary Figure 1. The stroke registry was queried for all hospitalizations for AIS between Oct 2007 and Sep 2020. Only the first hospitalization was included for each patient. Patients who suffered an in-hospital stroke or whose records could not be linked were excluded. The study population was split into a training set (patients admitted before the end of 2016) and a temporal test set (those admitted from 2017 onwards). All patients were traced in the DRD until AF was detected, death, the last visit within 1 year after the index stroke, or February 28, 2021, whichever came first.

Predictor and outcome variables

The class label (outcome) was AF, which was defined according to an AF ascertainment algorithm detailed in the Supplementary Methods in the Supplementary Material. According to the time sequence between AF detection and the index stroke (13), AF was further categorized as known AF before the index stroke, AF detected on admission, AF detected during the index stroke hospitalization, and AF detected after discharge (Figure 1). During the training phase, we trained ML models to predict which stroke is likely to stem from AF. Therefore, patients with all kinds of AF were retained in the training set. Because the study purpose was to build an ML model to assess the risk of NDAF poststroke, i.e., AF detected during the index stroke hospitalization and AF detected after discharge (Figure 1), patients who had known AF before the index stroke or AF detected on admission (34) were further excluded from the test set.

A total of 20 structured predictor variables (Supplementary Table 3), including age, sex, body mass index (BMI), vital signs, and results of routine blood tests, were chosen because they are readily available from EHRs upon admission. Missing values were imputed as mean values for continuous variables. Besides these structured variables, the free text extracted from the History of Present Illness section of the admission note was preprocessed through the following steps: spell checking, abbreviation expansion, removal of non-word symbols, removal of words suggestive of AF ("paroxysmal", "atrial", "fibrillation"), lowercase conversion, lemmatization, marking of negated words with the suffix "_NEG" using the Natural Language Toolkit mark_negation function, and stop-word removal.

The preprocessed text was then vectorized using the bag-of-words (BOW) approach with three different types of feature representation (Figure 2). We built a document-term matrix in which each column represents each unique feature (word) from the text corpus while the rows represent each document (present illness for each patient). The cells represent the counts of each word within each document (term frequency), the absence or presence of each word within each document (binary representation), or the term frequency with inverse document frequency (TF-IDF) weighting (35). Because medical terms are commonly comprised of two words or even more, we further experimented with adding word bigram features (two-word phrases) to the basic BOW model. To reduce noises such as redundant and less informative features as well as to improve training efficiency (36), we performed feature selection by filtering out words that appeared in <5% of all documents in the training set, followed by performing a penalized logistic regression with 10-fold cross-validation to identify the most predictive words (37).

Baseline models

For comparison with ML models, we only considered traditional risk scores that are based on variables available from

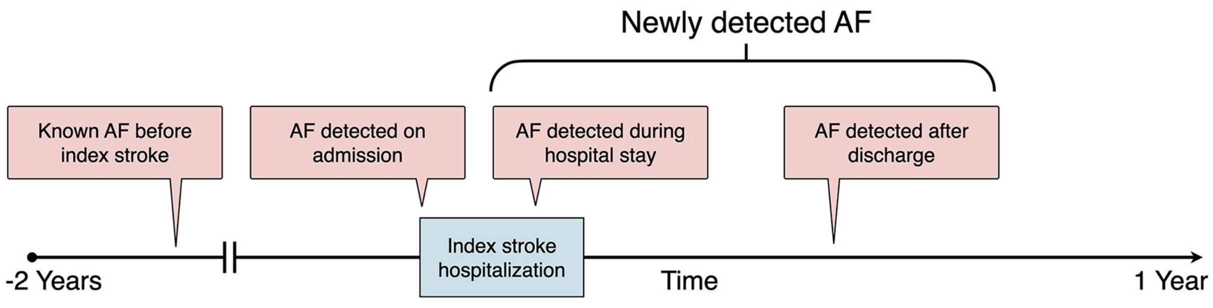

Newly detected AF
Known AF before index stroke
AF detected on admission
AF detected during hospital stay
AF detected after discharge
Index stroke hospitalization
-2 Years
Time
1 Year

FIGURE 1
Definition of AF categories according to the time sequence between AF detection and the index stroke. AF, atrial fibrillation.

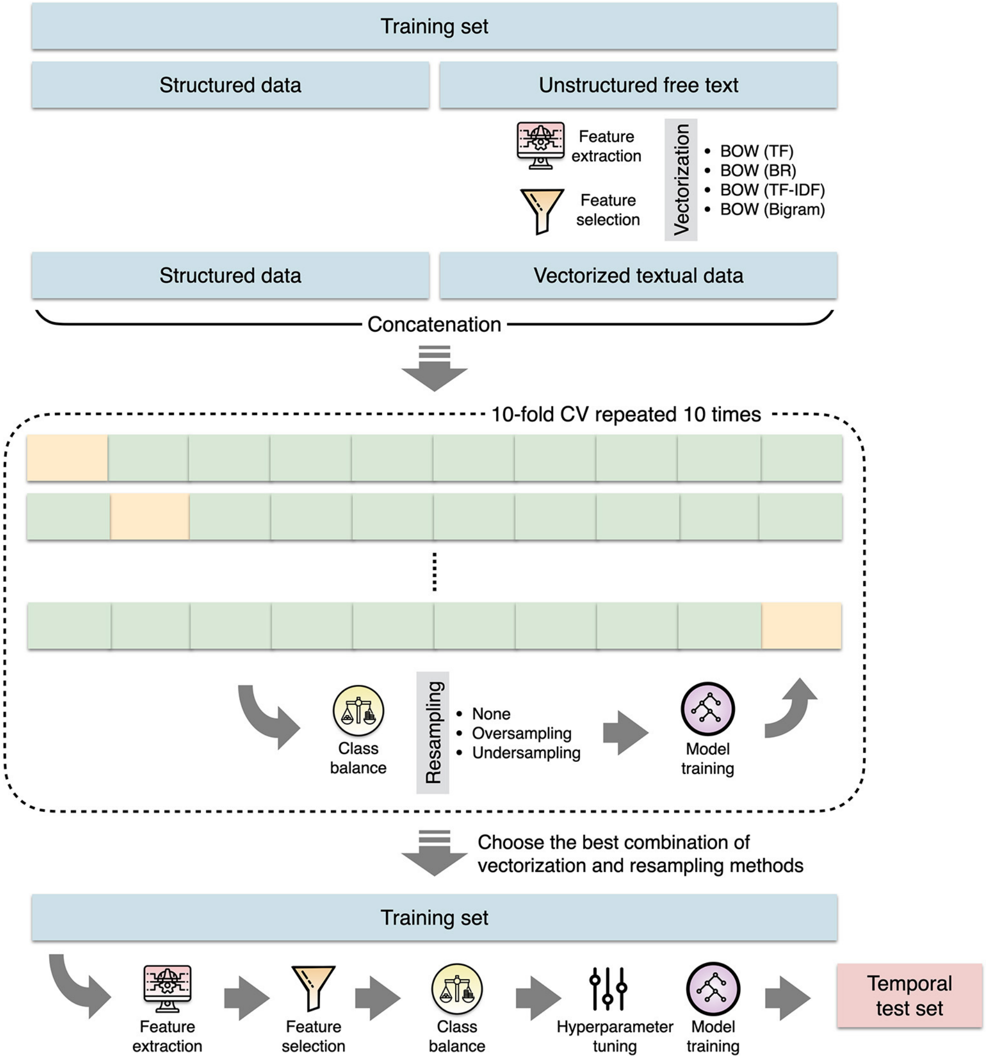


FIGURE 2
The process of machine learning model construction. BOW, bag-of-words; BR, binary representation; CV, cross validation; TF, term frequency; TF-IDF, term frequency with inverse document frequency.

EHRs upon admission. According to a validation study that evaluated eight such risk scores, two risk scores performed better than the others, demonstrating adequate discrimination and calibration (19). These two risk scores were thus used as the baseline models. The AS5F score, composed by age and NIHSS, was developed and validated in cohorts of patients who underwent extended Holter monitoring after AIS or TIA (22). The CHASE-LESS score was constructed from patients hospitalized for AIS in a claims database (31). It comprises seven components, including age, NIHSS, as well as the presence of coronary artery disease, congestive heart failure, hyperlipidemia, diabetes, and prior stroke or TIA.

Machine learning models

ML models were constructed by using structured data, vectorized textural data, or a combination of both (Figure 2). Because class imbalance might influence the classification performance, we experimented with resampling methods to maintain the ratio of majority and minority classes as 1:1, 2:1, or 3:1 (38). The extreme gradient boosting (XGB) algorithm was used to build classifiers. The XGB classifier trains a series of classification and regression trees where each successive tree attempts to correct the errors of the preceding trees.

During the training process, we first evaluated a suite of different combinations of text vectorization techniques and resampling methods without hyperparameter tuning. We repeated 10-fold cross-validation 10 times to obtain the performance estimates. The area under the receiver operating characteristic curve (AUC) was used as the evaluation metric because both positive and negative classes are important. After the optimal combination of text vectorization and resampling methods was determined, ML models were trained from the full training set through feature extraction, feature selection, class balancing, followed by hyperparameter tuning. Hyperparameter optimization for each model was performed by repeating 10-fold cross-validation 10 times. Model error was minimized in terms of AUC. We performed a grid search to find optimal hyperparameters following steps proposed in a prior study (39). After building the XGB classifiers, we used Shapley additive explanations (40) to interpret the output of the XGB classifiers. The experiments were carried out by using scikit-learn, XGBoost, imbalanced-learn, and SHAP libraries within Python 3.7 environment.

Statistical analysis

Categorical variables were reported with counts and percentages. Continuous variables were presented as means with standard deviations or medians and interquartile ranges. Differences between groups were tested by Chi-square tests for categorical variables and *t* tests or Mann-Whitney U tests for continuous variables, as appropriate.

The incidence rate of NDAF was expressed as events per 1,000 person-years. To assess the prediction performance of each prediction model, Cox proportional hazard regression analyses were performed by entering each risk score or the predicted probability output by each ML model as a continuous variable. Harrell's concordance index (C-index) was calculated to evaluate and compare model performance. The C-index ranges from 0.5 to 1.0, with 0.5 indicating random guess and 1 indicating perfect model discrimination. A model with a C-index value above 0.7 is considered acceptable for clinical use (41).

All statistical analyses were performed using Stata 15.1 (StataCorp, College Station, Texas) and R version 4.1.1 (R Foundation for Statistical Computing, Vienna, Austria). Two-tailed *p* values were considered statistically significant at <0.05.

Results

Characteristics of the study population

A total of 6,321 patients were eligible for this study (Supplementary Figure 1). The training set consisted of 4,604 patients who were admitted before the end of 2016. Among patients in the training set, 422 (9.2%) had known AF, 265 (5.6%) were diagnosed with AF on admission, and 232 (5.0%) developed NDAF during follow-up. Among 1,717 patients who were admitted from 2017 onwards, 122 and 103 were excluded because of having known AF before the index stroke and being diagnosed with AF on admission, respectively. Therefore, the temporal test set consisted of 1,492 patients. During a median follow-up of 10.2 months, 87 (5.8%) patients in the temporal test set were identified as having NDAF. Each patient had an average of 3.1 hospital visits per month during the follow-up period. The incidence rate of NDAF was 87.0 per 1,000 person-year. Table 1 lists the characteristics of the patients. Patients in the training set were older, more likely to be female, less likely to have diabetes mellitus, and tended to have hypertension, coronary artery disease, congestive heart failure, as well as prior stroke or TIA. They also had significantly higher NIHSS, AS5F, and CHASE-LESS scores.

Performance of prediction models

According to the estimates of AUC obtained from the 10 times of 10-fold cross-validation (Figure 3), ML models using a combination of both structured and unstructured data achieved higher AUCs than those using structured or unstructured data alone. Data resampling did not improve the performance of models. Text vectorization using BOW with TF-IDF weighting generally performed higher than the other

TABLE 1 Characteristics of the study population.

Characteristic	Training set ($N = 4{,}604$)	Temporal test set ($N = 1{,}492$)	*P*
Age, mean (SD)	69.2 (12.3)	68.0 (13.5)	0.002
Female	1,896 (41.2)	531 (35.6)	<0.001
Hypertension	3,705 (80.5)	1,119 (75.0)	<0.001
Diabetes mellitus	1,958 (42.5)	683 (45.8)	0.028
Hyperlipidemia	2,670 (58.0)	852 (57.1)	0.546
Coronary artery disease	560 (12.2)	103 (6.9)	<0.001
Congestive heart failure	228 (5.0)	25 (1.7)	<0.001
Prior stroke or TIA	1,143 (24.8)	274 (18.4)	<0.001
NIHSS, median (IQR)	5 (3-10)	5 (2-8)	<0.001
AS5F, median (IQR)	67.4 (59.2–76.5)	65.8 (56.9–74.2)	<0.001
CHASE-LESS, median (IQR)	6 (5-8)	6 (4-7)	<0.001

Data are numbers (percentage) unless specified otherwise.
IQR, interquartile range; NIHSS, National Institutes of Health Stroke Scale; SD, standard deviation; TIA, transient ischemic attack.

feature value representation methods. Therefore, we used the full original training set to build three ML models, that is, a model based on structured data (model A), a model based on textual data vectorized using BOW with TF-IDF weighting (model B), and a model based on both structured data and unstructured textual data vectorized using BOW with TF-IDF weighting (model C).

Table 2 lists the performance of prediction models. All the prediction models significantly predicted the risk of NDAF. Among the ML models, model C had the highest C-index (0.840), which was significantly higher than those of model A (0.791, $p = 0.009$) and model B (0.738, $p < 0.001$). Model C outperformed the AS5F (0.779, $p = 0.023$) and CHASE-LESS (0.768, $p = 0.005$) scores. The C-index of model A was comparable to those of AS5F ($p = 0.715$) and CHASE-LESS ($p = 0.487$) scores. Although model B attained the lowest C-index, its performance was also comparable to the AS5F ($p = 0.163$) and CHASE-LESS ($p = 0.282$) scores.

Model interpretation

Figure 4A shows the top 20 most important features in model C ordered by the mean absolute Shapley value, which indicates the global importance of each feature on the model output. Figure 4B presents the beeswarm plot depicting the Shapley value for every patient across these features, demonstrating each feature's contribution to the model output. According to the magnitude and direction of the Shapley value, patients who were female and those with increased age, high heart rate, elevated creatinine, elevated blood urea nitrogen, and high BMI were more likely to have NDAF. Patients with high triglyceride, platelet count, and pulse pressure were less likely to have NDAF. Words associated with an increased risk of NDAF included “unit”, “middle”, “cardiovascular”, “heart”, “electrocardiogram”, and “family”, whereas those associated with a decreased risk were “numbness”, “diabetes”, “day”, “visit”, and “ago”.

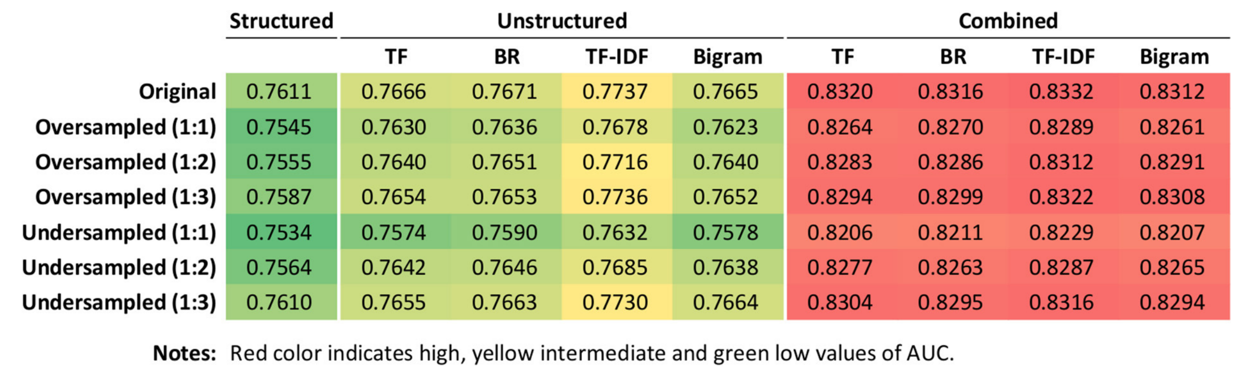
The top 20 most important features in model A and model B are shown in Supplementary Figures 2, 3, respectively. The important structured and unstructured predictors identified in model C were generally consistent with those identified separately in model A (structured data) and model B (unstructured textual data).

Discussion

We found that prediction of NDAF using routinely collected variables from EHRs was feasible. ML models performed better than or were comparable to existing risk scores. The ML model based on both structured variables and text had higher discriminability than those of AS5F and CHASE-LESS scores. Furthermore, by using the Shapley value to reveal the significance of features, we identified important predictors of NDAF that may help gain insight into clinical practice for stroke prevention.

Important predictors of newly detected atrial fibrillation

Many studies have investigated prediction models for NDAF in the general population (25, 42, 43) or in selected patient groups such as those with stroke or TIA (18, 21, 22, 31, 44, 45). Owing to the different characteristics of at-risk populations, it is arguable whether the relationships between the predictors and



	Structured	Unstructured				Combined			
		TF	BR	TF-IDF	Bigram	TF	BR	TF-IDF	Bigram
Original	0.7611	0.7666	0.7671	0.7737	0.7665	0.8320	0.8316	0.8332	0.8312
Oversampled (1:1)	0.7545	0.7630	0.7636	0.7678	0.7623	0.8264	0.8270	0.8289	0.8261
Oversampled (1:2)	0.7555	0.7640	0.7651	0.7716	0.7640	0.8283	0.8286	0.8312	0.8291
Oversampled (1:3)	0.7587	0.7654	0.7653	0.7736	0.7652	0.8294	0.8299	0.8322	0.8308
Undersampled (1:1)	0.7534	0.7574	0.7590	0.7632	0.7578	0.8206	0.8211	0.8229	0.8207
Undersampled (1:2)	0.7564	0.7642	0.7646	0.7685	0.7638	0.8277	0.8263	0.8287	0.8265
Undersampled (1:3)	0.7610	0.7655	0.7663	0.7730	0.7664	0.8304	0.8295	0.8316	0.8294

Notes: Red color indicates high, yellow intermediate and green low values of AUC.

FIGURE 3
Heat map showing AUC values across machine learning models with different combinations of text vectorization techniques and resampling methods. AUC, area under the receiver operating characteristic curve; BR, binary representation; TF, term frequency; TF-IDF, term frequency with inverse document frequency.

TABLE 2 Performance of prediction models for predicting newly detected atrial fibrillation.

Risk score	HR (95% CI)	*P*	Schoenfeld's global test	C-index (95% CI)
AS5F	1.10 (1.08–1.13)	<0.001	0.062	0.779 (0.734–0.825)
CHASE-LESS	1.49 (1.38–1.60)	<0.001	0.296	0.768 (0.721–0.816)
Model A (structured)	1.05 (1.04–1.06)	<0.001	0.764	0.791 (0.745–0.836)
Model B (unstructured)	1.04 (1.03–1.05)	<0.001	0.060	0.738 (0.688–0.788)
Model C (combined)	1.05 (1.04–1.06)	<0.001	0.600	0.840 (0.803–0.876)

CI, confidence interval; HR, hazard ratio.

NDAF are similar across patient groups. Among the identified structured predictor variables, some of them such as age and BMI were common to the general population and patients with stroke (18, 42), others are known predictors in the general population but have seldom been used to predict poststroke NDAF, while still others are controversial predictors that warrant further study. For example, chronic kidney disease is a positive predictor whereas hyperlipidemia is a negative predictor of NDAF in the general population (25, 43). This study echoes those findings by showing positive associations of NDAF with elevated creatinine, elevated blood urea nitrogen, as well as decreased triglyceride level (Figure 4B). On the other hand, the evidence on the relationship between heart rate and NDAF is conflicting (46).

The central hypothesis of this study is that clinical free text contains information that may be used to predict NDAF. We indeed identified several words that could help make predictions. The reason why some of these words were associated with the risk of NDAF may be obscure at first glance but could be revealed by examining each word in its context. For example, the word "unit" from the term "intensive care unit" and the word "middle" from the term "middle cerebral artery infarction" typically imply severe stroke, which is a known predictor of NDAF (22, 31). These results demonstrate that useful and informative predictors could be derived from unstructured text in EHRs without intervening human curation. Despite this, since clinicians may use different terms to describe the same condition in clinical text, the relationship between such terms might not be accurately represented. Concept-based feature extraction using specialized medical ontologies can be explored in future research (35).

Advantages of EHR-based machine learning models

Traditional prediction models used in clinical practice are generally built on limited predefined variables using logistic regression. Although such models have reasonable prediction performance, whether they are applicable in routine clinical practice and relevant to a specific context is yet to be determined (47). First, logistic regression models necessitate the assumptions of linear and additive relationships among predictors being fulfilled, while ML algorithms, especially tree-based models, are more effective in capturing potential nonlinear relationships and handling complex interactions between the predictor and outcome variables (48). Second,

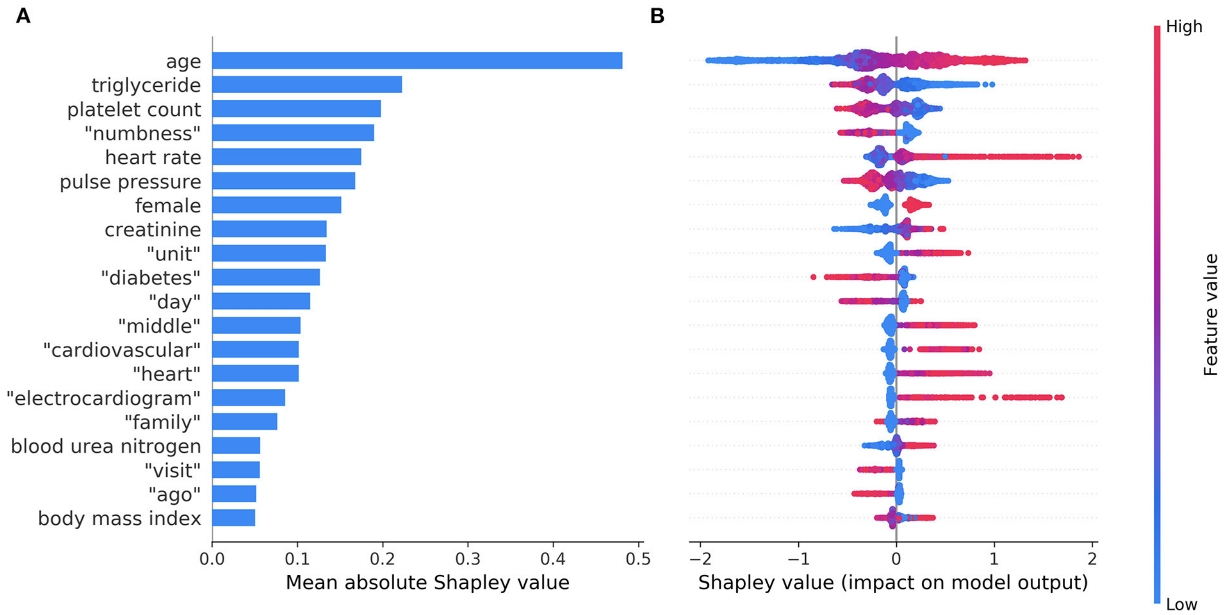


FIGURE 4
The top 20 most important features identified by the model based on both structured data and unstructured textual data. The mean absolute Shapley values that indicate the average impact on model output are shown in a bar chart **(A)**. The individual Shapley values for these features for each patient are depicted in a beeswarm plot **(B)**, where a dot's position on the x-axis denotes each feature's contribution to the model prediction for that patient. The color of the dot specifies the relative value of the corresponding feature.

considering the wide variety of data in EHRs, data-driven prediction modeling may allow identifying novel predictors in the context of insufficient prior knowledge of the real system (49). In this respect, ML is suitable for building complex models and analyzing noisy data such as that stored in EHRs (50). ML techniques were also applied to predict cardioembolic vs. non-cardioembolic stroke mechanism in patients with ESUS (51). Recently, deep learning techniques have been introduced to predict new-onset AF in the general population using structured primary care data or unstructured 12-lead ECG traces (52, 53).

Clinical applications and significance

Poststroke AF screening is essential for choosing the optimal strategy for secondary stroke prevention. However, to be resource efficient, extended ECG monitoring should be prioritized for patients at a high risk of NDAF. The developed ML model will be suited for assessing the risk of individual patients and assisting in personalized clinical decisions. Moreover, locally constructed prediction models may be more suitable for real-world clinical use than externally developed risk models (25). Since the prediction model was derived from EHRs, it is ideal to implement this model in the EHR as a decision support tool. With this tool, the calculation of risk estimates and the flagging of high-risk patients can be automated within the EHR, streamlining the process of risk stratification for poststroke AF screening.

Limitations

This study has several limitations. First, patients were traced through EHRs. Because patients might be diagnosed with AF outside the study hospital, some outcome misclassification was inevitable. Nevertheless, the frequent visits to the study hospital observed in this stroke population (>3 visits per month) might have alleviated this problem. Second, the diagnosis of AF was made in usual-care settings, where AF was detected almost exclusively by 12-lead ECG or 24-h Holter ECG. Advanced ECG monitoring *via* either implantable or external devices to detect subclinical or low-burden AF was not used. Consequently, the study findings are valid for relatively high-burden AF (54). Third, although data-driven ML modeling has its own advantages, the predictor-outcome relationships discovered from data does not mean causality. In other words, prediction accuracy should not be equated to causal validity (55). Fourth, as this is a single-site study, the generalizability of the study findings may be restricted. Variations in the terminology used in clinical documentation are to be expected across healthcare settings. However, the methods used here may allow other healthcare systems to develop their own customized versions of prediction models.

Conclusions

It is feasible to build an ML model to predict NDAF based on EHR data available at the time of hospital admission. Inclusion of information derived from clinical free text can significantly improve the model performance and may outperform risk scores developed using traditional statistical methods. These improvements may be due to both the modeling approach to delineate nonlinear decision boundaries and the use of textual features that help characterize nuances of disease presentation across patients. Despite these findings, further studies are required to confirm the approach's generalizability and the clinical usefulness of the prediction model.

Data availability statement

The data used in this study cannot be made available because of restrictions regarding the use of EMR data. Requests to access these datasets should be directed to Y-HH, yhhu@mgt.ncu.edu.tw.

Ethics statement

The studies involving human participants were reviewed and approved by Ditmanson Medical Foundation Chia-Yi Christian Hospital Institutional Review Board. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

Study concept and design: S-FS and Y-HH. Acquisition of data: S-FS and R-CP. Drafting of the manuscript: S-FS and K-LS. Study supervision: P-JL and Y-HH. Analysis and interpretation of data and critical revision of the manuscript for important intellectual content: all authors. All authors had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Funding

This research was supported in part by the Ditmanson Medical Foundation Chia-Yi Christian Hospital-National Chung Cheng University Joint Research Program (Grant number CYCH-CCU-2022-03).

Acknowledgments

The authors thank the help from the Clinical Data Center, Ditmanson Medical Foundation Chia-Yi Christian Hospital for providing administrative and technical support. This study is based in part on data from the Ditmanson Research Database (DRD) provided by Ditmanson Medical Foundation Chia-Yi Christian Hospital. The interpretation and conclusions contained herein do not represent the position of Ditmanson Medical Foundation Chia-Yi Christian Hospital. The authors also thank Ms. Li-Ying Sung for English language editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2022.941237/full#supplementary-material

References

1. Wang Y, Xu J, Zhao X, Wang D, Wang C, Liu L, et al. Association of hypertension with stroke recurrence depends on ischemic stroke subtype. *Stroke.* (2013) 44:1232–7. doi: 10.1161/strokeaha.111.000302

2. Kang K, Park TH, Kim N, Jang MU, Park S-S, Park J-M, et al. Recurrent stroke, myocardial infarction, and major vascular events during the first year after acute ischemic stroke: the multicenter prospective observational study about recurrence and its determinants after acute ischemic stroke I. *J Stroke Cerebrovasc Dis.* (2016) 25:656–64. doi: 10.1016/j.jstrokecerebrovasdis.2015.11.036

3. Hsieh C-Y, Wu DP, Sung S-F. Trends in vascular risk factors, stroke performance measures, and outcomes in patients with first-ever ischemic stroke in Taiwan between 2000 and 2012. *J Neurol Sci.* (2017) 378:80–4. doi: 10.1016/j.jns.2017.05.002

4. Lin B, Zhang Z, Mei Y, Wang C, Xu H, Liu L, et al. Cumulative risk of stroke recurrence over the last 10 years: a systematic review and meta-analysis. *Neurol Sci.* (2021) 42:61–71. doi: 10.1007/s10072-020-04797-5

5. Rücker V, Heuschmann PU, O'Flaherty M, Weingärtner M, Hess M, Sedlak C, et al. Twenty-year time trends in long-term case-fatality and recurrence

rates after ischemic stroke stratified by etiology. *Stroke.* (2020) 51:2778–85. doi: 10.1161/strokeaha.120.029972

6. Kolmos M, Christoffersen L, Kruuse C. Recurrent ischemic stroke – a systematic review and meta-analysis. *J Stroke Cerebrovasc Dis.* (2021) 30:105935. doi: 10.1016/j.jstrokecerebrovasdis.2021.105935

7. Flach C, Muruet W, Wolfe CDA, Bhalla A, Douiri A. Risk and secondary prevention of stroke recurrence. *Stroke.* (2020) 51:2435–44. doi: 10.1161/strokeaha.120.028992

8. Kamel H, Healey JS. Cardioembolic stroke. *Circ Res.* (2017) 120:514–26. doi: 10.1161/circresaha.116.308407

9. Go AS, Hylek EM, Phillips KA, Chang Y, Henault LE, Selby JV, et al. Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the anticoagulation and risk factors in atrial fibrillation (ATRIA) study. *JAMA.* (2001) 285:2370–5. doi: 10.1001/jama.285.18.2370

10. Yiin GSC, Howard DPJ, Paul NLM Li L, Luengo-Fernandez R, Bull LM, et al. Age-specific incidence, outcome, cost, and projected future burden of atrial fibrillation–related embolic vascular events. *Circulation.* (2014) 130:1236–44. doi: 10.1161/circulationaha.114.010942

11. Lip GYH, Hunter TD, Quiroz ME, Ziegler PD, Turakhia MP. Atrial fibrillation diagnosis timing, ambulatory ecg monitoring utilization, and risk of recurrent stroke. *Circ Cardiovasc Qual Outcomes.* (2017) 10:e002864. doi: 10.1161/circoutcomes.116.002864

12. Yaghi S, Bernstein RA, Passman R, Okin PM, Furie KL. Cryptogenic stroke. *Circ Res.* (2017) 120:527–40. doi: 10.1161/circresaha.116.308447

13. Sposato LA, Cipriano LE, Saposnik G, Vargas ER, Riccio PM, Hachinski V. Diagnosis of atrial fibrillation after stroke and transient ischaemic attack: a systematic review and meta-analysis. *Lancet Neurol.* (2015) 14:377–87. doi: 10.1016/s1474-4422(15)70027-x

14. Grond M, Jauss M, Hamann G, Stark E, Veltkamp R, Nabavi D, et al. Improved detection of silent atrial fibrillation using 72-hour holter ecg in patients with ischemic stroke. *Stroke.* (2013) 44:3357–64. doi: 10.1161/strokeaha.113.001884

15. Buck BH, Hill MD, Quinn FR, Butcher KS, Menon BK, Gulamhusein S, et al. Effect of implantable vs prolonged external electrocardiographic monitoring on atrial fibrillation detection in patients with ischemic stroke. *JAMA.* (2021) 325:2160–8. doi: 10.1001/jama.2021.6128

16. Noubiap JJ, Agbaedeng TA, Kamtchum-Tatuene J, Fitzgerald JL, Middeldorp ME, Kleinig T, et al. Rhythm monitoring strategies for atrial fibrillation detection in patients with cryptogenic stroke: A systematic review and meta-analysis. *Int J Cardiol Hear Vasc.* (2021) 34:100780. doi: 10.1016/j.ijcha.2021.100780

17. Jones NR, Taylor CJ, Hobbs FDR, Bowman L, Casadei B. Screening for atrial fibrillation: a call for evidence. *Eur Heart J.* (2019) 41:1075–85. doi: 10.1093/eurheartj/ehz834

18. Kishore AK, Hossain MJ, Cameron A, Dawson J, Vail A, Smith CJ. Use of risk scores for predicting new atrial fibrillation after ischemic stroke or transient ischemic attack—a systematic review. *Int J Stroke.* (2021) 174749302110458. doi: 10.1177/17474930211045880

19. Hsieh C-Y, Kao H-M, Sung K-L, Sposato LA, Sung S-F, Lin S-J. Validation of risk scores for predicting atrial fibrillation detected after stroke based on an electronic medical record algorithm: a registry-claims-electronic medical record linked data study. *Front Cardiovasc Med.* (2022) 9:888240. doi: 10.3389/fcvm.2022.888240

20. Ntaios G, Perlepe K, Lambrou D, Sirimarco G, Strambo D, Eskandari A, et al. Identification of patients with embolic stroke of undetermined source and low risk of new incident atrial fibrillation: The AF-ESUS score. *Int J Stroke.* (2020) 16:29–38. doi: 10.1177/1747493020925281

21. Muscari A, Barone P, Faccioli L, Ghinelli M, Trossello MP, Puddu GM, et al. Usefulness of the ACTEL score to predict atrial fibrillation in patients with cryptogenic stroke. *Cardiology.* (2020) 145:168–77. doi: 10.1159/000505262

22. Uphaus T, Weber-Krüger M, Grond M, Toenges G, Jahn-Eimermacher A, Jauss M, et al. Development and validation of a score to detect paroxysmal atrial fibrillation after stroke. *Neurology.* (2019) 92:e115–24. doi: 10.1212/wnl.0000000000006727

23. Healey JS, Wong JA. Pre-screening for atrial fibrillation using the electronic health record. *JACC Clin Electrophysiol.* (2019) 5:1342–3. doi: 10.1016/j.jacep.2019.08.019

24. Ding L, Liu C, Li Z, Wang Y. Incorporating artificial intelligence into stroke care and research. *Stroke.* (2020) 51:e351–4. doi: 10.1161/strokeaha.120.031295

25. Hulme OL, Khurshid S, Weng L-C, Anderson CD, Wang EY, Ashburner JM, et al. Development and validation of a prediction model for atrial fibrillation using electronic health records. *JACC Clin Electrophysiol.* (2019) 5:1331–41. doi: 10.1016/j.jacep.2019.07.016

26. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS ONE.* (2017) 12:e0174708. doi: 10.1371/journal.pone.0174708

27. Weissman GE, Hubbard RA, Ungar LH, Harhay MO, Greene CS, Himes BE, et al. Inclusion of unstructured clinical text improves early prediction of death or prolonged ICU stay. *Crit Care Med.* (2018) 46:1125–32. doi: 10.1097/ccm.0000000000003148

28. Sung S, Chen C, Pan R, Hu Y, Jeng J. Natural language processing enhances prediction of functional outcome after acute ischemic stroke. *J Am Heart Assoc.* (2021) 10:e023486. doi: 10.1161/jaha.121.023486

29. Marini C, Santis FD, Sacco S, Russo T, Olivieri L, Totaro R, et al. Contribution of atrial fibrillation to incidence and outcome of ischemic stroke. *Stroke.* (2005) 36:1115–9. doi: 10.1161/01.str.0000166053.83476.4a

30. Arboix A, García-Eroles L, Massons JB, Oliveres M, Pujades R, Targa C. Atrial fibrillation and stroke: clinical presentation of cardioembolic versus atherothrombotic infarction. *Int J Cardiol.* (2000) 73:33–42.

31. Hsieh C-Y, Lee C-H, Sung S-F. Development of a novel score to predict newly diagnosed atrial fibrillation after ischemic stroke: The CHASE-LESS score. *Atherosclerosis.* (2020) 295:1–7. doi: 10.1016/j.atherosclerosis.2020.01.003

32. Sung S-F, Hsieh C-Y, Hu Y-H. Early prediction of functional outcomes after acute ischemic stroke using unstructured clinical text: retrospective cohort study. *JMIR Med Inform.* (2022) 10:e29806. doi: 10.2196/29806

33. Hsieh F-I, Lien L-M, Chen S-T, Bai C-H, Sun M-C, Tseng H-P, et al. Get with the guidelines-stroke performance indicators: surveillance of stroke care in the taiwan stroke registry. *Circulation.* (2010) 122:1116–23. doi: 10.1161/circulationaha.110.936526

34. Sposato LA, Chaturvedi S, Hsieh C-Y, Morillo CA, Kamel H. Atrial fibrillation detected after stroke and transient ischemic attack: a novel clinical concept challenging current views. *Stroke.* (2022) 53:e94–103. doi: 10.1161/strokeaha.121.034777

35. Mujtaba G, Shuib L, Idris N, Hoo WL, Raj RG, Khowaja K, et al. Clinical text classification research trends: systematic literature review and open issues. *Expert Syst Appl.* (2019) 116:494–520. doi: 10.1016/j.eswa.2018.09.034

36. Deng X, Li Y, Weng J, Zhang J. Feature selection for text classification: a review. *Multimed Tools Appl.* (2018) 78:3797–816. doi: 10.1007/s11042-018-6083-5

37. Ma S, Huang J. Penalized feature selection and classification in bioinformatics. *Brief Bioinform.* (2008) 9:392–403. doi: 10.1093/bib/bbn027

38. Branco P, Torgo L, Ribeiro RP. A Survey of predictive modeling on imbalanced domains. *ACM Comput Surv (CSUR).* (2016) 49:1–50. doi: 10.1145/2907070

39. Ogunleye AA, Qing-Guo W. XGBoost model for chronic kidney disease diagnosis. *IEEE/ACM Trans Comput Biol Bioinform.* (2019) 17:2131–40. doi: 10.1109/tcbb.2019.2911071

40. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell.* (2020) 2:56–67. doi: 10.1038/s42256-019-0138-9

41. LaValley MP. Logistic regression. *Circulation.* (2008) 117:2395–9. doi: 10.1161/circulationaha.106.682658

42. Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D'Agostino RB, et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. *Lancet.* (2009) 373:739–45. doi: 10.1016/s0140-6736(09)60443-8

43. Liao J-N, Lim S-S, Chen T-J, Tuan T-C, Chen S-A, Chao T-F. Modified taiwan atrial fibrillation score for the prediction of incident atrial fibrillation. *Front Cardiovasc Med.* (2022) 8:805399. doi: 10.3389/fcvm.2021.805399

44. Chen Y-L, Wang H-T, Chen H-C, Liu W-H, Hsueh S, Chung W-J, et al. A risk stratification scoring system for new-onset atrial fibrillation after ischemic stroke. *Medicine.* (2020) 99:e20881. doi: 10.1097/md.0000000000020881

45. Ashburner JM, Wang X, Li X, Khurshid S, Ko D, Lipsanopoulos AT, et al. Re-CHARGE-AF: recalibration of the CHARGE-AF model for atrial fibrillation risk prediction in patients with acute stroke. *J Am Heart Assoc.* (2021) 10:e022363. doi: 10.1161/jaha.121.022363

46. Wang W, Alonso A, Soliman EZ, O'Neal WT, Calkins H, Chen LY, et al. Relation of resting heart rate to incident atrial fibrillation (From ARIC [atherosclerosis risk in communities] study). *Am J Cardiol.* (2018) 121:1169–76. doi: 10.1016/j.amjcard.2018.01.037

47. Drozdowska BA, Singh S, Quinn TJ. Thinking about the future: a review of prognostic scales used in acute stroke. *Front Neurol.* (2019) 10:274. doi: 10.3389/fneur.2019.00274

48. Orfanoudaki A, Chesley E, Cadisch C, Stein B, Nouh A, Alberts MJ, et al. Machine learning provides evidence that stroke risk is not linear: the non-linear Framingham stroke risk score. *PLoS ONE.* (2020) 15:e0232414. doi: 10.1371/journal.pone.0232414

49. Alaa AM, Bolton T, Angelantonio ED, Rudd JHF, Schaar M, van der. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS ONE.* (2019) 14:e0213653. doi: 10.1371/journal.pone.0213653

50. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *New Engl J Med.* (2019) 380:1347–58. doi: 10.1056/nejmra1814259

51. Kamel H, Navi BB, Parikh NS, Merkler AE, Okin PM, Devereux RB, et al. Machine learning prediction of stroke mechanism in embolic strokes of undetermined source. *Stroke.* (2020) 51:e203–10. doi: 10.1161/strokeaha.120.029305

52. Raghunath S, Pfeifer JM, Ulloa-Cerna AE, Nemani A, Carbonati T, Jing L, et al. Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ECG and help identify those at risk of atrial fibrillation–related stroke. *Circulation.* (2021) 143:1287–98. doi: 10.1161/circulationaha.120.047829

53. Nadarajah R, Wu J, Frangi AF, Hogg D, Cowan C, Gale C. Predicting patient-level new-onset atrial fibrillation from population-based nationwide electronic health records: protocol of FIND-AF for developing a precision medicine prediction model using artificial intelligence. *BMJ Open.* (2021) 11:e052887. doi: 10.1136/bmjopen-2021-052887

54. Aguilar M, Macle L, Deyell MW, Yao R, Hawkins N, Khairy P, et al. The influence of monitoring strategy on assessment of ablation success and post-ablation atrial fibrillation burden assessment: implications for practice and clinical trial design. *Circulation.* (2021) 145:21–30. doi: 10.1161/circulationaha.121.056109

55. Li J, Liu L, Le TD, Liu J. Accurate data-driven prediction does not mean high reproducibility. *Nat Mach Intell.* (2020) 2:13–5. doi: 10.1038/s42256-019-0140-2