Predicting Stroke Risk in Patients with Atrial Fibrillation Using Machine Learning

Seonwoo Jung
Chonnam National University
Eunjoo Lee
National Health Insurance Service
Minji Lee
Chonnam National University
Sejin Bae
National Health Insurance Service
Yeon-Yong Kim
National Health Insurance Service
Doheon Lee
Korea Advanced Institute of Science and Technology
Min-Keun Song
Chonnam National University Hospital
Sunyong Yoo (syyoo@jnu.ac.kr)
Chonnam National University

Research Article

Keywords: Atrial fibrillation, KNHIS, CHADS2

Posted Date: February 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-209356/v1

Title

Predicting Stroke Risk in Patients with Atrial Fibrillation Using Machine Learning

Seonwoo Jung[1,†], Eunjoo Lee[2,†], Minji Lee[1], Sejin Bae[2], Yeon-Yong Kim[2], Doheon Lee[3,4], Min-Keun Song[5,*] and Sunyong Yoo[1,*]

[1]Department of ICT Convergence System Engineering, Chonnam National University, Gwangju 61186, Republic of Korea

[2]Big Data Steering Department, National Health Insurance Service, Wonju 26464, Republic of Korea

[3]Bio-Synergy Research Center, Daejeon 34141, Republic of Korea

[4]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea

[5]Department of Physical & Rehabilitation Medicine, Chonnam National University Medical School & Hospital, Gwangju 61469, Republic of Korea.

†Cofirst author

*Corresponding author

E-mail addresses

Min-Keun Song (Corresponding author): drsongmk@jnu.ac.kr

Sunyong Yoo (Corresponding author): syyoo@jnu.ac.kr

Abstract

Atrial fibrillation (AF) is a well-known risk factor for stroke. Predicting the risk is important to prevent the first attack and re-attack of cerebrovascular diseases by determining the medication. Although several statistical methods have been developed to assess the stroke risk in AF patients, considerable improvement is needed in predictive performance. We propose a machine learning-based approach based on the massive and complex Korean National Health Insurance (KNHIS) data. We extracted 72-dimensional features, including demographics, health examination, and medical history information, of 754,949 patients with AF from KNHIS. Logistic regression was used to determine whether the extracted features had a statistically significant association with stroke occurrence. Then, we constructed the stroke risk prediction model based on a deep neural network. The extracted features were used as input, and the occurrence of stroke after the diagnosis of AF was the output used to train the model. When the proposed deep learning model was applied to 150,989 AF patients, it was confirmed that stroke risk was predicted with high accuracy, sensitivity, and specificity. As part of preventive medicine, this study could help AF patients prepare for stroke prevention based on predicted stoke associated feature and risk scores.

Introduction

Stroke is a fatal disease and the second and third leading cause of death and disability, respectively[1]. It can lead to various functional impairments such as motor weakness, sensory deficit, dysphagia, dysarthria, aphasia, cognitive impairment, and emotional disturbances[2-4]. Therefore, it is important to prevent it by predicting the risk and providing appropriate treatments. In particular, it is important to prevent re-attack by determining the prescription of anticoagulants in patients with atrial fibrillation (AF) during a cardioembolic stroke.

AF is a common risk factor of cardioembolic cerebral infarction[5]. It accounts for 7 to 31 percent of stroke patients aged 60 years or older[6-8]. Thromboembolism in the left atrium caused by AF would increase the risk of stroke by four to five times[7-9]. The recent population-based study presented AF as an independent predictor of 30-day and one-year mortality after a first ischemic stroke[10]. Approximately 17 percent of all deaths were attributable to the ischemic stroke with AF. Moreover, patients who have experienced an embolic attack have the most potent risk factor for recurrent stroke. The risk in the first few weeks after the initial attack is three to five percent based on patients[11,12]. The previous observation study showed that stroke with AF would affect the impact, quality of life in the elderly, and socioeconomic implications of the attendant[13]. Due to the high risk of recurrent embolism, the development of risk calculating methods for stroke with AF is in progress. In particular, because the pathophysiology of stroke in AF is different from that of non-AF, there is a need for a method that considers these characteristics[14-16].

Most previous studies for predicting stroke risk in patients with AF were based on statistical methods, such as $CHADS_2$ and CHA_2DS_2-VASc scores[17-19]. The $CHADS_2$ score would reflect the representation of incidence risk for stroke using five factors, including congestive heart failure, hypertension, more than 74 years of age, diabetes mellitus, and previous

67 cerebrovascular attack[11]. However, $CHADS_2$ has a limitation in that it is difficult to accurately
68 evaluate low-risk groups. In the low-risk group with a score of 0 to 1, it is not easy to determine
69 the anticoagulant regimen based on the score, as the risk varied greatly depending on the risk
70 factors[20]. To improve the predictive performance in the low-risk group, a CHA_2DS_2-VASc score
71 was proposed considering the presence or absence of vascular diseases, ages 65-74 years, and
72 female gender. CHA_2DS_2-VASc scores have guided many clinicians on using oral
73 anticoagulants as an indicator of bleeding risk, which could suggest its low use for stroke with
74 AF owing to high CHA_2DS_2-VASc scores[21]. It was devised to compensate for the defects of
75 $CHADS_2$, but there are other limitations. First, CHA_2DS_2-VASc scores are not enough to predict
76 the incidence of stroke. For example, only vascular diseases were considered, and other
77 mechanisms of thromboembolism were not considered. Second, the features were proposed
78 based on the situation 20 years ago. Some features may require new factors to adhere to the
79 current situation.

80 Herein, we present a machine learning-based method to predict the stroke risk in AF patients
81 based on the Korean National Health Insurance Service (KNHIS) data. Recent studies have
82 demonstrated that accumulated data of patients in electronic health record (EMR) can be
83 utilized to predict potential disease risk[22,23]. In Korea, more than 97% of the population is
84 covered by the KNHIS program and the remaining three percent are covered by a medical aid
85 program operated by the KNHIS[24]. The KNHIS contains information on Korean demographic,
86 health examination, and medical use/transaction information. Therefore, it was hypothesized
87 that the accumulated large-scale KNHIS information of the AF patients can be used to predict
88 the further risk of stroke. To handle the massive and complex KNHIS information, we adapted
89 a deep neural network that can be captured the patterns within the data by constructing multiple
90 hidden layers. The evaluation results showed that many stroke patients were identified with
91 high accuracy, sensitivity, and specificity.

Materials and methods

Data sources

This study used KNHIS data from January 1, 2005 to December 31, 2018. Since 1995, KNHIS, the single national health insurer, has provided health examinations for all Koreans. The KNHIS database contains complete health information about approximately 50 million Koreans[25]. In this study, case subjects were defined as patients with AF who were newly diagnosed with stroke, and control subjects were those with AF who had not been diagnosed with stroke. We used the International Classification of Disease, 10th revision (ICD-10) codes to identify patients with AF and those who had experienced stroke from the health claim records[26]. We found 754,949 patients diagnosed with AF (ICD-10: I48) between 2005 and 2013. Subsequently, we checked if the selected patients were hospitalized for stroke (ICD-10: I63) within five years after the diagnosis of AF. 62,226 stroke patients with AF and 692,723 non-stroke patients with AF were identified. Next, we collected demographic, health examination, and medical history information of subjects from the KNHIS database. Demographic information contains gender, age, occupational status, and income level. The medical history includes information on the occurrence of 43 diseases (e.g., hypertensive disease, hemolytic anemia, chronic gastritis, hyperlipidemia, and thyroid diseases) and the history of using seven antithrombotic agents by the subject in the past three years. The antithrombotic agent history was extracted based on the Anatomical Therapeutic Chemical (ATC) code B01. The medical history information in the KNHIS database is built using the medical bills that were claimed by medical service providers for the expenses. Health examination includes results of nine general laboratory tests (e.g., blood pressure, urinary protein, and obesity) and six questionnaires on lifestyle and behavior (e.g., smoking, exercise, and drinking). The study protocol was approved by the Institutional

Review Board of the National Health Insurance Service in Korea (NHIS-2020-4-109). The authors confirm that all methods were performed in accordance with relevant guidelines and regulations. The need for informed consent from participants was waived by the ethics committee of the Chonnam National University because this study involved routinely collected medical data that were anonymized at all stages to protect an individual's privacy.

Regression-based statistical analysis

Logistic regression is a statistical technique that estimates the causal relationship between categorical dependent variables and several independent variables and is divided into two types according to the number of categories of dependent variables[27]. A binary logistic regression is used when the dependent variable has two categories of 0 or 1, and polynomial logistic regression is used when the dependent variable is composed of two or more categories. The binary logistic regression, used as a statistical technique in this study, was expressed by defining logistic functions in reverse using logits as shown below in equations (1) and (2), to express linear relationships between independent and dependent variables[28].

$$p(x) = \sigma(t) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \quad (\because t = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) \quad (1)$$

$$g(p(x)) = \sigma^{-1}(p(x)) = logit(p(x)) = \ln\left(\frac{P(x)}{1-P(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \quad (2)$$

Regression coefficient, standard error, Wald chi-square, and *p*-value were used for binary logistic regression analysis with maximum likelihood estimation. The regression coefficient implies that the dependent variable increases or decreases in proportion to the estimated value when the independent variable increases by one unit[29]. If the coefficient is a positive value, it has a positive correlation and vice versa. As the coefficient was close to zero, the effect of the

independent variable decreased[29,30]. The standard error is the standard deviation of the sample means used to determine whether the regression coefficient occurs by accident, revealing the closeness of the sample mean values to the population mean[31]. The smaller the standard error, the closer it is to the population mean, and it spreads closer to the regression line, which implies that the probability of a regression coefficient being accidental is less likely to occur. This shows that the causal relationship between the independent variable and dependent variable is significant. The Wald chi-square is an index for evaluating the importance of each independent variable[27].

$$W = \left(\frac{\beta}{SE(\beta)}\right)^2 \quad (3)$$

where β is the coefficient, and SE is its standard error. The Wald chi-square refers to the ratio of the square of the regression coefficient to its standard error and is expressed as a chi-square distribution[27]. The higher the value, the lower the significance level, indicating that it is an important variable in explaining the dependent variable. The *p*-value is the probability that a value equal to or more than that in the sample is observed, assuming that the null hypothesis is correct[32]. Moreover, a *p*-value less than a certain significance level implies that the observed result is improbable under the null hypothesis and that there is a significant association between the dependent and the corresponding independent variables. However, a *p*-value greater than a certain significance level indicates that there is no significant association between the dependent and the corresponding independent variables.

Deep neural network for predicting the stroke risk in AF patients

In this study, we used a deep neural network to predict the stroke risk in AF patients based on KNHIS data (Fig. 1). The deep neural network is composed of multiple hidden layers between

an input layer and an output layer. The multiple hidden layers enable the modeling of complex nonlinear relationships through the learning function of a high-level layer formed by combining the features of the lower layer, and learning complex functions mapping the input to the output from data[33]. Among the 75 features extracted from KNHIS, we used 4 demographic information, 31 medical histories, and 13 health examination features, which were considered statistically significant through regression analysis. The dataset was divided into 6:2:2 as training, validation, and test set, respectively. We used 452,971 (stroke = 37,336, non-stroke = 415,635) samples for training the model, 150,989 (stroke = 12,445, non-stroke = 138,544) samples for validation, and 150,989 (stroke = 12,445, non-stroke = 138,544) samples for testing at each fold. Then, the self-attention mechanism was applied to the deep learning model. The self-attention mechanism improves the prediction performance by estimating the importance of the feature[34]. Input features were fed to the fully-connected and the softmax layers to calculate the self-attention scores.

$$a = softmax(g(X)) \quad (4)$$

where X is the selected input features, and $g(\cdot)$ is the fully-connected layer without activation. $g(\cdot)$ can be represented as below.

$$\mathrm{g}(x) = Wx + b \ (5)$$

where $W=[w_1, w_2,\ldots, w_n]$ is the weight matrix, and b is the bias of each unit. In this study, the output of linear operator $g(\cdot)$ is the same size as the input; therefore, $W\in\mathbb{R}^{48\times48}$ and $b\in\mathbb{R}^{48}$. $g(\cdot)$ is then fed into the softmax function which return a vector of numbers with equal to one.

$$softmax(z) = \frac{e^{z_i}}{\sum_i e^{z_i}} \quad (6)$$

Then, the component-wise multiplication between input features and self-attention score

vector was performed.

$$o = a \odot X \qquad (7)$$

where $\odot$ is the component-wise multiplication operator. Then, we concatenated the output vector o and input features x as $[o_i ; x_i]$. The concatenated vector was used to train the three-layer fully-connected neural network for predicting the stroke risk of AF patients.

The fully-connected network was constructed using several techniques. We applied the Rectified Linear Unit activation function for all hidden units to alleviate gradient vanishing. To compensate for the imbalance ratio between stroke groups and non-stroke groups, weight balancing was used to ensure that groups of strokes with a small number of samples contribute equally to overall losses. Batch normalization, which enables a stable learning by alleviating the difference in weight, was applied to the input layer to improve the learning speed, reduce overfitting, and avoid gradient vanishing[35]. As a loss function for gradient descent, a binary cross-entropy loss function was used for binomial classification problems. Also, the ADAM optimizer was used to optimize the loss function by adjusting the direction and step-size settings[36]. We applied early stopping which is a regularization technique to prevent the overfitting in the iterative procedure of gradient descent[37,38]. The model was trained for 2,000 epochs with early stopping (patience = 30).

Results

General characteristic of the study population

From June 2005 to March 2013, a total of 754,949 patients were diagnosed with AF, of which 62,226 (8.24%) were diagnosed with stroke five years after diagnosis of AF (Table 1). The

CHA_2DS_2-VASc score ranged from 0 to 9, indicating that the risk increases as the score increases. The mean CHA_2DS_2-VASc score of the non-stroke group was 2.15 points, and the stroke group was 3.01 points. As expected, it was confirmed that the stroke group had higher CHA_2DS_2-VASc scores than the non-stroke group. Next, we checked the individual risk factors of CHA_2DS_2-VASc scores, including five medical history factors, age, and sex. It was confirmed that the stroke group had a high proportion compared with the non-stroke group for the medical history of five diseases considered in the CHA_2DS_2-VASc score. The mean (±SD) age of the patients was 64.6±13.3 years in the non-stroke group and 71.5 ± 9.5 years in the stroke group. We also observed that the stroke group had a higher proportion of patients aged from 65 to 74 years and above 75 years than the non-stroke group. Regarding sex, the non-stroke group had a higher proportion of males (59.21%), whereas the stroke group had a higher proportion of females (50.34%). These results indicate that the CHA_2DS_2-VASc score reflects the characteristics of stroke because it gave a high score for the five medical history features, elderly, and women in the stroke group. However, the difference between the two groups was not significant, and the proportion of patients who scored five or higher in the stroke group did not reach 20%.

Statistical analysis

We used coefficient values of logistic regression to identify the features related to stroke occurrence. The *p*-value for each feature tests the null hypothesis that the feature does not correlate with the occurrence of stroke. In this study, we set the significance of the *p*-value as 0.001. The results showed that age, sex, and occupational status were important factors in demographic information. We identified important features from medical history, including thyroid diseases, other cardiac arrhythmias, chronic lower respiratory diseases, hemolytic

anemia, cancer, hemorrhoids, diabetes mellitus, hypertensive diseases, chronic kidney diseases, heart failure, hyperlipidemia, peripheral vascular disease, gout, noninflammatory gynecological problems, pulmonary embolism, and chronic gastritis. Also, vitamin K antagonists, direct thrombin inhibitors, heparin group, and enzymes were important features in the history of using antithrombotic agents.

Significant features found by the *p*-values through tests were analyzed based on coefficient (β), Wald chi-square (W), and odds ratio (OR) with 95% confidence interval (CI) (Table 2). Based on the regression coefficient values, we found that antithrombotic enzyme agent (β=1.114), transient cerebral ischemic attacks (β=0.7286), direct thrombin inhibitors (β =0.4592), platelet aggregation inhibitors excluding heparin (β=0.2959), vitamin K antagonist (β=0.2442), hemorrhagic stroke (β=0.1965) are significant features that have a positive correlation with the dependent variable. However, operation history (β=-0.4513), pulmonary embolism (β=-0.2693), and occupation status (employed) (β=-0.1518) are significant features that have a negative correlation with the dependent variable. Based on Wald chi-square, age (W=3650.57), transient cerebral ischemic attacks (W=3560.38), antithrombotic enzyme agent (W=1121.48), platelet aggregation inhibitors excluding heparin (W=818.36), vitamin K antagonist (W=627.71), sex (male) (W=282.44), hypertensive diseases (W=176.22), occupation status (employed) (W=155.46), operation history (W=148.14), diabetes mellitus (W=140.23), and direct thrombin inhibitors (W=107.80) are significant features. Through the odds ratio (95% confidence interval [CI]), we found that antithrombotic enzyme agents (OR=3.046), transient cerebral ischemic attacks (OR=2.072), direct thrombin inhibitors (OR=1.583), platelet aggregation inhibitors excluding heparin (OR=1.344), vitamin K antagonist (OR=1.277), and hemorrhagic stroke (OR=1.217) had significantly high odds. These statistically significant features were used to train the deep neural network.

253

254 **Predicting stroke risk in patients with AF**

255 Our method predicts the stroke risk in AF patients based on KNHIS data. We evaluated the
256 average area under the curve scores of the receiver operating characteristic (AUROC) and
257 accuracy to assess the predictive performance. We tested the performance for four different
258 types of input feature sets: (i) using all features; (ii) using demographic features only; (iii) using
259 medical history features only; and (iv) using health examination features only (Fig. 2a). From
260 the results, we found that using all features (AUROC=0.724) exhibited better performance than
261 using a single feature set only (AUROC=0.614~0.681). These results indicate that the proposed
262 model considers the complex associations of large-scale feature sets. Furthermore, we
263 compared the prediction performance of our method with the CHA_2DS_2-VASc scores (Fig. 2b).
264 For this, we first checked the CHA_2DS_2-VASc scores of the subjects in this study (Table 1).
265 When calculating the AUROC of the CHA_2DS_2-VASc scores, a value of 0.645 was confirmed.
266 These results showed that the prediction performance of the proposed method is better than that
267 of the CHA_2DS_2-VASc scores.

268 The model output can be interpreted as an approximate probability of stroke occurrence and
269 has a value between 0 and 1. In general, the decision threshold that predicts stroke occurrence
270 based on the model output value is often 0.5. However, the default threshold may not represent
271 an optimal interpretation of the predicted probabilities[39]. In this study, the class distribution of
272 the dataset is skewed, and predicted probabilities are not calibrated. This is a classification
273 problem with imbalanced classes[40]. To solve this, we identified the optimal threshold value of
274 the model output to judge the occurrence of stroke. We calculated the F1-scores, which is the
275 harmonic mean of precision and recall, by changing the threshold of the model output. The best
276 performance (F1-score = 0.225) was when the threshold value was 0.513. Next, the accuracy

of our method was compared using all features with the method using a single feature set only. The result showed that using all features has the best performance (accuracy=0.842) compared to using a single feature set only (accuracy=0.781–0.807).

Discussion

AF is the most common sustained arrhythmia. $CHADS_2$ and CHA_2DS_2-VASc scores are the most popular methods for predicting stroke risks in patients with AF. However, these scores may not be enough to predict the incidence of stroke as they only use five to seven features based on limited information. Previous studies demonstrated that the pathogenesis of stroke in AF patients is complex and involves various factors, such as hypertension, diabetes, dementia, and obesity[41,42]. Moreover, based on KNHIS data, this study found that 32 features were statistically significantly associated with stroke. Therefore, more accurate predictions will be possible if the information from stroke patients with AF can be completely utilized.

The EMR data accumulated in the hospital applies to this approach because it contains various medical information about patients. However, sharing or releasing EMR data is very difficult owing to privacy and confidentiality issues[43,44]. Therefore, there is a limit to analyzing past medical information of patients who have used several hospitals. In recent years, the Observational Health Data Sciences and Informatics (OHDSI) project has been attempting to standardize and expand the EMR information of hospitals; however, it is currently being conducted only for a few hospitals, and technical and institutional improvements are needed[45]. In Korea, the records of diagnosis and prescriptions generated by all medical institutions are collected by KNHIS. Therefore, it is possible to develop and apply a model with high coverage by utilizing KNHIS data. In this study, a machine learning method was applied based on this data. The validation results showed that performance improved when predictions were made

using various types of information, including demographic, medical history, and health examination. Most risk prediction methods were developed in different cohort studies[46]. They are not suitable for the Korean population as their clinical trial cohorts include information on people from different races. This study is significant as the prediction model was developed by considering the characteristics of the Korean population with high accuracy and coverage.

Conclusions

This study proposes a new model to predict stroke risk in patients with AF. To prevent stroke, a system must be established to warn of risks. This study predicted stroke risk in AF patients based on a machine learning approach by utilizing the massive and complex KNHIS data. The validation results showed that the proposed machine learning model has high accuracy, sensitivity, and specificity compared to CHA_2DS_2-VASc scores. The outcomes of this study are significant because, as part of preventive medicine, patients with AF can prepare for stroke prevention based on the predicted risk values.

Acknowledgements

This research was supported by the Bio-Synergy Research Project (NRF-2012M3A9C4048758) of the Ministry of Science, ICT, and Future Planning, through the National Research Foundation, and supported by the National Research Foundation of Korea grant funded by the Korea government (MSIT) (NRF-2020R1C1C1006007).

Author contributions

S.Y. proposed the objective and motivation of this work and designed overall method. S.J., E.L., Y.K. and S.B. performed data-preprocessing. S.J. and M.L. performed preliminary study. D.L., M.S. and S.Y. helped to write the main manuscript text and provided comments that improve introduction and method parts. S.J. and E.L. performed evaluation process. E.J. and M.S. provided some ideas in discussion. M.S., and S.Y. supervised this work.

Competing interests

The author(s) declare no competing interests.

References

1 Johnson, W., Onuma, O., Owolabi, M. & Sachdev, S. Stroke: a global response is needed. *Bull World Health Organ* **94**, 634-634A, doi:10.2471/BLT.16.181636 (2016).

2 Kim, K.-J., Kim, H.-Y. & Chun, I.-A. Correlations between the sequelae of stroke and physical activity in Korean adult stroke patients. *Journal of Physical Therapy Science* **28**, 1916-1921, doi:10.1589/jpts.27.1916 (2016).

3 Mukherjee, D., Levin, R. L. & Heller, W. The Cognitive, Emotional, and Social Sequelae of Stroke: Psychological and Ethical Concerns in Post-Stroke Adaptation. *Topics in Stroke Rehabilitation* **13**, 26-35, doi:10.1310/tsr1304-26 (2006).

4 Schneider, A. T. *et al.* Trends in Community Knowledge of the Warning Signs and Risk Factors for Stroke. *JAMA* **289**, 343-346, doi:10.1001/jama.289.3.343 (2003).

5 Boehme AK, Esenwa C, Elkind MS. Stroke Risk Factors, Genetics, and Prevention. Circ Res. 2017;120(3):472-495. doi:10.1161/CIRCRESAHA.116.308398

6 Go, A. S. *et al.* Prevalence of Diagnosed Atrial Fibrillation in AdultsNational Implications for Rhythm Management and Stroke Prevention: the AnTicoagulation and Risk Factors In Atrial Fibrillation (ATRIA) Study. *JAMA* **285**, 2370-2375, doi:10.1001/jama.285.18.2370 (2001).

7 Waldo, A. L., Becker, R. C., Tapson, V. F., Colgan, K. J. & Committee, N. S. Hospitalized patients with atrial fibrillation and a high risk of stroke are not being provided with adequate anticoagulation. *J Am Coll Cardiol* **46**, 1729-1736, doi:10.1016/j.jacc.2005.06.077 (2005).

8 The Effect of Low-Dose Warfarin on the Risk of Stroke in Patients with Nonrheumatic Atrial Fibrillation. *New England Journal of Medicine* **323**, 1505-1511,

doi:10.1056/nejm199011293232201 (1990).
9 Friberg, L. & Bergfeldt, L. Atrial fibrillation prevalence revisited. *Journal of Internal Medicine* **274**, 461-468, doi:10.1111/joim.12114 (2013).
10 Marini, C. *et al.* Contribution of atrial fibrillation to incidence and outcome of ischemic stroke: results from a population-based study. *Stroke* **36**, 1115-1119 (2005).
11 The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke. International Stroke Trial Collaborative Group. *Lancet (London, England)* **349**, 1569-1581 (1997).
12 Saxena, R., Lewis, S., Berge, E., Sandercock, P. A. & Koudstaal, P. J. Risk of early death and recurrent stroke and effect of heparin in 3169 patients with acute ischemic stroke and atrial fibrillation in the International Stroke Trial. *Stroke* **32**, 2333-2337, doi:10.1161/hs1001.097093 (2001).
13 Dalen, J. E. & Alpert, J. S. Silent atrial fibrillation and cryptogenic strokes. *The American Journal of Medicine* **130**, 264-267 (2017).
14 Kamel, H., Okin, P. M., Elkind, M. S. & Iadecola, C. Atrial fibrillation and mechanisms of stroke: time for a new model. *Stroke* **47**, 895-900 (2016).
15 D'Souza, A., Butcher, K. S. & Buck, B. H. The multiple causes of stroke in atrial fibrillation: thinking broadly. *Canadian Journal of Cardiology* **34**, 1503-1511 (2018).
16 Violi, F., Soliman, E. Z., Pignatelli, P. & Pastori, D. Atrial fibrillation and myocardial infarction: a systematic review and appraisal of pathophysiologic mechanisms. *Journal of the American Heart Association* **5**, e003347 (2016).
17 Malone, D. C. *et al.* PRM26 The Use of Claims-Based CHA2DS2-VASc and ATRIA Scores to Predict Stroke/Systemic Embolism and Bleeding Rates Among Anticoagulated Patients With Atrial Fibrillation (AFIB) in a Pharmacy-Benefit Management (PBM) Environment. *Value in Health* **15**, 464 (2012).
18 Potpara, T. S. & Olesen, J. B. Comparing the ATRIA, CHADS2, and CHA2DS2-VASc Scores for Stroke Prediction in Atrial Fibrillation. *Journal of the American College of Cardiology* **67**, 2316-2317 (2016).
19 van den Ham, H. A., Klungel, O. H., Singer, D. E., Leufkens, H. G. & van Staa, T. P. Comparative Performance of ATRIA, CHADS2, and CHA2DS2-VASc Risk Scores Predicting Stroke in Patients With Atrial Fibrillation. *Journal of the American College of Cardiology* **66//SUP**, 1851-1859 (2015).
20 Olesen, J. B., Torp-Pedersen, C., Hansen, M. L. & Lip, G. Y. H. The value of the CHA~2DS~2-VASc score for refining stroke risk stratification in patients with atrial fibrillation with a CHADS~2 score 0-1: A nationwide cohort study. *Thrombosis and Haemostasis* **107**, 1172-1179 (2012).
21 Roldán, V. *et al.* The HAS-BLED score has better prediction accuracy for major bleeding than CHADS2 or CHA2DS2-VASc scores in anticoagulated patients with atrial fibrillation. *Journal of the American College of Cardiology* **62**, 2199-2204 (2013).
22 Perotte, A., Ranganath, R., Hirsch, J. S., Blei, D. & Elhadad, N. Risk prediction for chronic

kidney disease progression using heterogeneous electronic health record data and time series analysis. *Journal of the American Medical Informatics Association* **22**, 872-880 (2015).

23 Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE journal of biomedical and health informatics* **22**, 1589-1604 (2018).

24 Shin, D. W., Cho, B. & Guallar, E. Korean National Health Insurance database. *JAMA internal medicine* **176**, 138-138 (2016).

25 Kwon, S. Payment system reform for health care providers in Korea. *Health policy and planning* **18**, 84-92 (2003).

26 Wilchesky, M., Tamblyn, R. M. & Huang, A. Validation of diagnostic codes within medical services claims. *Journal of clinical epidemiology* **57**, 131-141 (2004).

27 Park, H.-A. An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. *jkan* **43**, 154-164, doi:10.4040/jkan.2013.43.2.154 (2013).

28 Peng, C.-Y. J., Lee, K. L. & Ingersoll, G. M. An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research* **96**, 3-14, doi:10.1080/00220670209598786 (2002).

29 Guthery, F. S. & Bingham, R. L. A primer on interpreting regression models. *The Journal of Wildlife Management* **71**, 684-692 (2007).

30 Peng, C.-Y. J., So, T.-S. H., Stage, F. K. & John, E. P. S. The use and interpretation of logistic regression in higher education journals: 1988–1999. *Research in higher education* **43**, 259-293 (2002).

31 McHugh, M. L. Standard error: meaning and interpretation. *Biochemia medica: Biochemia medica* **18**, 7-13 (2008).

32 Wasserstein, R. L. & Lazar, N. A. The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician* **70**, 129-133, doi:10.1080/00031305.2016.1154108 (2016).

33 Bengio, Y. Learning Deep Architectures for AI. *Foundations and Trends® in Machine Learning* **2**, 1-127, doi:10.1561/2200000006 (2009).

34 Hyunho, K. & Hojung, N. hERG-Att: Self-Attention-Based Deep Neural Network for Predicting hERG Blockers. *Computational Biology and Chemistry*, 107286 (2020).

35 Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

36 Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

37 Prechelt, L. Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks* **11**, 761-767 (1998).

38 Yao, Y., Rosasco, L. & Caponnetto, A. On early stopping in gradient descent learning. *Constructive Approximation* **26**, 289-315 (2007).

39 Brownlee, J. *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. (Machine Learning Mastery, 2020).

40 Fernández, A. *et al. Learning from imbalanced data sets*. (Springer, 2018).
41 Kim, Y.-H. & Roh, S.-Y. The mechanism of and preventive therapy for stroke in patients with atrial fibrillation. *Journal of Stroke* **18**, 129 (2016).
42 Sanoski, C. A. Prevalence, pathogenesis, and impact of atrial fibrillation. *American Journal of Health-System Pharmacy* **67**, S11-S16 (2010).
43 Terry, N. P. & Francis, L. P. Ensuring the privacy and confidentiality of electronic health records. *U. Ill. L. Rev.*, 681 (2007).
44 Price, W. N. & Cohen, I. G. Privacy in the age of medical big data. *Nature medicine* **25**, 37-43 (2019).
45 Hripcsak, G. *et al.* Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Studies in health technology and informatics* **216**, 574 (2015).
46 Senoo, K., Lane, D. & Lip, G. Y. Stroke and bleeding risk in atrial fibrillation. *Korean circulation journal* **44**, 281-290 (2014).

Figure legends

Figure 1. A systematic overview of the deep neural network-based model that predicts stroke risk in AF patients. Demographic, medical history, and health examination information was used as input, and occurrence of stroke in AF patients was used as output. We calculated attention scores for the input features and concatenated the attention scores with input features. The stroke risk was predicted by a three-layer fully-connected neural network with non-linear activation function.

Figure 2. AUROC value of models generated from different datasets and methods. (a) We compared the AUROC performance of the proposed method for four different input datasets, including all features, demographic feature only, medical history features only, and health examination features only. (b) We compared the performance of our method with the CHA_2DS_2-VASc scores.

462 **Tables**

463 **Table 1. Baseline characteristics of the study cohort.** We analyzed patients with AF according

464 to the stroke and non-stroke groups.

	Non-stroke patients (n = 692,723)	Stroke patients (n = 62,226)
CHA_2DS_2-VASc scores		
0	105,511 (15.23)	3,416 (5.49)
1	169,613 (24.48)	8,912 (14.32)
2	156,863 (22.64)	13,174 (21.17)
3	124,433 (17.96)	14,055 (22.59)
4	77,692 (11.22)	11,071 (17.79)
5	36,049 (5.2)	6,210 (9.98)
6	15,303 (2.21)	3,427 (5.51)
7	5,879 (0.85)	1,565 (2.52)
8	1,275 (0.18)	356 (0.57)
9	105 (0.02)	40 (0.06)
Medical history		
Heart failure	92,580 (13.36)	11,301 (18.16)
Hypertension	335,126 (48.38)	36,252 (58.26)
Diabetes mellitus	122,587 (17.70)	13,802 (22.18)
Stroke/TIA/thromboembolism	79,456 (11.47)	15,759 (25.33)
Vascular disease	47,126 (6.8)	5,561 (8.94)
Age		
Age ≥75	106,976 (15.44)	16,682 (26.80)
Age 65-74	184,629 (26.65)	19,907 (31.99)
Sex		
Male	410,197 (59.21)	30,900 (49.66)
Female	282,526 (40.79)	31,326 (50.34)

465

466

467 **Table 2. The result of logistic regression with maximum likelihood estimation.** We
468 calculated coefficient, standard error, and Wald chi-square for the input features.

Feature	Coefficient	Wald chi-square	Odds ratio (95% CI)
Age	0.0427	3650.57	1.044 (1.042-1.045)
Sex (Male)	-0.1071	282.44	0.807 (0.787-0.828)
Occupation status (Unemployed)	0.0733	54.16	1.009 (0.985-1.035)
Occupation status (Employed)	-0.1518	155.46	0.806 (0.78-0.833)
Thyroid diseases (ICD-10:E00~E07)	-0.0658	15.24	0.936 (0.906-0.968)
Other cardiac arrhythmias (ICD-10: I44, I45, I47, I49)	-0.1266	88.18	0.881 (0.858-0.905)
Hemolytic anemias (ICD-10: D55~D59)	-0.1024	15.40	0.903 (0.858-0.95)
Cancer (ICD-10: C00~C97, D00~D90, D37~D48)	-0.1502	79.59	0.861 (0.833-0.889)
Hemorrhoids (ICD-10: I84, K64)	-0.0955	19.54	0.909 (0.871-0.948)
Diabetes mellitus (ICD-10: E10~E14)	0.1473	140.23	1.159 (1.131-1.187)
Hypertensive disease (ICD-10: I10~I16)	0.1465	176.22	1.158 (1.133-1.183)
Chronic kidney disease (ICD-10: N18, N19)	-0.147	20.26	0.863 (0.81-0.92)
Hyperlipidemia (ICD-10: E78)	-0.0634	14.76	0.939 (0.909-0.969)
Peripheral vascular disease (ICD-10: I70~I73)	0.0999	35.50	1.105 (1.069-1.142)
Chronic gastritis (ICD-10: K21, K25~K29)	-0.0562	36.52	0.945 (0.928-0.963)
Hemorrhagic stroke (ICD-10: I60~I62)	0.1965	29.99	1.217 (1.134-1.306)
Transient cerebral ischemic attacks (ICD-10: G45)	0.7286	3560.38	2.072 (2.023-2.122)
Gout (ICD-10: M10)	0.0813	12.74	1.085 (1.037)
Benign prostatic hyperplasia (ICD-10: N40)	-0.0855	37.64	0.918 (0.893-0.943)
Chronic lower respiratory diseases (ICD-10: J40-J47)	-0.0486	27.40	0.953 (0.935-0.97)

Noninflammatory gynecological problems (ICD-10: N81, N84~N90, N93, N95)	-0.0823	16.55	0.921 (0.885-0.958)
Pulmonary embolism (ICD-10: I26)	-0.2693	13.88	0.764 (0.663-0.88)
Heart failure (ICD-10: I50)	0.0878	42.95	1.092 (1.063-1.121)
Vitamin K antagonist (ATC code: B01AA)	0.2442	627.71	1.277 (1.252-1.301)
Heparin group (ATC code: B01AB)	0.1145	94.64	1.121 (1.096-1.147)
Platelet aggregation inhibitors excluding heparin (ATC code: B01AC)	0.2959	818.36	1.344 (1.317-1.372)
Antithrombotic enzyme agents (ATC code: B01AD)	1.114	1121.48	3.046 (2.854-3.252)
Direct thrombin inhibitors (ATC code: B01AE)	0.4592	107.80	1.583 (1.451-1.726)
Hospitalization history	0.0447	18.78	1.046 (1.025-1.067)
Operation history	-0.4513	148.14	0.637 (0.592-0.685)
The number of patient days	0.000331	28.98	1 (1-1)
Total medical costs	2.92E-09	17.23	1 (1-1)

Figures

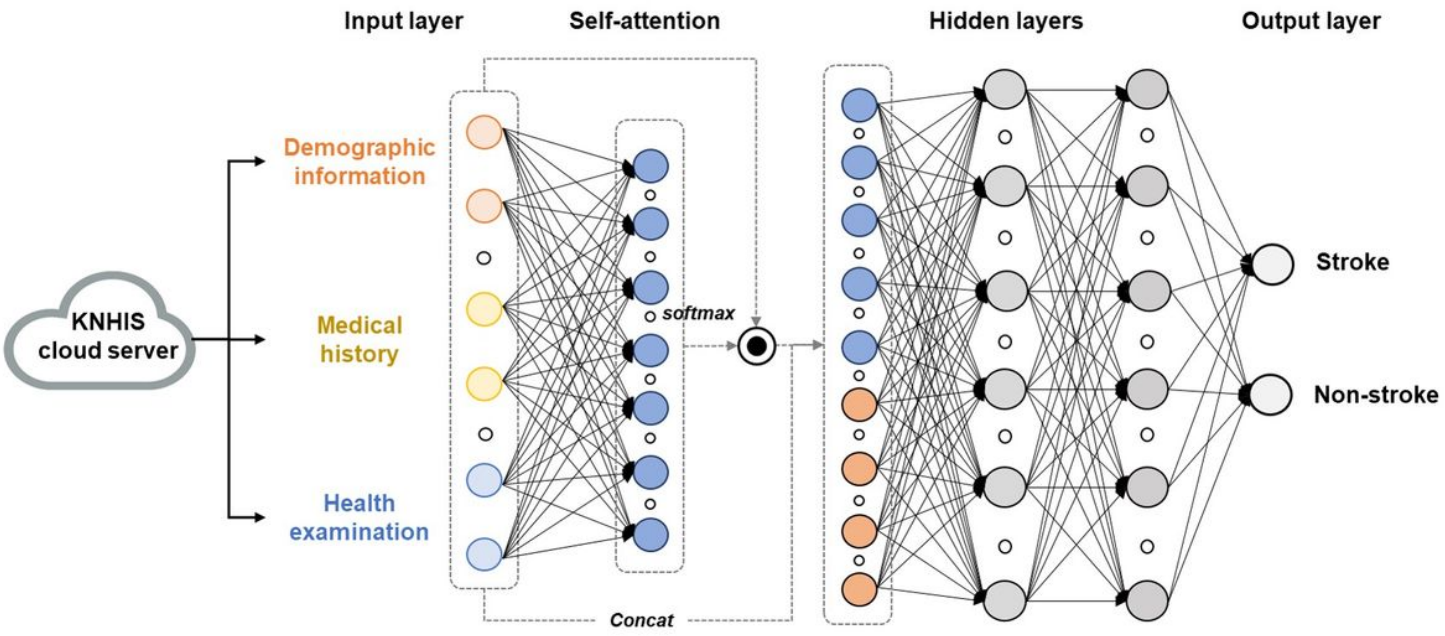

Figure 1

A systematic overview of the deep neural network-based model that predicts stroke risk in AF patients. Demographic, medical history, and health examination information was used as input, and occurrence of stroke in AF patients was used as output. We calculated attention scores for the input features and concatenated the attention scores with input features. The stroke risk was predicted by a three-layer fully-connected neural network with non-linear activation function.

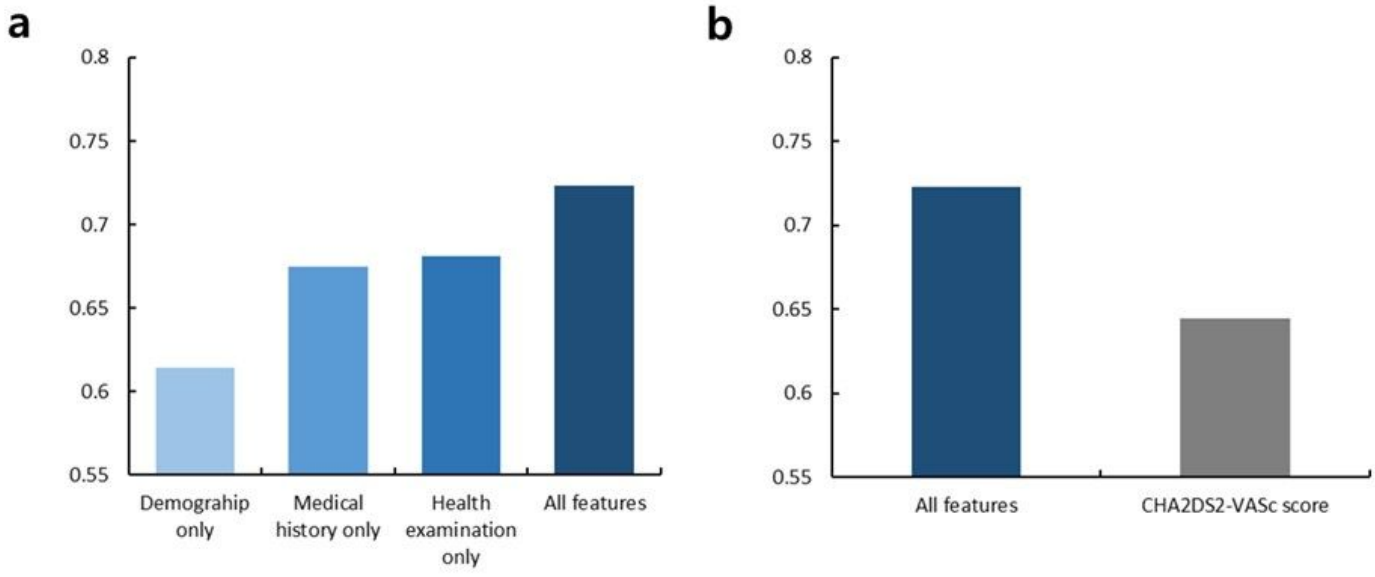

Figure 2

AUROC value of models generated from different datasets and methods. (a) We compared the AUROC performance of the proposed method for four different input datasets, including all features, demographic feature only, medical history features only, and health examination features only. (b) We compared the performance of our method with the CHA2DS2-VASc scores.