Develop and validate a nomogram for predicting stroke inrheumatoidarthritis patients by electronic medical record data in northern China

Fangran Xin
First Affiliated Hospital of China Medical University
Bowen Yang
First Affiliated Hospital of China Medical University
Lingyu Fu (fulingyucmu@sina.com)
First Affiliated Hospital of China Medical University
Haina Liu
First Affiliated Hospital of China Medical University
Tingting Wei
First Affiliated Hospital of China Medical University
Cunlu Zou
Neusoft Research of Intelligent Healthcare Technology
Bingqing Bai
First Affiliated Hospital of China Medical University

Research article

Keywords: Rheumatoid arthritis, stroke, lipids, inflammatory markers, development and validation nomogram

Posted Date: August 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-50669/v1

Abstract

Background: to develop and validate a serum lipid and inflammatory marker model based on the nomogram for the prediction of stroke risk in rheumatoid arthritis patients.

Methods: This study was conducted among 313 rheumatoid arthritis with stroke patients and 1827 rheumatoid arthritis patients divided into develop and validation cohorts from the First Affiliated Hospital of China Medical University during January 2011 to December 2018. Logistic regression analysis was used to create a nomogram of predictive model of stroke risk in rheumatoid arthritis patients, after comparing with other machine algorithms. The performance of the nomogram was evaluated by discrimination, calibration and decision curve analysis, also compared with the Framingham Risk Score in predicting stroke in rheumatoid arthritis patients.

Results: the nomogram was performed by logistic regression algorithm, and predictors of which included the stratifications of sex, age, systolic blood pressure, C-reactive protein, erythrocyte sedimentation rate, total cholesterol, low density lipoprotein cholesterol and the distribution of being accompanied with hymed, diabetes, atrial fibrillation and coronary heart disease history, which exhibited a well goodness fit and a good agreement. The analysis with area under the curve, the net reclassification index, the integrated discrimination improvement and clinical use, suggested that this is an easy-to-use nomogram compared with the Framingham Risk Score.

Conclusion: This study presents a risk nomogram that incorporates the traditional risk factors, serum lipids and inflammatory markers which can be used to predict stroke in rheumatoid arthritis patients.

Introduction

Rheumatoid arthritis (RA), a globally distributed disease, is one of the leading causes of work disability, with morbidity rates of 0.32%~0.36% in China, where the highest is 0.5% in northeast China[1]. RA can be classified into two major types of pathological lesions, namely, synovitis and pannus[2]. Synovitis is the basic damage to a joint, while pannus involves pathological lesions in extra-articular joints, which can worse the prognosis of RA and increase the risk of cardiovascular disease (CVD) and other diseases, such as stroke.

Elevated inflammatory levels and lipid abnormalities in RA patients are independent risk factors of atherosclerosis, stroke and other cardiovascular diseases[3-6]. We postulate that the risk of stroke among RA patients may be closely related to elevated levels of erythrocyte sedimentation rate (ESR), high density lipoprotein cholesterol (HDL)[7], total cholesterol (TC), triglyceride (TG), anti-cyclic citrullinated peptide antibody (anti-CCP)[8], low density lipoprotein cholesterol (LDL), and C-reactive protein (CRP)[8-11]. Typically, in clinical practice, these objective and quantitative descriptors would be routinely detected in RA patients and available in the electronic medical records (EMR). However, these biomarkers are so common that clinicians often ignore clinical importance and even don’t know how to combine them together for clinical

decision. Since most chronic diseases are caused by several weak risk factors acting together, statistically combining their effects may produce a more robust prediction of risk than considering these factors independently. Risk prediction tools are increasingly common in clinical medicine[12-14], with well-known models, such as the *Framingham Risk Score* for predicting cardiovascular events[15, 16], used for determining clinical guidelines, thus, these tools are sometimes powerful enough to change clinical management and decision-making.

Our previous study had suggested that the elevated ESR, LDL levels, and much higher CRP levels ≥230mg/L were independent risk factors or RA patients in developing stroke in our study population[17]. Finally, in this study, we aimed to developed and validated a risk nomogram of predictors that incorporates serum lipids and inflammatory markers for the individual prediction of stroke in RA patients, based on the Framingham Risk Score.

1. Patients And Methods

1.1 Selection of the subjects

A total of 8389 RA patients, with 9.04% prevalence of stroke (758 RA with stroke patients) were filtered from the inpatient department of rheumatology and immunology of the First Affiliated Hospital of China Medical University during January 2011 to December 2018 in this study. According to the inclusion and exclusion criteria, 313 RA with stroke patients and1827 RA patients were included into studies (shown as figure s1, see in additional file 1), all aged 18 and older, from EMR of a third-senior hospital in Liaoning province. EMRs were classified and coded by using the International Classification of Diseases Tenth Revision of the Beijing clinical version (RA: M05.x~06.x; stroke (ischemic and hemorrhagic): I60 I60.1-I60.0 I61 I61.0-I61.9 I69.0 I69.1 I63 I63.0-I63.9 I69.3). The study conformed to the principles outlined in the Declaration of Helsinki and was conducted under the guidelines of the Institutional Review Board approved by the ethics committee of the medical science research institute of the First Affiliated Hospital, China Medical University (approval number: AF-SOP-07-1.0-01). All subjects gave written informed consent for the use of their data.

This study criteria included the American College of Rheumatology 1987/2010[18] for RA and the CVD criteria adopted at the Fourth Academic Conference by the Chinese Neuroscience Society in 1995 [19] for stroke. RA with stroke cohort, which patients should meet the follow inclusion criteria: ⑴) the patients conform to the above stroke and RA diagnostic criteria; ⑵) Based on the first record time of RA and stroke in the EMRs, if the record of stroke was later than that of RA, we believed that the patient has developed stroke after being diagnosed with RA and included in the RA with stroke cohort. ⑶) the patients have been detected at least one time laboratory assessments when they were in hospital at the first time (i.e, serum inflammatory-, antibody-, complement-, lipid-assays); ⑷) over 18 years old. RA cohort (without stroke), which patients should met the follow inclusion criteria: ⑴) conform to the above RA diagnostic criteria; ⑵) over 18 years old. RA with stroke cohort and RA cohort were excluded if they met the following criteria: ⑴) patients who still suffered from other connective tissue diseases, including systemic lupus

erythematosus, scleroderma, dry syndrome and vasculitis;) RA patients with coexisting ankylosing spondylitis and gout arthritis. Finally, we selected 70% of the RA with stroke and RA patients as the develop cohort randomly, with the rest comprising the validation cohort[20].

1.2 Data collection

All of data were screened from EMR, mainly including personal information, such as age, gender, height and weight, metabolic indices including serum TC, TG, LDL, HDL and fasting blood-glucose (FBG), and serologic profiles including CRP, ESR, rheumatoid factor (RF), complement3 (C3), complement4 (C4) and anticyclic citrullinated peptide (anti-CCP) antibodies, and coronary heart disease (CHD), atrial fibrillation (AF), left ventricular hypertrophy (LVH), cardiovascular disease (CVD) history records. The medication history in record was also included, that is, hypotensive medicine (hy-med), biologic disease modifying anti-rheumatic drugs (Bio-med). All laboratory tests were carried out using overnight fasting venous blood samples and conducted with clinical standard operating procedures for inspection items. In addition, when the results of multiple laboratory tests at different time points were assessed during the initial data filtering, the first laboratory test results were selected at first admission due to stroke among RA patients (RA with stroke cohort) and selected at first admission of RA patients without stroke (RA cohort).

1.3 Statistical analysis.

All reported statistical significance levels were set at 0.05 with two-sided. The categorical data were expressed as percentages by cohort, and some continuous predictors (ie., age, SBP and CRP) were categorized after assessed using consensus approaches or guidelines and previously published articles. In addition, the absence of some features in clinical medical records was inevitable (which accounted for less 20% in this study), and we used multiple imputation, based on 5 replications and a chained equation approach method (predictive mean matching, PMM for quantitative data, and linear regression for categorical data), to account for missing data in SPSS 23.0 software. Adopting Chi-Square test or Fisher exact test to compare the difference in participant characteristics between develop cohort and validation cohort. The univariable association between RA with stroke group and RA group with serum biochemical marker levels was assessed in the develop cohort using univariate LR, based on variables associated with stroke which were assessed by clinical importance, scientific knowledge, and predictors identified in previously published articles to develop the risk model of RA patients with stroke[21], then validated in the validation cohort.

1.3.1. Developing model between the RA with stroke group and RA group in the develop cohort

The model was built by binary logistic regression (LR) as a simple model with unadjusted, and a complex model adjusted by sex and age, considering the disparity between male and female morbidity in RA patients and the obvious aging of stroke patients. Further, to provide the clinician with a quantitative tool to predict the individual probability of stroke, we built the nomogram for the prediction of stroke in RA patients based on multivariable LR analysis in the develop cohort. All the analysis was conducted with R

software version 3.6.2 (packages mainly include rms, Hmisc, dca, PredictABEL. R packages. http: // www. Rproject. org) .

1.3.2. Comparing several machine learning models between the RA with stroke group and RA group in the develop and validation cohorts

Machine learning models are conducted based on scikit-learn which is an open source machine learning library, using Bayesian optimization method to implement algorithm optimization, and cross-validation method, N-folds=5, to complete algorithm evaluation during the optimization process. We used 6 kinds of machine algorithms running three 30-minute sessions, including LR, Support Vector Machine (SVM), Random Forest (RF), xgboost (XGB), gradient boosting decision tree (GBDT), k-Nearest Neighbors (KNN), to compare algorithms and evaluate the simple model and complex model in develop and validation cohorts respectively, which were compared by evaluation metrics as follows: accuracy, precision, recall, f1-score, balance error (ber).

1.3.3. Comparing the performance of developed models with Framingham Risk model and validation of which in the validation cohort

The performance of the nomogram of the model was assessed by discrimination and calibration. Calibration curves, the accuracy of point estimates of the LR function, accompanied with the Hosmer-Lemeshow test to assess if the model calibrated perfectly or not. The discrimination of the nomograms was evaluated using the Harrell's concordance index (C-index), the predictive accuracy for individual outcomes (discriminating ability), is equivalent to area under the curve (AUC), and compared among the Framingham Risk Score in predicting stroke and our prediction models. In addition, the calibration was calculated via a bootstrap method with 1000 resamples in the develop cohort. Internal validation was performed using the validation cohort. The LR formula formed in the develop cohort was applied to all patients of the validation cohort. The net reclassification index (NRI) indicates the proportion of patients correctly reclassified by a new model compared with an existing or standard model, while the integrated discrimination improvement (IDI) indicates the change in difference in average predicted probabilities between those who combined with stroke and those who did not in a new and existing model[22]. Furtherly, NRI and IDI between the Framingham Risk Score in predicting stroke and our prediction models were assessed based on low risk (0~20%), medium risk (20%~59%), high risk (60%~100%).

1.3.4. Clinical Use

In the develop cohort, DCA was conducted to evaluate the clinical usefulness of the nomogram by quantifying the net benefits at different threshold probabilities and was used to identify the predictive models with the best discriminative abilities[23]. In addition, net benefit was defined as the proportion of true positives minus the proportion of false positives, weighted by the relative harm of false-positive and false-negative results [24]. Simple and complex models were used to predict risk stratification of 1000 people with a bootstrap resample by the clinical impact curve.

2. Results

2.1 Baseline demographics and clinical characteristics.

There were 218 RA with stroke patients and 1136 RA patients included in develop cohort, and there were 95 RA with stroke patients and 486 RA patients included in the validation cohort. The clinicopathologic characteristics of the patients are listed in Table1. The baseline clinicopathologic data were similar between the develop and validation cohorts. As shown in Table 2, univariate LR analysis of RA patients developing to stroke indicated that the stratifications of sex, age, SBP, CRP, ESR, TC, LDL and the distribution of being accompanied with hy-med, Diabetes, AF, CVD and CHD history, were significantly different between RA with stroke group and RA group ($P < 0.05$) in the develop cohort.

2.2 Development of an Individualized Prediction Model

All variables used in this analysis were based on table 2, which were carefully chosen to ensure parsimony and practicality of the final model (noted in methods section), finally identified the following 10 variables that had the strongest association with stroke risk: including the stratifications of sex, age, SBP, CRP, ESR, TC, LDL and the distribution of being accompanied with hy-med, diabetes, AF and CHD history. Based on the final complex model of multivariate analysis shown in **figure**1, sex (0.63[0.45-0.91]), age (for 66-79 vs.18-65, 1.2.01 [1.38-2.93], for ≥80 vs. 18-65, 4.20 [2.68-6.58]), AF (2.27[1.08-4.68]), CHD (2.49[1.70-3.64]), hy-med (2.08[1.44-3.00]), SBP stratification (for 140-159 vs.<120, 1.64 [1.04-2.61], for 160-179 vs. <120, 2.44 [1.13-5.20]), CRP stratification (for ≥64.32 vs. <9.06, 1.67 [1.05-2.68]), ESR stratification (for ≥84.8 vs. <29, 1.64 [1.03-2.62]), TC stratification (for ≥6.2 vs.<5.2, 0.35 [0.17-0.70], LDL stratification (for 3.4- 4.1 vs.<3.4, 4.45 [2.35-8.68], for ≥4.1 vs.<3.4, 4.22 [1.66-10.69]) were independently associated with stroke among RA patients, the final simple model of multivariate analysis was also shown in **figure**1. Comprehensively considering vital indicators of machine learning models, shown as **figure**2 and table s1(see in additional file 1), the results indicated that the effect of LR model building the nomogram comparing with the other machine learning models were at the same level even better, and the LR algorithm was effective and feasible for the prediction of current data, simultaneously indicating the complex model was better than the simple model. Finally, the nomogram was performed based on the complex model incorporating the above independent predictors, shown in **figure**3, which showed the score of the influencing factor levels, the personal total cumulative score, and the predicted risk value of the individual outcome event for RA patients.

2.3 Assessing the Performance and Internal Validation of the Stroke Nomogram

Figure4 depicting the flexible calibration curve, indicated good agreement between prediction and observation in the develop and validation cohorts (slope=1, intercept=0 all with simple and complex model). Furtherly, we comprehensively assessed and compared the performance of developed models with Framingham Risk model, shown as table3. The Hosmer-Lemeshow test yielded a non-significant statistic, suggesting that there was no departure from perfect fit in the simple model vs. the complex model in the develop cohort ($P = 0.385$ vs. 0.097) and revealing a good agreement in the probability of

stroke between prediction and observation in the develop cohort. Based on the AUCs, the discrimination performance of models, the complex model (AUC, 95% CI: 0.784 [0.750- 0.818], P<0.001) had a better diagnostic value than the simple model (AUC, 95% CI: 0.747 [0.711-0.784], P<0.001), P=0.0016 and had no significant statistic difference comparing with the Framingham Risk model (AUC, 95% CI: 0.808 [0.778-0.893], P<0.001), P=0.0631 in predicting the development of stroke in RA patients. In addition, the simple model and complex model improved the correctly reclassified based on NRI (11.59[2.90, 20.29], 20.30[12.54, 28.05] separately) and IDI (1.71[-0.77, 4.18], 5.65[3.41, 7.88] separately) comparing with the Framingham Risk model in predicting stroke.

2.4 Clinical Use

The DCA for the nomogram of the develop cohort, presented in **figure s2** (see in additional file 1), indicated that if the threshold probability of a patient or doctor was about 15%, using the risk nomogram to predict stroke could add benefit compared to either the treat-all-patients scheme or the treat-none scheme. When the threshold value was 15%-55%, net benefit was comparable on the basis of the risk nomogram, suggesting that the benefit of the complex model (blue line) was higher than that in simple model (red line) in predicting the risk of stroke in RA patients.

3 Discussion

To our knowledge, this is the first prediction model for the risk of RA patients developing stroke. By using EMR data from hospitalized patients in northern China, we developed and validated a nomogram for the individualized prediction of stroke in RA patients which incorporated several factors, sex, age, SBP, CRP, ESR, TC, LDL and the distribution of being accompanied with hy-med, AF and CHD history, between the Framingham Risk Score in predicting stroke and our prediction models .

Prediction models use multiple predictors to estimate the absolute probability or risk that a certain outcome is present or will occur within a specific time period in an individual [25, 26]. Recently, some studies have taken advantage of prediction models to create multi-markers for clinical decisions, such as Yan-qi Huang, et al. with a radiomics nomogram that incorporates the radiomics signature, CT-reported lymph node status and clinical risk factors to preoperatively predict lymph node metastasis in patients with colorectal cancer [13]. The Framingham Risk Score for predicting stroke events was performed with the data followed for 10 years by using Cox proportional hazards regression model [27], while ours were aim to explore and utilize the data of Real World study. Our study evaluated six machine learning models, which suggested that LR algorithm performed well in the evaluation, also confirmed LR algorithm has a better generalization. Multivariable analyses that incorporate individual markers into marker panels have been embraced in recent studies, similarly, in our study, the model incorporating serum lipids, inflammatory markers and connecting multiple individual features demonstrated adequate discrimination in a develop cohort which was then improved in the validation cohort. Our prediction model also has a good calibration. While, it is unclear which of two models is more preferable[23]. By the Hosmer-Lemeshow, the developed models claimed a well goodness of fit. We didn't find the significant statistics difference

between complex model and Framingham Risk model by the AUC analysis, however evaluated the 20.30% [12.54, 28.05] patients correctly reclassified by complex model than Framingham Risk model with 5.65% [3.41, 7.88] of IDI. DCA assessment, in theory, can inform on model effectiveness, or which of several alternative models should be used [28]. To this end, we used DCA to address the heterogeneity across different institutions in the clinical data collection and further to select the best model. When the threshold probability of a patient or doctor was >15%, the higher net-benefit of the complex and simple model were superior to either the treat-all-patients scheme or the treat-none scheme, and this was best with the complex model when the threshold probability was 15%-55%. Therefore, the discrimination, calibration, NRI, IDI and clinical use measures, suggested that this easy-to-use nomogram can effectively predict the risk of stroke among RA patients. That is, our predictive model has strong clinical value for clinical decisions for RA patients.

Some previous studies identified a number of demographic and clinical characteristics which should affect the risk of RA patients developing stroke, mainly increased lipid metabolism levels, high inflammatory levels, and other traditional CVD risk factors [29-31]. However, these studies were conducted in Caucasians, lacking data from Asians. Meanwhile, these studies' results were not consistent, Zhang J et al.[29] supported the hypothesis that RA-related systemic inflammation played a role in determining cardiovascular risk and a complex relationship between LDL and cardiovascular risk, be alike that some studies suggested the "lipid paradox" of LDL[7, 32], and some other studies [9, 33] suggested that TC, LDL, TG levels were useful and limited for prediction of stroke in RA than in the general population. Our results emphasized the effects of serum TC, LDL levels in predicting the development of stroke among RA patients, reserved the traditional risk factors (hy-med, AF and CHD history) confirmed in previous studies about the Framingham Risk Score[16, 27, 34]. For this particular Population of RA, our findings underscored the important contribution of systemic inflammation to the development of stroke in RA patients, founding CRP and ESR inflammatory factors to be independent risk factors for stroke in RA patients as well as confirmed in published papers[4, 35]. Notably, considering the above factors played critical roles in RA patients, several weak risk factors combining their effects may produce a more reliable prediction of risk than the consideration of a single risk factor. Thus, in order to explain the comprehensive effect above clinical factors, we developed the nomogram incorporating several independent predictors to inform individual patient care to prevent stroke in RA patients.

Our study still had several limitations. First, the data used in the clinical prediction model was derived from a single center in a hospital and the design had better be a randomized controlled prospective trial owning to a gap on the classification of stroke in the RA with stroke patients even if the stroke was present before the onset of RA. Then, the model should be externally validated at additional sites even though we computed the C-index for the prediction nomogram via bootstrapping validation and assessing NRI of a bootstrap resample with 1,0000 people for multiple validation. Thirdly, recently there is increasing evidence of an association between RA and stroke, it is unclear if simply building a model that applies the traditional risk factors, serum lipids and inflammatory markers to predict outcomes ideally. This disagrees with the theory that scientific inferences should be based on evidence from as many

sources and individuals as possible, an accepted principle that is often used in intervention studies. Nevertheless, well-designed randomized clinical trials or cohort following investigation are necessary to further comparing the predictive power of our model with FRS for stroke in patients with RA. Clinical prediction models are increasingly used to complement clinical reasoning and decision-making in modern medicine. The adoption of such models by professionals can guide their decision making and improve patient outcomes and the cost effectiveness of care. Prediction models are not developed to replace doctors, but to provide objective estimates of health outcome risks for both individuals (patients) and healthcare providers to assist their subjective interpretations, intuitions and guidelines [36, 37].

In conclusion, our study presented an effective nomogram that incorporates the traditional risk factors, serum lipids and inflammatory markers that can be used to predict stroke in patients with RA. Our study can provide a theoretical basis for improving the prognosis of RA patients and preventing the onset of stroke.

Supplementary Information

Additional file 1: Figure S1. Flow chart showing the selection process of study participants. (Abbreviation: RA, rheumatoid arthritis; EMR, the electronic medical record.)

Additional file 1: Figure S2. Decision curve analysis for serum lipids, inflammatory markers, and serological status in rheumatoid arthritis and stroke patients of the simple and complex model in the primary cohorts. The y-axis represents the net benefit, the x-axis represents the risk threshold of stroke in RA patients. The red line represents the nomogram of predictors in simple model. The blue line represents the complex model with addition of sex and age. The gray line represents the assumption that all patients have stroke. Thin black line represents the assumption that no RA patients developing stroke. The net benefit was calculated by subtracting the proportion of all patients who are false positive from the proportion who are true positive, weighting by the relative harm of forgoing treatment compared with the negative consequences of an unnecessary treatment.

Additional file 1: Table S1 Model evaluation (F1-score) results based on the number of features across 6 models. Abbreviation: GBDT, gradient boosting decision tree; KNN, k-Nearest Neighbors; LR, logistic regression; RF, Random Forest; XGB, xgboost; SVM, Support Vector Machine.

Abbreviations

RA, Rheumatoid Arthritis; EMR, electronic medical records; OR (95%CI), odd ratios, 95% confidence intervals; SBP, Systolic Blood Pressure; CHD, coronary heart disease; AF, atrial fibrillation; LVH, left ventricular hypertrophy ; CVD, cardiovascular disease; hy-med, hypotensive medicine; Bio-med, biologic disease modifying anti-rheumatic drugs; CCP^+, positive anti-cyclic citrullinated peptide antibody; RF^+, positive rheumatoid factor; CRP,C-reactive protein; EMR, electronic medical records; ESR, erythrocyte sedimentation rate; C3, complement 3; C4, complement 4; FBG, fasting blood glucose; TC, total

cholesterol; LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol; TG, triglycerides. GBDT, gradient boosting decision tree; KNN, k-Nearest Neighbors; LR, logistic regression; RF, Random Forest; XGB, xgboost; SVM, Support Vector Machine. AUC, the area under the receiver operating characteristic curve; NRI, the net reclassification index; IDI, the integrated discrimination improvement, NA, not available; ref., the reference level.

Declarations

Acknowledgements

We thank Yiducloud (Beijing) Technology Ltd for supporting part of the data extraction and processing also thank Neusoft for supporting analysis and evaluation by leveraging their RealMedSci that is an automatic medical analysis platform.

Authors' contributions

All the authors listed have reviewed the final version of the manuscript and approved it for publication. FRX analyzed and interpreted the data and drafted the manuscript. LYF designed the study. FRX, TTW, HNL, BWY, and BQB collected the data. CLZ provided the technical support of RealMedSci, an automatic medical analysis platform. LYF contributed to the interpretation of the results and critical revision of the manuscript for important intellectual content and approved the final version of the manuscript. All authors have read and approved the final manuscript. FRX and LYF were the study guarantors.

Funding

This work was supported by grants from the Program of the National Natural Science Foundation of China [81673246]

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available for ethical and privacy reasons but are available from corresponding author on reasonable request.

Ethics approval and consent to participate

The study conformed to the principles outlined in the Declaration of Helsinki and was conducted under the guidelines of the Institutional Review Board approved by the ethics committee of the medical science research institute of the First Affiliated Hospital, China Medical University (approval number: AF-SOP-07-1.0-01). All subjects gave written informed consent for the use of their data.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. FengchunZhang Zhanguo Li. *Rheumatoid Arthritis.* People's Medical Publishing House, 1 th ed.,2009.
2. Yongjian Xu Junbo Ge. *Internal Medicine.* People's Medical Publishing House, 8 th ed.,2013.
3. S. J. Wiseman, S. H. Ralston, and J. M. Wardlaw. Cerebrovascular Disease in Rheumatic Diseases: A Systematic Review and Meta-Analysis. *Stroke* 2016;47: 943-950.
4. R. H. Mackey, L. H. Kuller, K. D. Deane, B. T. Walitt, Y. F. Chang, V. M. Holers, W. H. Robinson, R. P. Tracy, M. A. Hlatky, C. B. Eaton, S. Liu, M. S. Freiberg, M. B. Talabi, E. B. Schelbert, and L. W. Moreland. Rheumatoid Arthritis, Anti-Cyclic Citrullinated Peptide Positivity, and Cardiovascular Disease Risk in the Women's Health Initiative. *Arthritis Rheumatol* 2015;67: 2311-2322.
5. T. H. Liou, S. W. Huang, J. W. Lin, Y. S. Chang, C. W. Wu, and H. W. Lin. Risk of Stroke in Patients with Rheumatism: A Nationwide Longitudinal Population-Based Study. *Sci Rep* 2014;4: 5110.
6. I. Navarro-Millan, S. Yang, S. L. DuVall, L. Chen, J. Baddley, G. W. Cannon, E. S. Delzell, J. Zhang, M. M. Safford, N. M. Patkar, T. R. Mikuls, J. A. Singh, and J. R. Curtis. Association of Hyperlipidaemia, Inflammation and Serological Status and Coronary Heart Disease among Patients with Rheumatoid Arthritis: Data from the National Veterans Health Administration. *Ann Rheum Dis* 2016;75: 341-347.
7. D. Dursunoglu, H. Evrengul, B. Polat, H. Tanriverdi, V. Cobankara, A. Kaftan, and M. Kilic. Lp(a) Lipoprotein and Lipids in Patients with Rheumatoid Arthritis: Serum Levels and Relationship to Inflammation. *Rheumatol Int* 2005;25: 241-245.
8. J. Lindhardsen, G. H. Gislason, O. Ahlehoff, O. R. Madsen, and P. R. Hansen. Increased Risk of Stroke and Atrial Fibrillation in Rheumatoid Arthritis - a Nationwide Cohort Study. *Eur Heart J* 2010;31: 156-156.
9. A. G. Semb, T. K. Kvien, A. H. Aastveit, I. Jungner, T. R. Pedersen, G. Walldius, and I. Holme. Lipids, Myocardial Infarction and Ischaemic Stroke in Patients with Rheumatoid Arthritis in the Apolipoprotein-Related Mortality Risk (Amoris) Study. *Ann Rheum Dis* 2010;69: 1996-2001.
10. J. Zhang, L. Chen, E. Delzell, P. Muntner, W. B. Hillegass, M. M. Safford, I. Y. Millan, C. S. Crowson, and J. R. Curtis. The Association between Inflammatory Markers, Serum Lipids and the Risk of Cardiovascular Events in Patients with Rheumatoid Arthritis. *Ann Rheum Dis* 2014;73: 1301-1308.
11. Goodson NJ, Symmons DP, and Scott DG. Baseline Levels of Creactive Protein and Prediction of Death from Cardiovascular Disease in Patients with Inflammatory Polyarthritis: A Ten-Year Follow up Study of a Primary Care-Based Inception Cohort. *Arthritis Rheum* 2005;52: 3-9.
12. Chunjuan Wang, Yingying Yang, Yuesong Pan, Xiaoling Liao, Xiaochuan Huo, Zhongrong Miao, Yongjun Wang, and Yilong Wang. Validation of the Simplified Stroke-Thrombolytic Predictive Instrument to Predict Functional Outcomes in Chinese Patients. *Stroke* 2018;49: 2773-2776.

13. Y. Q. Huang, C. H. Liang, L. He, J. Tian, C. S. Liang, X. Chen, Z. L. Ma, and Z. Y. Liu. Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer. *J Clin Oncol* 2016;34: 2157-2164.
14. H. Y. Kim, E. J. Jang, B. Park, T. Y. Kim, S. A. Shin, Y. C. Ha, and S. Jang. Development of a Korean Fracture Risk Score (Kfrs) for Predicting Osteoporotic Fracture Risk: Analysis of Data from the Korean National Health Insurance Service. *PLoS One* 2016;11: e0158918.
15. ScD Nancy R. Cook, Nina P. Paynter, PhD, Charles B. Eaton, MD, JoAnn E. Manson, MD, DrPH, Lisa W. Martin, MD, Jennifer G. Robinson, MD, MPH, Jacques E. Rossouw, MD, Sylvia Wassertheil-Smoller, PhD, Paul M. Ridker, MD. Comparison of the Framingham and Reynolds Risk Scores for Global Cardiovascular Risk Prediction in the Multiethnic Women's Health Initiative. *Circulation* 2012;125: 9.
16. Xuebiao Weia Open Access Yaowang Lina, Anping Cai, Xing Yang, Yingling Zhou and Danqing Yu. Framingham Risk Score for the Prediction of Coronary Artery Disease in Patients with Chronic Rheumatic Heart Disease. *Diagnosis (Berl)* 2014;1: 6.
17. F. Xin, L. Fu, H. Liu, Y. Xu, T. Wei, and M. Chen. Exploring Metabolic and Inflammatory Abnormalities in Rheumatoid Arthritis Patients Developing Stroke Disease: A Case-Control Study Using Electronic Medical Record Data in Northern China. *Clin Rheumatol* 2019;38: 1401-1411.
18. Edworthy Sm Arnett Fc, Bloch Da, et al. The American Rheumatism Association 1987 Revised Criteria for the Classification of Rheumatoid Arthritis. *Arthritis Rheum* 1988;31: 315-324.
19. Chinese Neuroscience Association. Diagnosis of Various Cerebrovascular Diseases(1995. *Journal of clinical and experimental medicine* 2013;12: 1.
20. Z. Lei, J. Li, D. Wu, Y. Xia, Q. Wang, A. Si, K. Wang, X. Wan, W. Y. Lau, M. Wu, and F. Shen. Nomogram for Preoperative Estimation of Microvascular Invasion Risk in Hepatitis B Virus-Related Hepatocellular Carcinoma within the Milan Criteria. *JAMA Surg* 2016;151: 356-363.
21. Y. Kim, G. A. Margonis, J. D. Prescott, T. B. Tran, L. M. Postlewait, S. K. Maithel, T. S. Wang, D. B. Evans, I. Hatzaras, R. Shenoy, J. E. Phay, K. Keplinger, R. C. Fields, L. X. Jin, S. M. Weber, A. I. Salem, J. K. Sicklick, S. Gad, A. C. Yopp, J. C. Mansour, Q. Y. Duh, N. Seiser, C. C. Solorzano, C. M. Kiernan, K. I. Votanopoulos, E. A. Levine, G. A. Poultsides, and T. M. Pawlik. Nomograms to Predict Recurrence-Free and Overall Survival after Curative Resection of Adrenocortical Carcinoma. *JAMA Surg* 2016;151: 365-373.
22. J. L. Lund, T. M. Kuo, M. A. Brookhart, A. M. Meyer, A. F. Dalton, C. E. Kistler, S. B. Wheeler, and C. L. Lewis. Development and Validation of a 5-Year Mortality Prediction Model Using Regularized Regression and Medicare Data. *Pharmacoepidemiol Drug Saf* 2019;28: 584-592.
23. A. J. Vickers, and E. B. Elkin. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med Decis Making* 2006;26: 565-574.
24. G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (Tripod): The Tripod Statement. *Bmj* 2015;350: g7594.
25. J. Ranstam, J. A. Cook, and G. S. Collins. Clinical Prediction Models. *Br J Surg* 2016;103: 1886.

26. M. A. Hadi, and S. J. Closs. Applications of Mixed-Methods Methodology in Clinical Pharmacy Research. *Int J Clin Pharm* 2016;38: 635-640.
27. R. B. D'Agostino, P. A. Wolf, A. J. Belanger, and W. B. Kannel. Stroke Risk Profile: Adjustment for Antihypertensive Medication. The Framingham Study. *Stroke* 1994;25: 40-43.
28. M.; Glasziou Hunink, P.; Siegel, J. *Decision-Making in Health and Medicine: Integrating Evidence and Values.* New York: Cambridge University Press, ed.,2001.
29. J. Zhang, L. Chen, E. Delzell, P. Muntner, W. B. Hillegass, M. M. Safford, I. Y. Millan, C. S. Crowson, and J. R. Curtis. The Association between Inflammatory Markers, Serum Lipids and the Risk of Cardiovascular Events in Patients with Rheumatoid Arthritis. *Ann Rheum Dis* 2014;73: 1301-1308.
30. P. H. Dessein, B. I. Joffe, and A. E. Stanwix. Inflammation, Insulin Resistance, and Aberrant Lipid Metabolism as Cardiovascular Risk Factors in Rheumatoid Arthritis. *J Rheumatol* 2003;30: 1403-1405.
31. L. R. Baghdadi, R. J. Woodman, E. M. Shanahan, and A. A. Mangoni. The Impact of Traditional Cardiovascular Risk Factors on Cardiovascular Outcomes in Patients with Rheumatoid Arthritis: A Systematic Review and Meta-Analysis. *PLoS One* 2015;10: e0117952.
32. E. Myasoedova, C. S. Crowson, H. M. Kremers, V. L. Roger, P. D. Fitz-Gibbon, T. M. Therneau, and S. E. Gabriel. Lipid Paradox in Rheumatoid Arthritis: The Impact of Serum Lipid Measures and Systemic Inflammation on the Risk of Cardiovascular Disease. *Ann Rheum Dis* 2011;70: 482-487.
33. Namrata Dhillon, and Kimberly Liang. Prevention of Stroke in Rheumatoid Arthritis. *Current treatment options in neurology* 2015;17: 356-356.
34. D. H. O'Leary, J. F. Polak, R. A. Kronmal, T. A. Manolio, G. L. Burke, and S. K. Wolfson, Jr. Carotid-Artery Intima and Media Thickness as a Risk Factor for Myocardial Infarction and Stroke in Older Adults. Cardiovascular Health Study Collaborative Research Group. *The New England journal of medicine* 1999;340: 14-22.
35. van Leuven SI, Franssen R, and Kastelein JJ. Systemic Inflammation as a Risk Factor for Atherothrombosis. *Rheumatology* 2008; 47: 3-7.
36. K. G. Moons, P. Royston, Y. Vergouwe, D. E. Grobbee, and D. G. Altman. Prognosis and Prognostic Research: What, Why, and How? *Bmj* 2009;338: b375.
37. T. G. McGinn, G. H. Guyatt, P. C. Wyer, C. D. Naylor, I. G. Stiell, and W. S. Richardson. Users' Guides to the Medical Literature: Xxii: How to Use Articles About Clinical Decision Rules. Evidence-Based Medicine Working Group. *Jama* 2000;284: 79-84.

Tables

Table 1 Participant Characteristics of both primary cohort and validation cohort

Variables	Cohort No.(%)		χ^2	P
	Train Cohort (1354)	Test Cohort (581)		
RA with stroke	218 (16.10)	95 (16.35)	0.019	0.891
Sex, female	1021 (75.41)	428 (73.67)	0.655	0.419
Age, year				
18-65	757 (55.91)	308 (53.01)	1.386	0.500
66-79	426 (31.46)	194 (33.39)		
≥80	171 (12.63)	79 (13.60)		
SBP, mmHg				
120	422 (31.17)	194 (33.39)	7.511	0.111
120-139	592 (43.72)	240 (41.31)		
140-159	287 (21.20)	114 (19.62)		
160-179	46 (3.40)	24 (4.13)		
≥180	7 (0.52)	9 (1.55)		
Smoking	176 (13.00)	84 (14.46)	0.744	0.388
Diabetes	190 (14.03)	97 (16.70)	2.282	0.131
CHD	208 (15.36)	104 (17.90)	1.937	0.164
AF	40 (2.95)	17 (2.93)	0.001	0.973
LVH*	1 (0.07)	1 (0.17)	-	0.510
CVD	436 (32.20)	198 (34.08)	0.651	0.420
hy-med	240 (17.73)	153 (26.33)	18.615	<0.001
Bio	22 (1.62)	12 (2.07)	0.457	0.499
CCP^+	827 (61.08)	337 (58.00)	1.604	0.205
RF^+	908 (67.06)	381 (65.58)	0.403	0.526
CRP, mg/L				
10	319 (23.56)	157 (27.02)	3.769	0.152
≥9.06nd64.32	670 (49.48)	287 (49.40)		
≥64.32	365 (26.96)	137 (23.58)		
ESR, mm/H				
29	351 (25.92)	137 (23.58)	2.368	0.306
≥29nd84.80	662 (48.89)	280 (48.19)		
≥84.8	341 (25.18)	164 (28.23)		
C3, g/L				
0.95	345 (25.48)	155 (26.68)	0.890	0.641
≥0.95nd1.34	673 (49.7)	293 (50.43)		
≥1.34	336 (24.82)	133 (22.89)		
C4, g/L				
0.18	378 (27.92)	146 (25.13)	5.450	0.066
≥0.18nd0.28	656 (48.45)	315 (54.22)		
≥0.28	320 (23.63)	120 (20.65)		
FBG, mmol/L				
4.84	336 (24.82)	154 (26.51)	2.052	0.358
≥4.84nd6.33	670 (49.48)	295 (50.77)		
≥6.33	348 (25.70)	132 (22.72)		
TC, mmol/L				
5.2	1131 (83.53)	479 (82.44)	6.259	0.044
≥5.2nd66.2	163 (12.04)	61 (10.50)		
≥6.2	60 (4.43)	41 (7.06)		
LDL, mmol/L				
3.4	1106 (81.68)	463 (79.69)	6.278	0.043
≥3.4nd4.1	182 (13.44)	73 (12.56)		
≥4.1	66 (4.87)	45 (7.75)		
HDL, mmol/L				
≥1.55	125 (9.23)	46 (7.92)	4.119	0.128
≥1.04nd1.55	570 (42.10)	273 (46.99)		
1.04	659 (48.67)	262 (45.09)		
TG, mmol/L				
1.7	1113 (82.20)	460 (79.17)	2.697	0.260
≥1.7nd2.3	139 (10.27)	73 (12.56)		
≥2.3	102 (7.53)	48 (8.26)		

Data are represented as number and proportion, statistics were calculated by Chi-Square test; *: statistics were calculated by Fisher exact test ; **Abbreviation**: RA, Rheumatoid Arthritis; SBP, Systolic Blood Pressure; CHD, coronary heart disease; AF, atrial fibrillation; LVH, left ventricular hypertrophy ;

CVD, cardiovascular disease; hy-med, hypotensive medicine; Bio-med, biologic disease modifying anti-rheumatic drugs; CCP+, positive anti-cyclic citrullinated peptide antibody; RF+, positive rheumatoid factor; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; C3, complement 3; C4, complement 4; FBG, fasting blood glucose; TC, total cholesterol; LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol; TG, triglycerides.

Table 2 Univariate Logistic Regression Analysis of RA patients developing to stroke in the Training Cohort

Variables	RA (1136)	RA with stroke (218)	*OR (95%CI)*	*P*
Sex, female vs. male	873 (76.85)	148 (67.89)	0.64 (0.46-0.88)	0.005
Age, year				
66-79 vs. 18-65	336 (29.58)	90 (41.28)	2.90 (2.05-4.10)	<0.001
≥80 vs. 18-65	107 (9.42)	64 (29.36)	6.48 (4.33-9.68)	<0.001
SBP, mmHg				
120-139 vs.<120	503 (44.28)	89 (40.83)	1.45 (0.99-2.12)	0.057
140-159 vs.<120	224 (19.72)	63 (28.90)	2.30 (1.52-3.48)	<0.001
160-179 vs.<120	29 (2.55)	17 (7.80)	4.79 (2.45-9.39)	<0.001
≥180 vs.<120	4 (0.35)	3 (1.38)	6.13 (1.33-28.25)	<0.001
Smoking	145 (12.76)	31 (14.22)	1.13 (0.75-1.72)	0.558
Diabetes	142 (12.50)	48 (22.02)	1.98 (1.37-2.85)	<0.001
CVD*	218 (19.19)	218 (100)	-	<0.001
CHD	130 (11.44)	78 (35.78)	4.31 (3.09-6.01)	<0.001
AF	22 (1.94)	18 (8.26)	4.56 (2.40-8.65)	<0.001
LVH*	0 (0)	1 (0.46)	-	0.161
hy-med	164 (14.44)	76 (34.86)	3.17 (2.29-4.39)	<0.001
Bio	21 (1.85)	1 (0.46)	0.25 (0.03-1.83)	0.236
CCP^+	698 (61.44)	129 (59.17)	0.91 (0.68-1.22)	0.529
RF^+	766 (67.43)	142 (65.14)	0.90 (0.67-1.24)	0.510
CRP, mg/L				0.007
≥9.06 and 64.32 vs. 10	570 (50.18)	100 (45.87)	1.19 (0.81, 1.76)	
≥64.32 vs. 10	288 (25.35)	77 (35.32)	1.81 (1.20-2.74)	
ESR, mm/H				0.037
≥29 and 84.80 vs. 29	553 (48.68)	109 (50)	0.58 (0.38-0.88)	
≥84.8 vs. 29	275 (24.21)	66 (30.28)	0.82 (0.59-1.15)	
C3, g/L				0.516
≥0.95 and 1.34 vs. 0.95	567 (49.91)	106 (48.62)	0.85 (0.61-1.20)	
≥1.34 vs. 0.95	286 (25.18)	50 (22.94)	0.80 (0.53-1.20)	
C4, g/L				0.786
≥0.18 and 0.28 vs. 0.18	550 (48.42)	106 (48.62)	0.95 (0.67-1.33)	
≥0.28 vs. 0.18	272 (23.94)	48 (22.02)	0.87 (0.58-1.30)	
FBG, mmol/L				0.652
≥4.84 and 6.33 vs. 4.84	556 (48.94)	114 (52.29)	1.12 (0.78-1.60)	
≥6.33 vs. 4.84	296 (26.06)	52 (23.85)	0.96 (0.63-1.46)	
TC, mmol/L				0.038
≥5.2 and 66.2 vs. 5.2	143 (12.59)	20 (9.17)	0.73 (0.45-1.20)	
≥6.2 vs. 5.2	44 (3.87)	16 (7.34)	1.90 (1.05-3.43)	
LDL, mmol/L				0.004
≥3.4 and 4.1 vs. 3.4	138 (12.15)	44 (20.18)	1.87 (1.28-2.73)	
≥4.1 vs. 3.4	53 (4.67)	13 (5.96)	1.44 (0.77-2.70)	
HDL, mmol/L				0.774
≥1.04 and 1.55 vs.≥1.55	483 (42.52)	87 (39.91)	0.89 (0.53-1.50)	
1.04 vs.≥1.55	549 (48.33)	110 (50.46)	0.99 (0.60-1.66)	
TG, mmol/L				0.186
≥1.7 and 2.3 vs. 1.7	114 (10.04)	25 (11.47)	1.11 (0.70-1.77)	
≥2.3 vs. 1.7	92 (8.10)	10 (4.59)	0.55 (0.28-1.08)	

Data are represented as number and proportion. statistics were conducted by univariate Logistic Regression. Abbreviation: RA, Rheumatoid Arthritis; OR (95%CI), odd ratios, 95% confidence intervals; SBP, systolic blood pressure; CHD, coronary heart disease; AF, atrial fibrillation; LVH, left ventricular hypertrophy ; CVD, cardiovascular disease; hy-med, hypotensive medicine; Bio-med, biologic disease modifying anti-rheumatic drugs; CCP^+, positive anti-cyclic citrullinated peptide antibody; RF^+, positive rheumatoid factor; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; C3, complement 3; C4, complement 4; FBG, fasting blood glucose; TC, total cholesterol; LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol; TG, triglycerides.

Table 3 Apparent Performance and Internal Validation of the Stroke Nomogram

		Framingham Risk	Simple Model	Complex Model
Hosmer-Lemeshow test				
	χ^2	NA	8.517	13.456
	P	NA	0.385	0.097
AUC (95% CI)		0.808(0.778, 0.839)*&	0.747 (0.711, 0.784)*#	0.784 (0.750, 0.818)#&
NRI (95% CI)		ref.	11.59 (2.90, 20.29)	20.30 (12.54, 28.05)
IDI (95% CI)		ref.	1.71 (-0.77, 4.18)	5.65 (3.41, 7.88)

*: P=0.0013, statistically significant associations between Framingham risk and simple model; #: P=0.0016, statistically significant associations between complex model and simple model; &: P=0.0631, statistically significant associations between complex model and Framingham risk. Abbreviation: AUC, the area under the receiver operating characteristic curve; NRI, the net reclassification index; IDI, the integrated discrimination improvement, NA, not available; ref., the reference level.

Figures

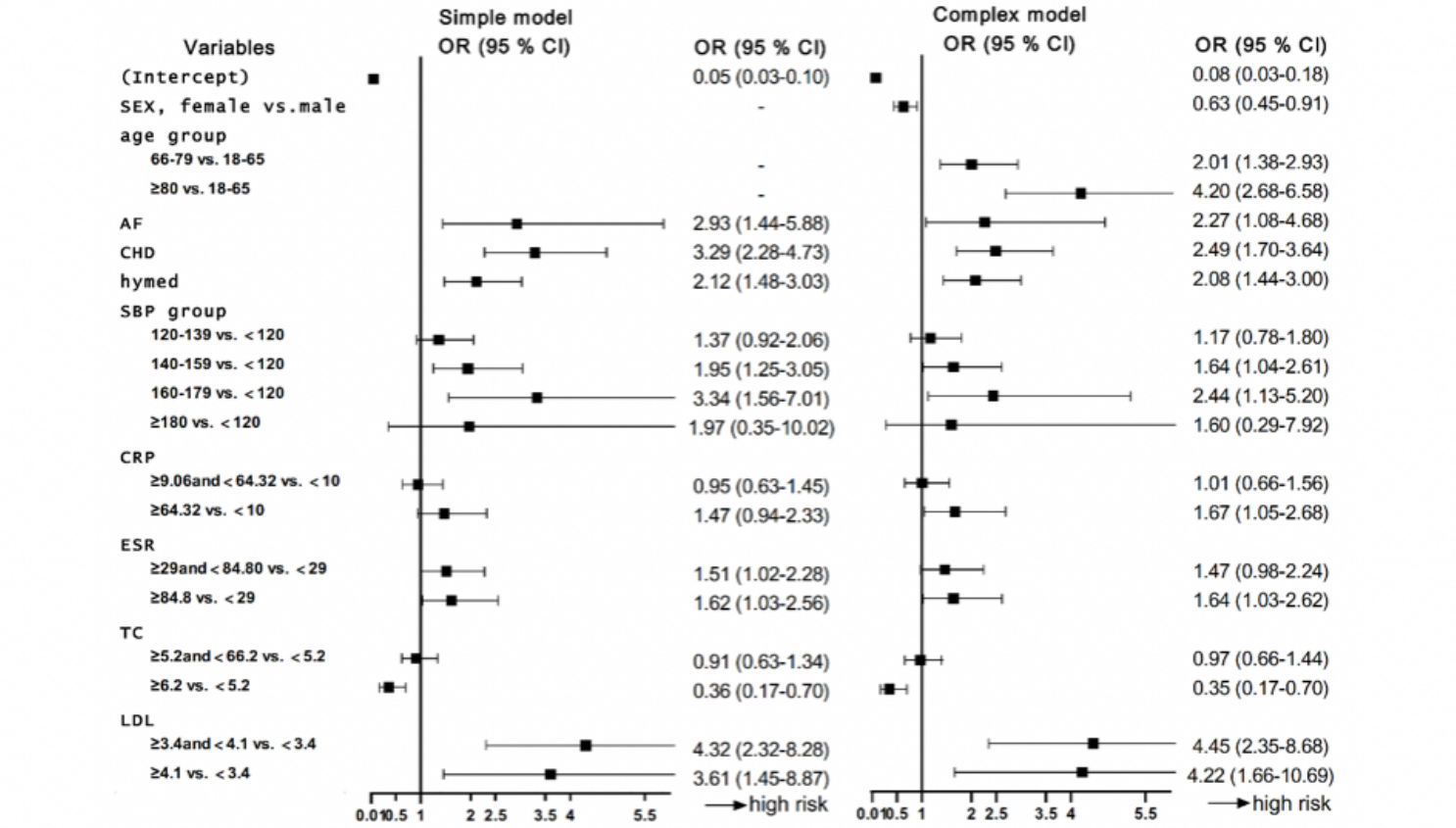

Figure 1

Multivariate logistic regression analysis for RA patients developing to stroke in training cohort. Abbreviation: SBP, systolic blood pressure; CHD, coronary heart disease; AF, atrial fibrillation; hy-med, hypotensive medicine; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; TC, total cholesterol; LDL, low-density lipoprotein cholesterol; OR (95%CI), odd ratios, 95% confidence intervals.

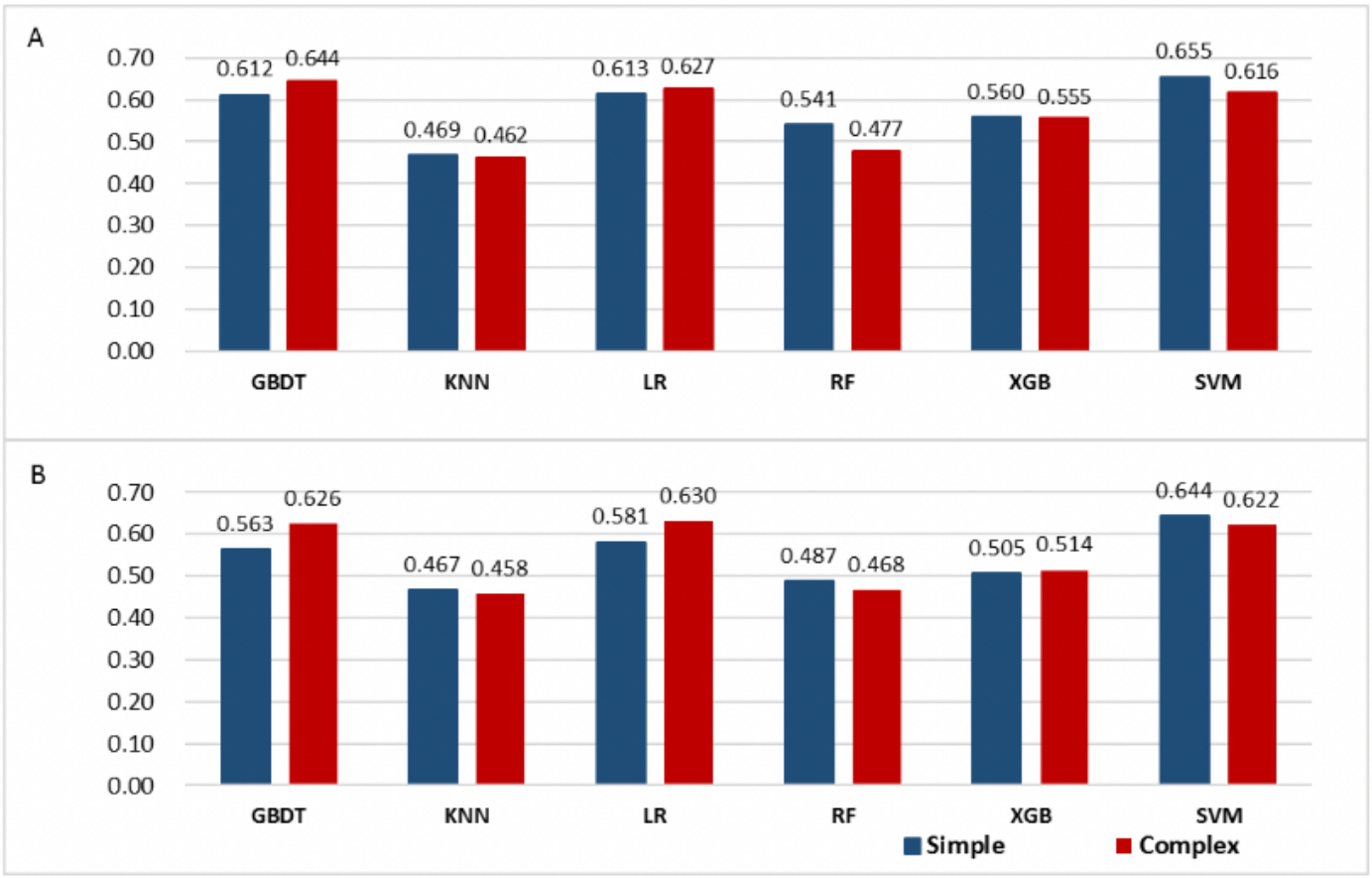

Figure 2

Model evaluation (F1-score) results based on the number of features across 6 models (A: primary cohort; B: validation cohort). Abbreviation: GBDT, gradient boosting decision tree; KNN, k-Nearest Neighbors; LR, logistic regression; RF, Random Forest; XGB, xgboost; SVM, Support Vector Machine.

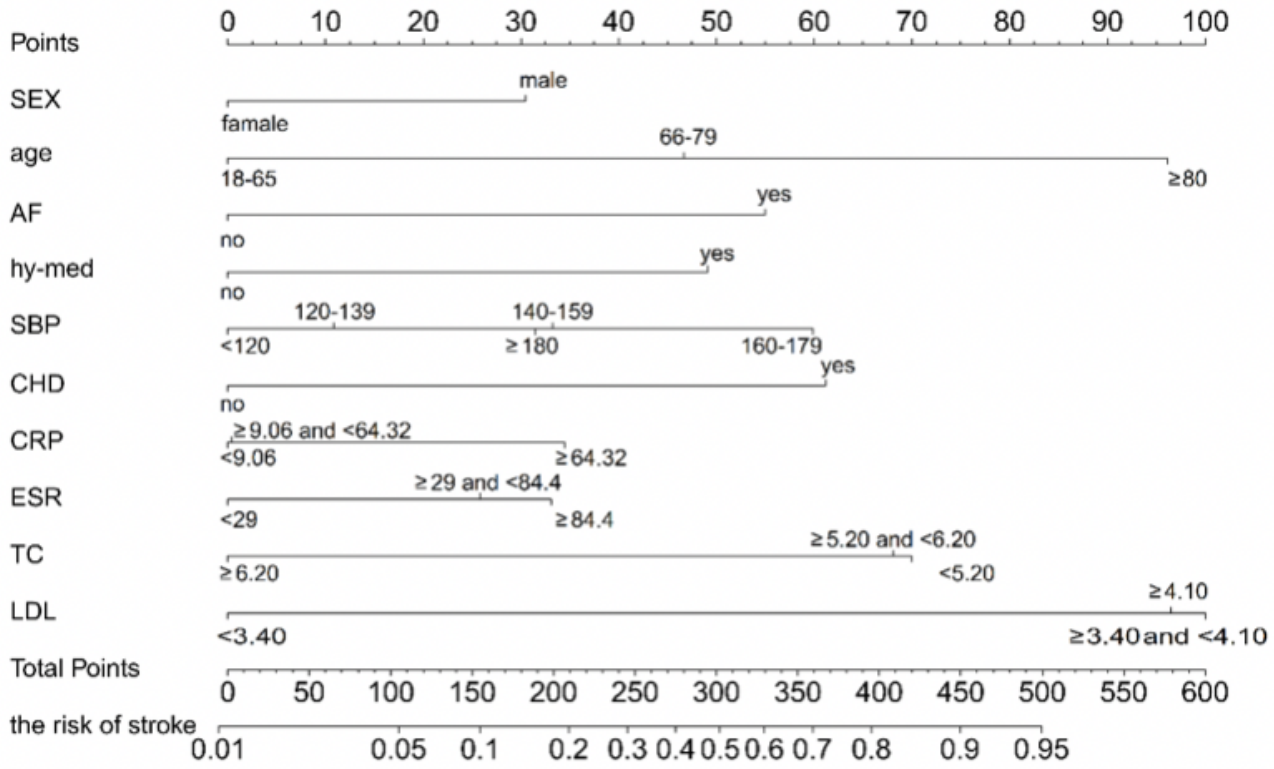

Figure 3

A developed stroke nomogram in the train cohort. Abbreviation: SBP, systolic blood pressure; CHD, coronary heart disease; AF, atrial fibrillation; hy-med, hypotensive medicine; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; TC, total cholesterol; LDL, low-density lipoprotein cholesterol. For example, a 70 year-old (47 points)male (30 points) RA patient with a AF (55) and CHD (62) history 60 mm/H ESR (27 points), 5mmol/L TC (65 points) arrived at a total point value of 286, which given an estimated probability of 46% for stroke developing.

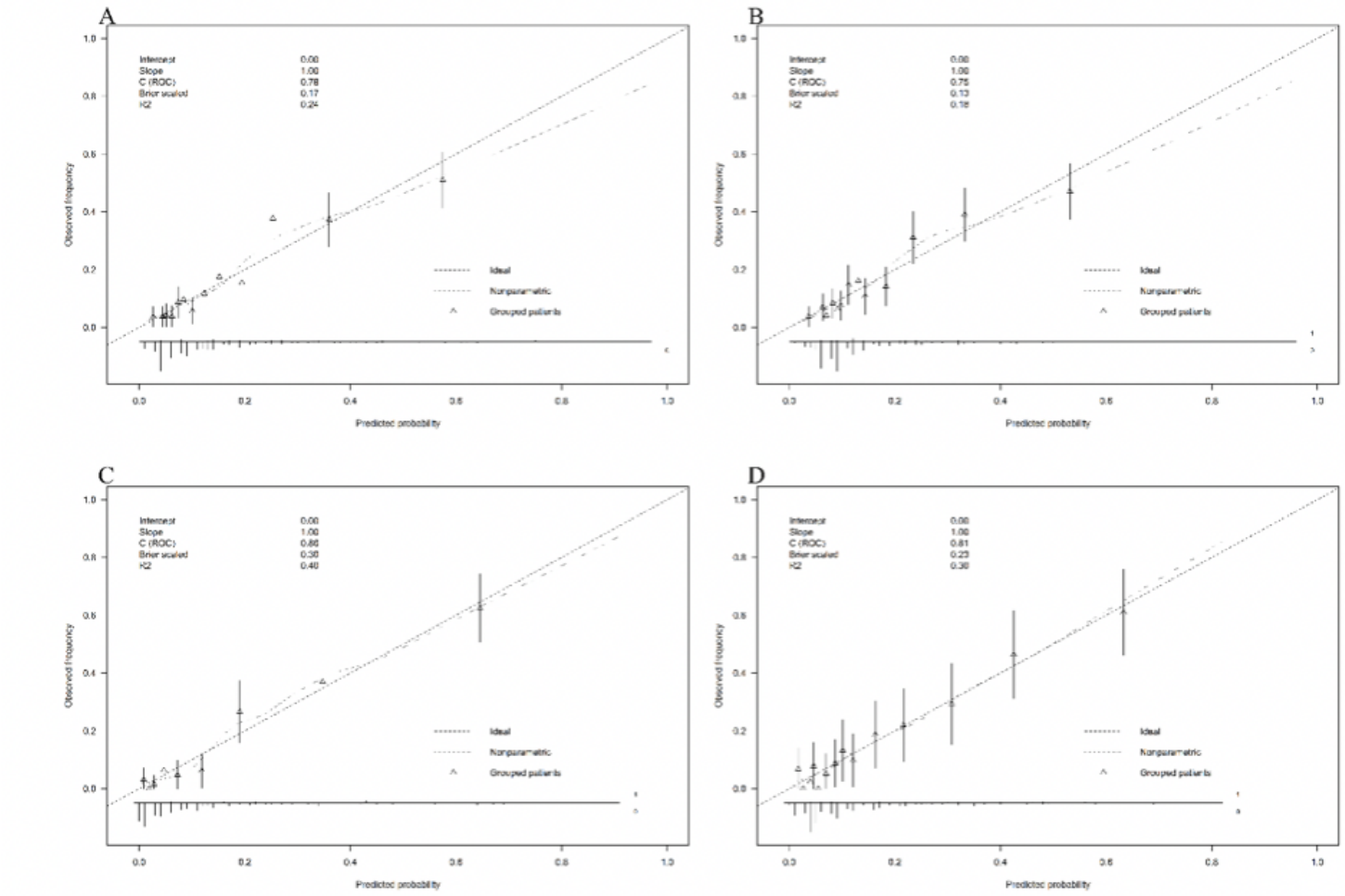

Figure 4

Calibration curves of (A) the complex model of the develop cohort, (B) the simple model in the develop cohort, (C) in the complex model of the validation cohort, (D) the simple model in the validation cohort. Calibration curves depicted the calibration of each model in terms of the agreement between the predicted risks of stroke and observed outcomes of stroke. The y-axis represents the actual stroke. The x-axis represents the predicted stroke risk. The diagonal gray line represents a perfect prediction by an ideal model. The dotted line represents the performance of the nonparametric nomogram, of which a closer fit to the diagonal gray line represents a better prediction.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx