Machine-Learning-Derived Model for the Stratification of Cardiovascular risk in Patients with Ischemic Stroke

George Ntaios, MD,[a] Dimitrios Sagris, MD,[a] Athanasios Kallipolitis,[b] Efstathia Karagkiozi,[a] Eleni Korompoki, MD,[c] Efstathios Manios, MD,[c] Vasileios Plagianakos, PhD,[d] Konstantinos Vemmos, MD,[c] and Ilias Maglogiannis, PhD[b]

Background Stratification of cardiovascular risk in patients with ischemic stroke is important as it may inform management strategies. We aimed to develop a machine-learning-derived prognostic model for the prediction of cardiovascular risk in ischemic stroke patients. *Materials and Methods:* Two prospective stroke registries with consecutive acute ischemic stroke patients were used as training/validation and test datasets. The outcome assessed was major adverse cardiovascular event, defined as non-fatal stroke, non-fatal myocardial infarction, and cardiovascular death during 2-year follow-up. The variables selection was performed with the LASSO technique. The algorithms XGBoost (Extreme Gradient Boosting), Random Forest and Support Vector Machines were selected according to their performance. The evaluation of the classifier was performed by bootstrapping the dataset 1000 times and performing cross-validation by splitting in 60% for the training samples and 40% for the validation samples. *Results:* The model included age, gender, atrial fibrillation, heart failure, peripheral artery disease, arterial hypertension, statin treatment before stroke onset, prior anticoagulant treatment (in case of atrial fibrillation), creatinine, cervical artery stenosis, anticoagulant treatment at discharge (in case of atrial fibrillation), and statin treatment at discharge. The best accuracy was measured by the XGBoost classifier. In the validation dataset, the area under the curve was 0.648 (95%CI:0.619−0.675) and the balanced accuracy was 0.58 ± 0.14. In the test dataset, the corresponding values were 0.59 and 0.576. *Conclusions:* We propose an externally validated machine-learning-derived model which includes readily available parameters and can be used for the estimation of cardiovascular risk in ischemic stroke patients.

Key Words: Ischemic stroke—Cardiovascular risk—Risk stratification—Machine learning

Introduction

Patients with ischemic stroke are considered as very high risk patients for recurrent cardiovascular events and mortality.[1] Nevertheless, even within this very high risk group, there is considerable variation of the cardiovascular risk, with some patients being at the extreme edge of the spectrum.[2] Further-refined risk-stratification of patients with ischemic stroke is important as it may guide management strategies such as intensive treatment of modifiable risk factors and prioritization of high-cost strategies.[1] Also, it may inform treatment decisions in patients where the treatment-related risk-benefit balance is delicate like in patients with indication for antithrombotics and high bleeding risk[3] or those who may benefit more from an intensive follow-up schedule. Additionally, it may have a positive impact on the motivation of patient adherence to secondary prevention strategies.

From the [a]Department of Internal Medicine, University of Thessaly, Greece; [b]Department of Digital Systems, University of Piraeus, Greece; [c]Department of Clinical Therapeutics, University of Athens, Greece; and [d]Department of Computer Science and Biomedical Informatics, University of Thessaly, Greece.

Received April 28, 2021; revision received July 12, 2021; accepted July 18, 2021.

Corresponding author. E-mail: gntaios@med.uth.gr.

1052-3057/$ - see front matter

https://doi.org/10.1016/j.jstrokecerebrovasdis.2021.106018

In this context, we aimed to develop a machine-learning-derived prognostic model for the prediction of cardiovascular risk in patients with ischemic stroke.

Methods

The study was performed according to Good Clinical Practice guidelines and the Helsinki Declaration, registered at Clinicaltrials.gov (NCT04189497) and approved by the Institutional Review Board of the Larissa University Hospital, Greece. Patient informed consent was waived as this was a retrospective analysis of de-identified data. This analysis is reported according to recently published recommendations for reporting machine-learning analyses in clinical research.[4] The analysis was performed with the sklearn, pandas and numpy libraries of the Python programming language.

Outcome definition

The outcome assessed was major adverse cardiovascular event, defined as non-fatal stroke, non-fatal myocardial infarction, and cardiovascular death during 2-year follow-up. Cardiovascular death included death due to stroke, myocardial infarction, heart failure or cardiogenic shock, sudden death, or any other death due to other cardiovascular causes.

Data source

The training/validation and test datasets contained all consecutive acute ischemic stroke patients admitted at the Alexandra University Hospital (Athens/Greece) during 1993−2011 and the Department of Internal Medicine at the Larissa University Hospital (Larissa/Greece) during 2013−2020, respectively. To maximize the power of our study, we included all available patients from these two prospective stroke registries.

Preprocessing

An extended set of parameters was prospectively registered for each patient including demographics, medical history, vascular risk factors, previous treatment, stroke severity at admission, laboratory results, imaging data, in-hospital treatment, and medication at discharge. Patients were followed up prospectively at the outpatient clinic at 1, 3 and 6 months after hospital discharge and yearly thereafter. We excluded parameters with >30% missing values. Remaining missing values were imputed by the k-nearest neighbors (KNN) algorithm with $n = 5$.

Model training

The selection of the variables included in the classifier was performed with the LASSO technique, which ranks the variables in terms of importance by penalizing each variable with a restriction on the sum of absolute values of the coefficients, while performing regression to reduce the freedom of the predictive model.

The correlation between the selected variables were summarized in Point biserial correlation matrix (Supplemental material). The following three algorithms were selected according to their performance: XGBoost (Extreme Gradient Boosting); Random Forest; and Support Vector Machines. Fine tuning of these classifiers after testing their performance with different values of the basic parameters, led to the configuration presented in the supplemental material.

We also evaluated the performance of the classifier for the prediction of 5-years outcome in the training/validation dataset, but not in the test dataset, given that only a minority of patients in the Larissa Stroke Registry completed 5 year follow-up at the time of the analysis.

Model validation

The evaluation of the classifier was performed by bootstrapping the dataset 1000 times and performing cross-validation by splitting in 60% for the training samples and 40% for the validation samples. The Area under ROC curve (AUC) was measured for all iterations to provide 95% confidence intervals. In addition to the bootstrapping evaluation technique, we performed a second evaluation scheme, namely 10-fold cross validation, to assess the balanced accuracy of the classifier, defined as (sensitivity +specificity)/2. We also employed 2-dimensional visualization using Principal Component Analysis (PCA) to

Table 1. *Baseline characteristics of patients in the training/validation and test datasets.*

Variables	Training/validation dataset ($n = 2832$)	Test dataset ($n = 561$)
Age, median (IQR)	71 (63−79)	80 (75−85)
Female gender (%)	1098 (38.8)	280 (49.8)
Hypertension (%)	1984 (70.1)	435 (77.4)
Heart failure (%)	218 (7.7)	29 (5.2)
Peripheral artery disease (%)	123 (4.3)	17 (3)
Carotid Stenosis (%)		
No stenosis or not reported	831 (29.3)	150 (26.7)
< 30%	1242 (43.9)	278 (49.5)
30-69%	171 (6)	46 (8.2)
≥ 70%	310 (10.9)	50 (9.8)
Occlusion	278 (9.8)	32(5.7)
Atrial fibrillation (%)	974 (34.4)	232 (41.3)
Anticoagulants prior (%)	191 (6.7)	65 (11.6)
Anticoagulant at discharge	806 (28.5)	104 (18.5)
Statin use prior to Stroke (%)	186 (6.6%)	196 (34.9)
Statin at discharge (%)	587 (20.7%)	526 (93.6)
MACE at 2 years (%)	433 (15.3)	83 (14.8)
MACE: major adverse cardiovascular event		

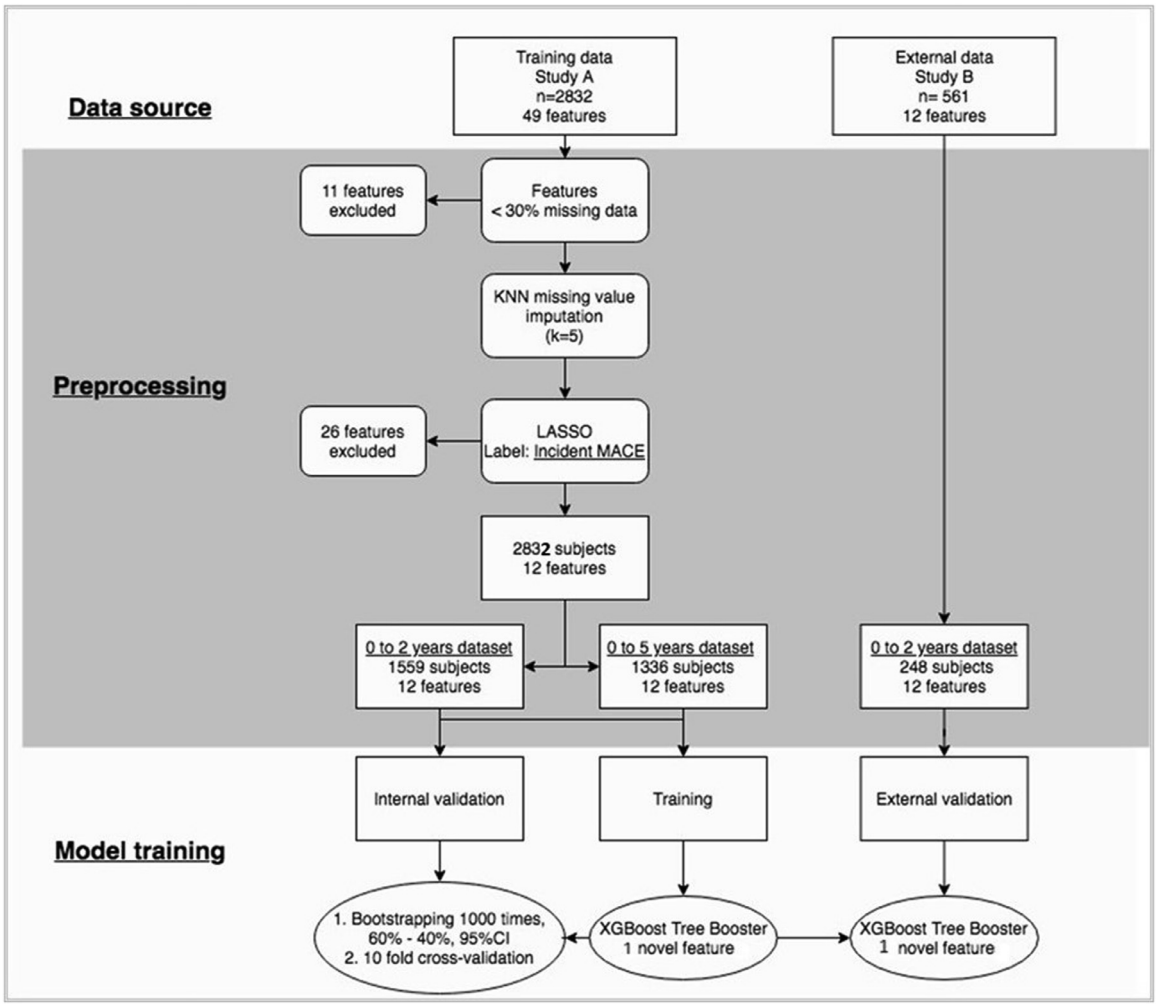

Fig. 1. *Disposition of patients and parameters included in the analysis.*

visually investigate separability between patients who had an outcome event and those who did not.

Availability of the classifier

The XGBoost classifier is available for online use at http://83.212.75.102:3002.

Results

The baseline characteristics of patients in the training/validation and test datasets are summarized in Table 1. In the test dataset the included subjects were older, there were more females, higher rates of hypertension and atrial fibrillation and a higher use of statins. The workflow of the analysis is summarized in Fig. 1. The variables which were included in the analysis are listed in supplemental table 1. The parameters which were selected by the LASSO technique and included in the classifier were: age; gender; AF; heart failure; peripheral artery disease; arterial hypertension; statin prior to admission; anticoagulant prior to admission (in case of AF); serum creatinine; cervical artery stenosis; anticoagulant at discharge (in case of AF); and statin at discharge. The ranking of these parameters by means of 5-fold cross-validated Lasso is presented in the supplemental material. We found low degree of correlations between the variables (supplemental Fig. 2).

The configuration of the classifiers is summarized in the supplemental material (supplemental table 2). The best accuracy was measured by utilizing the XGBoost classifier. The two-dimensional PCA is visualized in the supplemental material (supplemental Fig. 3). In the training/validation dataset, the AUC of the model was 0.648 (95%CI:0.619−0.675) and the balanced accuracy was 0.58 ± 0.14 for the prediction of 2 year outcome. The corresponding values were 0.655 (95%CI:0.626−0.682) and 0.59 ± 0.19 for the 5 year outcome prediction. In the test dataset, the AUC was 0.59 and the balanced accuracy was 0.576 for the prediction of the 2 year outcome (Fig. 2).

Discussion

The available tools for the stratification of cardiovascular risk in unselected ischemic stroke patients have limitations. The SMART score was developed to stratify the

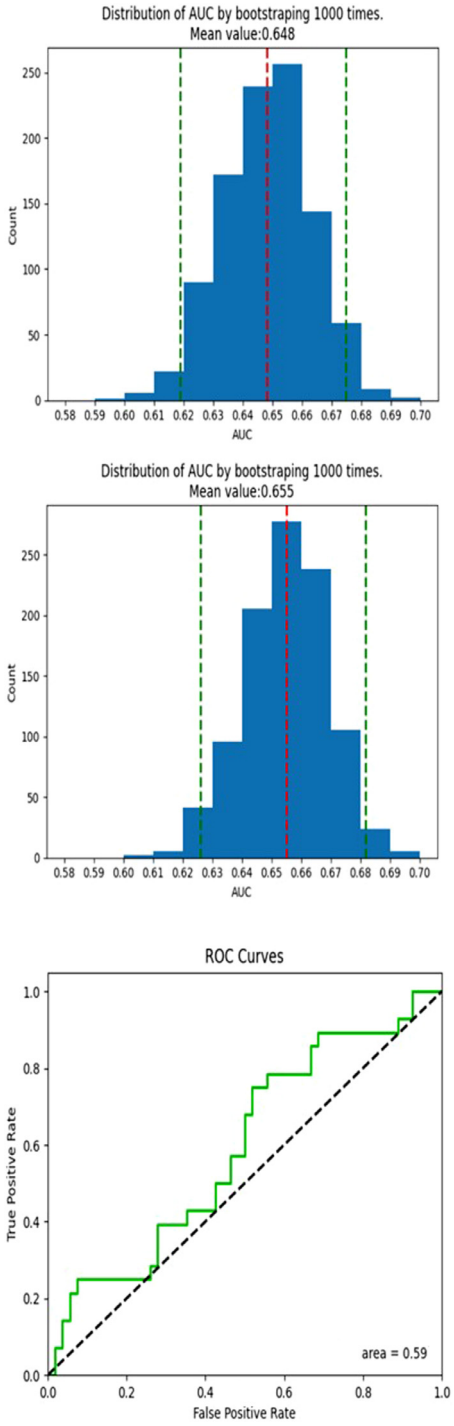

Fig. 2. *Distribution of AUC by bootstrapping the initial dataset 1000 times and splitting 60−40% for the prediction of outcome at 2 years (upper panel) and 5 years (middle panel) in the training/validation dataset. ROC curve of the model for the prediction of outcome at 2 years in the test dataset (lower panel).*

cardiovascular risk in patients with various clinical manifestations of arterial disease, but not specifically in ischemic stroke patients.[2] This might be of importance, given that unlike coronary and peripheral artery disease, stroke is an etiologically heterogeneous syndrome.[5] The CHA_2DS_2VASc score, originally developed for risk stratification of AF patients, can predict stroke outcomes also in non-AF ischemic stroke patients, but with modest discriminatory performance.[6–8] The Essen Stroke Risk score (ESRS) was shown to stratify the risk of stroke recurrence or major vascular events in ischemic stroke patients in the CAPRIE trial, but AF patients were excluded.[9]

The AUC of our model in the derivation cohort was 0.646. Although not outstanding, our score performed better than the CHA_2DS_2VASc and the ESRS in our training/validation datasets (c-statistic of 0.578 and 0.544 respectively), as we reported previously.[10] Also, the AUC of our model is comparable and (at least numerically) higher than the AUC of other guideline-recommended risk stratification tools. In particular, the widely used CHA_2DS_2VASc score yielded an AUC of 0.606 in the derivation cohort for the risk stratification of stroke or systemic embolism in AF patients.[11] Further validation of the proposed model is welcome.

A strength of the proposed model is that it includes readily available parameters. The inclusion of more sophisticated parameters like biomarkers and advanced cardiovascular imaging could possibly enhance its prognostic performance but would limit its applicability. The limitations of this study are the moderate size of the test dataset and the retrospective design of the analysis, although the patients were registered prospectively, and data were analyzed in pre-specified manner. In addition, there were differences in patient characteristics between the two datasets, mainly due to their difference in time span (1993−2011 for the training/validation dataset and 2013−2020 for the test dataset), which might have influenced the results. Moreover, the high proportion of missing data for some features with well-established association with cardiovascular prognosis like LDL-cholesterol at baseline or follow-up, led to their exclusion from the analysis. Still, two of the features which were included in the final model i.e., statin treatment before stroke onset and at discharge, provide relative information.

In conclusion, we propose an externally-validated machine−learning derived model which includes readily available parameters and can be used for the stratification of cardiovascular risk in ischemic stroke patients. For example, this score could be used to identify patients who need closer follow-up to ensure better adherence and persistence to therapeutic strategies; patients who need more intensive therapeutic interventions e.g. lower LDL-cholesterol levels or lower blood pressure levels. In addition, a higher score in a specific patient could serve as a further motivation to adhere more closely to the indicated treatment.

Funding source

AMGEN.

Disclosures

None related.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.jstrokecerebrovasdis.2021.106018.

References

1. Mach F, Baigent C, Catapano AL, Koskinas KC, Casula M, Badimon L, et al. 2019 ESC/EAS Guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk: the task force for the management of dyslipidaemias of the European society of cardiology (ESC) and European atherosclerosis society (EAS). Eur Heart J 2020;41:111-188.
2. Dorresteijn JA, Visseren FL, Wassink AM, Gondrie MJ, Steyerberg EW, Ridker PM, et al. Development and validation of a prediction rule for recurrent vascular events based on a cohort study of patients with arterial disease: the SMART risk score. Heart 2013;99:866-872.
3. Ntaios G, Lip GY. Difficult situations in anticoagulation after stroke: between Scylla and Charybdis. Curr Opin Neurol 2016;29:42-48.
4. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. Circ Cardiovasc Qual Outcomes 2020;13:e006556.
5. Ntaios G, Hart RG. Embolic stroke. Circulation 2017;136:2403-2405.
6. Ntaios G, Lip GY, Makaritsis K, Papavasileiou V, Vemmou A, Koroboki E, et al. $CHADS_2$, $CHA_2S_2DS_2$ - VASc, and long-term stroke outcome in patients without atrial fibrillation. Neurology 2013;80:1009-1017.
7. Andersen SD, Gorst-Rasmussen A, Lip GY, Bach FW, Larsen TB. Recurrent stroke: the value of the CHA_2DS_2-VASc score and the essen stroke risk score in a nationwide stroke cohort. Stroke. 2015;46:2491-2497.
8. Ntaios G, Vemmos K, Lip GY, Koroboki E, Manios E, Vemmou A, et al. Risk stratification for recurrence and mortality in embolic stroke of undetermined source. Stroke 2016;47:2278-2285.
9. Weimar C, Diener HC, Alberts MJ, Steg PG, Bhatt DL, Wilson PW, et al. The Essen stroke risk score predicts recurrent cardiovascular events: a validation within the reduction of atherothrombosis for continued health (REACH) registry. Stroke 2009;40:350-354.
10. Georgiopoulos G, Ntaios G, Stamatelopoulos K, Manios E, Korompoki E, Vemmou E, et al. Comparison of risk scores for the prediction of the overall cardiovascular risk in patients with ischemic stroke: the Athens stroke registry. J Stroke Cerebrovasc Dis 2019;28:104415.
11. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest 2010;137:263-272.