Predicting stroke risk in Chinese hypertensive population using machine learning

X. Huang[1], T.Y. Cao[2], Y.P. Wei[3], B. Xu[4], H.Y. Wu[5], Y.Q. Wu[1], X.S. Cheng[1], X.P. Xu[3], L.S. Liu[3]

[1] Second Affiliated Hospital of Nanchang University, Department of Cardiology, Nanchang, China; [2] University of California Santa Barbara, Biological anthropology, Santa Barbara, United States of America; [3] China Agricultural University, Beijing, China; [4] Duke University, Trinity College of Arts and Sciences, durham, United States of America; [5] Evergreen Medical Institute, Shenzhen, China

Funding Acknowledgement: Type of funding sources: Public Institution(s). Main funding source(s): Key R&D Projects, Jiangxi [20203BBGL73173] National Key Research and Development Program [2016YFE0205400]

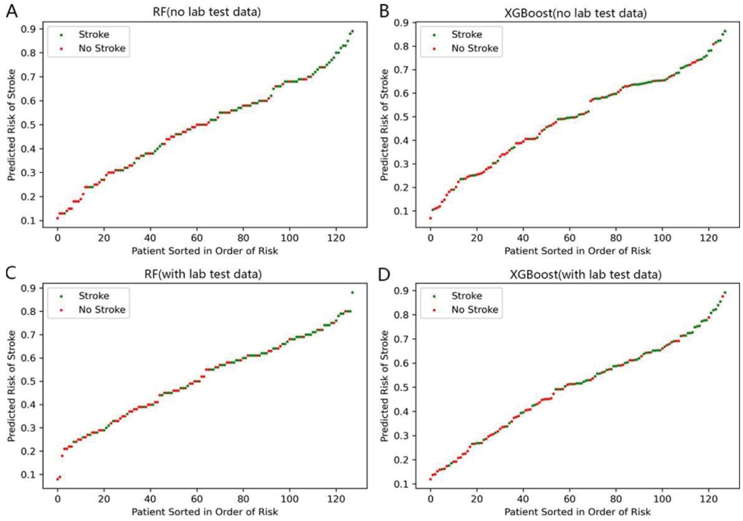
Background: Stroke is the leading cause of death in China, and the stroke burden is especially high in rural areas. Risk prediction is essential for primary prevention of stroke. However, uncertainty remains about the optimal methodology for analyzing stroke risk. In this study, we aim to determine the most effective stroke prediction method in a targeted population and establish a general framework and pipeline for future analysis.

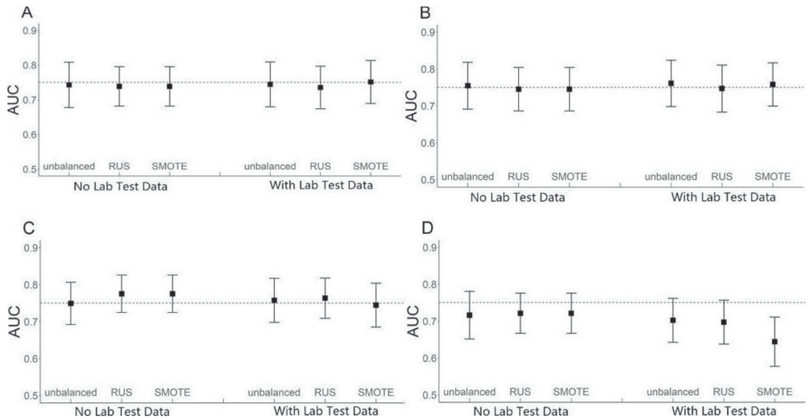
Purpose: 1) to determine the most effective stroke prediction method in a targeted population and 2) to establish a general framework and pipeline for future analysis.

Methods: Data were obtained from the China Stroke Primary Prevention Trial (CSPPT), a randomized, double-blind, multi-center clinical trial. 20,702 hypertensive patients without prior history of stroke were included in the study. The primary outcome was new nonfatal and fatal stroke (ischemic or hemorrhagic) occurring between baseline and follow-up (a median of 4.2 years). All suspected stroke cases were collected and further validated by the event adjudication committee. We compared two regression models (logistic regression and step wise logistic regression) and two machine learning methods (extreme gradient boosting and random forest). All models were trained using questionnaire data with and without laboratory data, then analyzed and compared. The primary outcome was defined as first stroke. Accuracy, sensitivity, specificity and AUCs (area under receiver operating characteristic curve) were used to assess each model. AUCs were used to evaluate the performance of each analysis method.

Results: In our data set with 20,702 samples and 127 variables, the highest AUCs (0.775 (0.725–0.826)) were observed with RUS (random under sampling) applied to RF (random forest). Before applying data balancing techniques, all analysis methods showed very low sensitivity (around 0.01), very high accuracy (around 0.97), and very high specificity (around 1.00). The mean AUCs were 0.741 (0.678–0.803). After data balancing techniques were applied, we observed an increase in sensitivity and decreases in accuracy and specificity. Different data balancing techniques had different effects on analysis methods. No significant effect on AUCs was observed; the range of increase and decrease was around 0.01. Similar overall patterns were observed when training with laboratory test data added. The mean AUCs were 0.739 (0.679–0.799) and 0.734 (0.674–9.795) for all models using data with and without laboratory test respectively. The 10 most important variables as determined by the model were selected as stroke risk predictors for all analysis models.

Conclusion: The most effective stroke prediction method in this Chinese rural hypertensive population is RUS applied to RF. The optimal analysis method and variable selection depends on data-specific features.


A
RF(no lab test data)
B
XGBoost(no lab test data)
C
RF(with lab test data)
D
XGBoost(with lab test data)
Stroke
No Stroke
Predicted Risk of Stroke
Patient Sorted in Order of Risk


A
B
C
D
AUC
unbalanced
RUS
SMOTE
No Lab Test Data
With Lab Test Data