Title: Prediction model for stroke in ageing population in India

Authors: Pravin Sahadevan[1], Vineet Kumar Kamal[1,2]*

Authors Affiliation: [1]ICMR-National Institute of Epidemiology, Chennai, India, [2]All India Institute of Medical Sciences (AIIMS), Kalyani, India

*Corresponding author. Department of Biostatistics, AIIMS, Kalyani, West Bengal 741250, India.

E-mail address: vineetstats@gmail.com.

BACKGROUND

Stroke is one of the leading causes of death worldwide. Annually, 15 million people worldwide suffer from stroke. Stroke is the fourth leading cause of death in India. In India, where cases of strokes are rising, there is a need to explore non-invasive cheap methods for the prediction of early strokes. There is no prediction model for stroke developed explicitly for the Indian population. So, we tried to build a prediction model for stroke for the Indian population. We aim to develop a prediction model for stroke to estimate the risk of stroke and to identify the risk factors for stroke.

OBJECTIVE

To develop and validate a prediction model to identify individuals at increased risk of stroke in general Indian population of age 45 years and above. To Compare the performance of the classification prediction model using machine learning algorithms for stroke.

METHODS

The data was taken from Longitudinal Ageing Study in India (LASI) Wave 1 (2017- 2018). A biennial panel survey representative of the older and elderly population aged 45 years and above for India and its states and union territories. The LASI adopted a multistage stratified area probability cluster sampling design. The sample size with complete information on stroke was 65,395 participants. The outcome variable was stroke (Yes/No). Models were developed on randomly drawn 70% of the data (training set), and their performance was evaluated on the remaining 30% of the data (the test set). In this study, we applied Logistic regression for the prediction of stroke, and we also did internal validation using the Bootstrap technique and external validation in the test set for Logistic regression model. The calibration slope, Brier score, and area under the curve were used to assess the performance of Logistic regression in development and internal and external validation. We also compared the effectiveness of Machine Learning (ML) algorithms like Logistic regression, Random Forest and Naïve Bayes in the

prediction of stroke. The performance of the ML algorithms was evaluated in terms of sensitivity, specificity and accuracy.

RESULTS

The predictors identified to be associated with stroke by Logistic regression model were females [adjusted OR:2.00 (1.70 to 2.35)], diabetes [adjusted OR:1.80 (1.52 to 2.12)], chronic heart disease [adjusted OR:1.86 (1.48 to 2.36)], high cholesterol [adjusted OR:1.54 (1.20 to 2.00)], family history of stroke [adjusted OR:3.17 (2.57 to 3.90)], smoking [adjusted OR: (1.24 (1.04 to 1.49)], physical activity [adjusted OR:2.78 (2.20 to 3.51)] and alcohol usage [adjusted OR: (1.23 (1.02 to 1.47)]. When we compared the ML algorithms, we found that Random Forest had the highest accuracy of 76.89%, highest specificity of 77.30% and Logistic regression had the highest sensitivity of 72.02% and area under curve of 0.77 compared to all other ML algorithms in the test set.

CONCLUSION

We found that Logistic regression model had good discrimination ability and was not very well calibrated in stroke prediction. Random Forest had the best accuracy in the prediction of stroke when compared to other ML algorithms.

Keywords: Prediction model, Stroke, LASI, Machine learning.