Risk Prediction of Stroke: A Prospective Statewide Study on Patients in Maine

Le Zheng, Yue Wang, Shiying Hao, Karl G. Sylvester, Xuefeng B. Ling
Departments of Surgery
Stanford University
Stanford, CA, USA

Bo Jin, Chunqing Zhu, Hua Jin, Dorothy Dai, Haihua Xu, Frank Stearns, Eric Widen
HBISolutions Inc.
Palo Alto, CA, USA

Andrew Young Shin
Departments of Pediatrics
Stanford University
Stanford, CA, USA

Devore S. Culver, Shaun T. Alfreds, Todd Rogow
HealthInfoNet
Portland, ME, USA

***Abstract*— Predicting the future risks of stroke for patients is in high demands. In this paper, we proposed a model predictive of risks of stroke in future 1 year's period for patients across all age, all payor, and all disease groups in Maine, using demographics and clinical histories extracted from Electronic Medical Record (EMR) and clinical notes provided by Health Information Exchange (HIE). A retrospective cohort of 180,196 patients and a prospective cohort of 347,504 patients were constructed for model development and validation, respectively. A logistic regression model based on multivariate analysis was built for risk prediction. The model had a c-statistic of 0.887 in prospective testing, resulting in a sensitivity of 0.410 at a positive predictive value (PPV) of 0.262. Integration of this early-warning system into online patient monitoring platforms enables better management of population with chronic conditions.**

Keywords—stroke prediction; EMR; statistical learning; risk assessment

I. INTRODUCTION

Stroke, accounting for 6.7 million deaths in 2012, was the second most frequent cause of death in the world. Prevalence of stroke in U.S. adults was around 2.6% (6.6 million) in 2012 [1]. Assessing patient risks of future stroke using big data analytics [2] can help the health care provider manage patient health status, and drive proper interventions to improve the life quality of patients.

Several risk models of stroke have been proposed in previous studies [3-7]. However, most of them focused on the patients in specific groups, such as patients within specific age range [3], or located in specific areas and countries [4, 5]. A model estimating stroke risks for patients with full demographic is still needed.

In this paper, a risk model predictive of stroke in future 1 year targeting at patients in Maine was proposed and validated prospectively. The model was derived based on Electronic Medical Records (EMR) and clinical notes provided by Health Information Exchange (HIE). To our knowledge, it is the first model of predicting the risk of stroke on patients across all age, all payor, and all disease groups.

II. METHODS

A retrospective cohort of 180,196 patients from HealthInfoNet (HIN) operated by HIE was assembled with the associated demographics and clinical histories between January 1st, 2012 and December 31st, 2012, to derive a model predicting the risks of having stroke between January 1st, 2013 and December 31st, 2013. To validate, a prospective cohort of 347,504 patients with demographics and clinical histories between January 1st, 2013 and December 31st, 2013 was constructed, to predict the risks of stroke between January 1st, 2014 and December 31st, 2014.

Initially, around 30,000 demographic and clinical features were extracted from EMR and notes. Natural language processing (NLP) techniques were applied to collect clinical histories from notes. A random forest algorithm [8] was applied to the retrospective cohort to gauge the risk of stroke for each patient. Features were ranked ranking according to the MSE increase due to the random permutation of each feature. Top 100 features were selected. 64 features that didn't correlated to stroke were then removed by manual review. Finally, 26 features having significant p-values (<0.05) by multivariate analysis were selected for model development. These features included 19 diagnoses, 2 prescription medications, counts of emergency department visits, costs and chronic diseases, and 5 NLP features. Results of multivariate analysis of these features in discriminating patients with stroke from those without stroke in future 1 year were shown in Table I.

Prior to the modelling process, the retrospective cohort was randomized into training, calibration, and blind-testing sub-cohorts, with the ratio of patients with stroke to those without stroke maintained at the same level in each sub-cohort. A

978-1-4673-6799-8/15/$31.00 ©2015 IEEE

logistic regression model was built with the training sub-cohort. The output of the model was a risk score (ranging between 0 and 1) describing the probability of having stroke in future 1 year. Continuous scores were converted into a binary classification (threshold = 0.3) with the calibration sub-cohort. The threshold was chosen so that both positive predictive value (PPV) and sensitivity reached at acceptable levels. The model was then validated with the blind-testing sub-cohort, and tested on the prospective cohort.

III. Model Performances

The c-statistics for the retrospective and prospective predictions were 0.892 and 0.887, respectively. At a PPV of 0.262, the model correctly identified 41.0% (3,028 of 7,387) of prospective patients who had stroke in future 1 year (Table II). Such prospective performance highlights the effectiveness of our model in identifying an impressive percentage of population having stroke over a large, independent cohort.

We also tested performance of the Framingham study model with our prospective cohort [9, 10]. A c-statistics of 0.836 was achieved. The receiver operating characteristic (ROC) curves (Fig. 1) illustrate that our model had a c-statistics comparable to the Framingham model. Furthermore, our model was applied to patients across all age, all payor, and all disease groups, while the Framingham model was applied to patients with age 54+. Prospective results indicated that our model derived for patients with full demographics using statewide EMR and clinical notes had accuracy comparable to other commercial models derived for special patient groups.

IV. Conclusion

A risk model predictive of stroke in future 1 year was developed based on EMR and clinical notes, for patients across all payers, all diagnoses, and all age groups in Maine. Its effectiveness was supported by the c-statistics, PPV, and sensitivity in prospective testing. Implementation of this model onto a real-time monitoring platform of statewide population can provide healthcare providers with early warnings of health status of population, which benefits timely administration of population with chronic conditions.

Acknowledgment

We thank and express our gratitude to the hospitals, medical practices, physicians and nurses participating in Maine's HIE. We also thank the biostatistics colleagues at the Department of Health Research and Policy of Stanford University for critical discussions.

[1] D. Mozaffarian, E. J. Benjamin, A. S. Go, D. K. Arnett, M. J. Blaha, M. Cushman, *et al.*, "Heart disease and stroke statistics--2015 update: a report from the American Heart Association," *Circulation,* vol. 131, pp. e29-322, Jan 27 2015.

[2] N. V. Chawla and D. A. Davis, "Bringing big data to personalized healthcare: a patient-centered framework," *J Gen Intern Med,* vol. 28 Suppl 3, pp. S660-5, Sep 2013.

[3] P. Morillas, V. Pallares, L. Facila, J. L. Llisterri, M. E. Sebastian, M. Gomez, *et al.*, "The CHADS2 Score to Predict Stroke Risk in the Absence of Atrial Fibrillation in Hypertensive Patients Aged 65 Years or Older," *Rev Esp Cardiol (Engl Ed),* vol. 68, pp. 485-91, Jun 2015.

[4] H. Yong, J. Foody, J. Linong, Z. Dong, Y. Wang, L. Ma, *et al.*, "A systematic literature review of risk factors for stroke in China," *Cardiol Rev,* vol. 21, pp. 77-93, Mar-Apr 2013.

[5] X. Chen, L. Zhou, Y. Zhang, D. Yi, L. Liu, W. Rao, *et al.*, "Risk factors of stroke in Western and Asian countries: a systematic review and meta-analysis of prospective cohort studies," *BMC Public Health,* vol. 14, p. 776, 2014.

[6] D. E. Singer, Y. Chang, L. H. Borowsky, M. C. Fang, N. K. Pomernacki, N. Udaltsova, *et al.*, "A new risk scheme to predict ischemic stroke and other thromboembolism in atrial fibrillation: the ATRIA study stroke risk score," *J Am Heart Assoc,* vol. 2, p. e000250, Jun 2013.

[7] S. Hoffmann, U. Malzahn, H. Harms, H. C. Koennecke, K. Berger, M. Kalic, *et al.*, "Development of a clinical score (A2DS2) to predict pneumonia in acute ischemic stroke," *Stroke,* vol. 43, pp. 2617-23, Oct 2012.

[8] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News,* vol. 2, pp. 18-22, 2002.

[9] P. A. Wolf, R. B. D'Agostino, A. J. Belanger, and W. B. Kannel, "Probability of stroke: a risk profile from the Framingham Study," *Stroke,* vol. 22, pp. 312-8, Mar 1991.

[10] 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,* vol. 25, pp. 40-3, Jan 1994.

TABLE I. MULTIVARIATE ANALYSIS OF VARIABLES PREDICTIVE OF STROKE OCCURRENCE FOR PATIENTS IN THE TRAINING SUB-COHORT

Features	Level	Odds Ratio (95% CI)
Age	0-50	Ref [a]
	50-60	4.15 (3.92, 4.40)
	60-70	8.47 (8.01, 8.95)
	70-80	12.3 (11.6, 13.1)
	80+	17.7 (16.5, 18.9)
Sex	M vs F	0.85 (0.81, 0.88)
Systolic Blood Pressure	Normal	Ref
	Pre-Hypertension	1.01 (0.97, 1.05)
	Hypertension	1.31 (1.10, 1.55)
BMI	Underweight	Ref
	Normal	1.72 (1.51, 1.96)
	Overweight	1.94 (1.71, 2.21)
	Obese	1.87 (1.65, 2.13)
Heart Rate	Low	Ref
	Normal	6.74 (4.53, 10.03)
	High	6.87 (4.53, 10.44)
Smoke	1 vs 0	1.41 (1.34, 1.48)
Stroke History	1 vs 0	19.8 (18.3, 21.3)
Cardiovascular Disease	1 vs 0	1.65 (1.56, 1.74)
Atrial Fibrillation	1 vs 0	1.43 (1.28, 1.6)
Disorders of Lipid Metabolism	1 vs 0	1.08 (1.04, 1.12)
Diabetes Mellitus Without Complication	1 vs 0	1.1 (1.04, 1.17)
Diabetes Mellitus With Complication	1 vs 0	1.28 (1.19, 1.36)
Diabetes Type 2 Without Complication	1 vs 0	1.02 (1.01, 1.03)
Congestive Heart Failure	1 vs 0	0.62 (0.55, 0.71)
Cardiac Dysrhythmias	1 vs 0	1.02 (1.01, 1.03)
Administrative/ Social Admission	1 vs 0	1.06 (1.01, 1.11)
Skin and Subcutaneous Tissue Infections	1 vs 0	0.83 (0.80, 0.86)
Hypertension NOS	1 vs 0	0.97 (0.96, 0.98)
Coronary Athero NOS	1 vs 0	1.09 (1.06, 1.12)
Nervous System Disorders	1 vs 0	1.34 (1.27, 1.41)
Mental Health and Substance Abuse	1 vs 0	1.56 (1.35, 1.80)
Use of Anticoag	1 vs 0	1.02 (1.01, 1.03)
Use of Guanfacine HCL	1 vs 0	1.00 (1.001, 1.007)
Use of Warfarin Sodium	1 vs 0	1.00 (1.000, 1.002)
Number of Chronic Disease	1 vs 0	1.12 (1.11, 1.13)
Count of Past 1 Year Emergency Visit	1 vs 0	1.07 (1.06, 1.09)

[a] Reference factor for the multivariate analysis

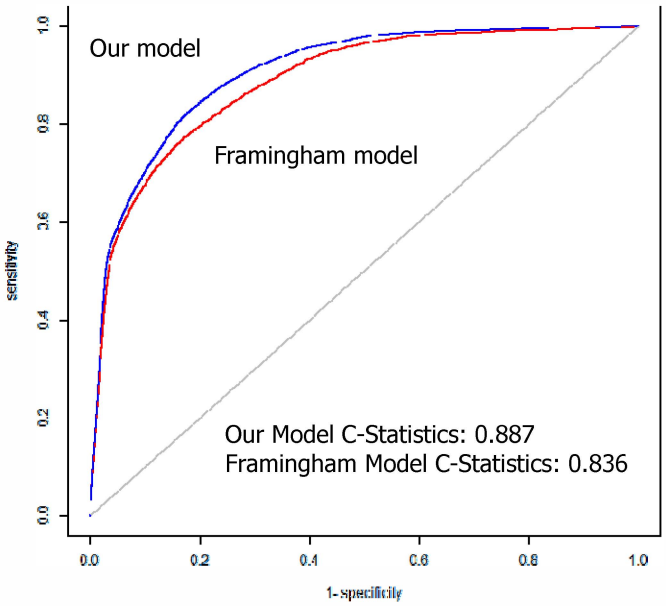

Fig. 1. ROC curve comparison of our model and Framingham model in patients classification in the prospective cohort

TABLE II. PROSPECTIVE RESULTS OF OUR MODEL

	No. of patients with stroke	No. of patients without stroke	
No. of patients classified as stroke	3,028	8,519	PPV: 0.262
No. of patients classified as non-stroke	4,359	331,598	NPV: 0.987
	Sensitivity: 0.410	Specificity: 0.975	