Machine Learning predicts Ischemic Stroke complications among Type 2 Diabetes Mellitus Patients using a Real-World Database.

Phan Thanh Phuc[1, 5], **Nguyen Phung Anh**[2,3], **Jason C. Hsu**[1,2,3,4*]

1. International Ph.D. Program in Biotech and Healthcare Management, College of Management, Taipei Medical University, Taipei, Taiwan;
2. Clinical Data Center, Office of Data Science, Taipei Medical University, Taipei, Taiwan;
3. Research Center of Health Care Industry Data Science, College of Management, Taipei Medical University, Taipei, Taiwan;
4. Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei Medical University, Taipei, Taiwan;
5. University Medical Center at Ho Chi Minh City, Ho Chi Minh City, Viet Nam.

Corresponding author: Jason C. Hsu, Ph.D.

Email: jasonhsu@tmu.edu.tw

Postal Address: 11 Fl., 172-1 Keelung Rd., Sec. 2, Da’an Dist., Taipei 106, Taiwan (R.O.C.)

Phone: 1-886-985518678

Abstract

OBJECTIVES: Diabetes Mellitus is an exceedingly predominant chronic disease worldwide. Stroke is progressively identified as a clinically considerable complication of type 2 diabetes (T2DM). This study aims to: (i) develop machine learning (ML) algorithms to predict the ischemic stroke among T2DM patients; (ii) personalize medicine for specific anti-diabetic medication groups. The population was collected from the Taipei Medical University Clinical Research Database (TMUCRD) combining data from three medical centers.

METHODS: Index 2008 data as wash-out-period—newly diagnosed T2DM patients from 2009 to 2019 as our cohort study including structured data (such as basic patient information, visits, tests, diagnosis results, treatment, etc.), unstructured data (such as physician records, discharge records). We divided the dataset into training and testing to obtain robust models and mimic the sample selection bias. The stratified 5-fold cross-validate was applied to optimize the ML models' performance [i.e., logistic regression (LR), random forest (RF), gradient boosting (GBM), and extreme gradient boosting (xGB)]. The performance was measured by Area Under the Curve (AUC), sensitivity, specificity, precision, recall, and F1 score.

RESULTS: Our population has 9,279 T2DM patients, whose training cohort is 4,697 patients, and the testing cohort is 4,582 patients. ML algorithms were performed to predict ischemic stroke risk among T2DM with their Area Under the Curve (AUC) improving from 0.79 for xGB to 0.85 for GB. The critical factors obtained from the best ML models were stroke history, aspirin, age, sulfonylureas, and contact laxatives. To personalize the medicine models for anti-diabetic medications, the AUC is applied for GLP1 (0.86), Metformin (0.83), SGLT (0.82), and Sulfonylureas (0.80) accordingly.

CONCLUSION: We successfully developed ML algorithms to predict the risk of ischemic stroke among T2DM to contribute to a clinical prognostic of the risk of a cardiovascular event and stratify the probability of complication according to the treatment line of T2DM.

Keywords: diabetes; ischemic stroke; machine learning; prediction model, TMUCRD, real-world database, patient-level prediction