Machine Learning to Predict the Ischemic Stroke among Type 2 Diabetes Mellitus Patients using Taipei Medical University Clinical Research Database

PHAN THANH PHUC [a], PHUNG ANH NGUYEN [b,d] and JASON C. HSU [a,b,c,d,1]

[a] International PhD Program in Biotech and Healthcare Management, College of Management, Taipei Medical University, Taipei, Taiwan;
[b] Clinical Data Center, Office of Data Science, Taipei Medical University, Taipei, Taiwan;
[c] Research Center of Health Care Industry Data Science, College of Management, Taipei Medical University, Taipei, Taiwan;
[d] Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei Medical University, Taipei, Taiwan.

1. Introduction

Diabetes mellitus is correlated with several complications, morbidity, and mortality (1). Studies reported that ischemic stroke highly associated with diabetes (2). Ischemic stroke has been recognized as a clinically important complication of type 2 diabetes patients (T2DM). Evidence also showed that T2DM had been associated with a 2.5-time increased risk of ischemic stroke (3). Risk prediction models for DM complications/comorbidities have substantial capacity to support the decision-making process regarding the patients' clinical management (4).

This study aims to develop machine learning algorithms to predict the risk of ischemic stroke among T2DM patients using various predictors such as patients' characteristics, disease history, laboratory tests, and medication.

2. Method

2.1. Data source and study population

The dataset was collected from the Taipei Medical University Clinical Research Database (TMUCRD) in this study. It combines comprehensive data of the three hospitals, including structured data (such as basic patient information, visits, tests, diagnosis results, treatment, surgery, and medication, etc.), unstructured data (such as physician records, pathology reports, radiological reports, discharge records), and index 2008 data as wash-out-period—newly diagnosed T2DM patients from 2009 to 2019 as our cohort study.

We identified individuals who were visited outpatient clinics (OPD) with the diagnosis of diabetes, DM (International Classification of Disease, Clinical Modification, Ninth Revision [ICD9-CM] codes 250, and Tenth Revision [ICD10-CM] codes E11. Subsequently, only the identified subjects who had at least one prescription with antidiabetic drugs (Anatomical Therapeutic Chemical [ATC] codes A10) during the treatment were included in our study.

2.2. Outcome

All patients were monitored from the date of taking antidiabetic drugs to the date the patients were admitted to hospitals with ischemic stroke (ICD9-CM codes 433, 434, 436, and ICD10-CM codes I60, I61, I62) during one-year follow-up. Data were censored if patients lost to follow-up or the end of the study, i.e., December 31, 2020. Patients with censored data or who did not admit the hospitals with ischemic stroke were defined as non-stroke (5).

[1] Jason C. Hsu, International Ph.D. program in Biotech and Health Management, College of Management, Taipei Medical University, Taipei, Taiwan; 11F., No.172-1, Sec. 2, Keelung Rd., Daan Dist., Taipei City 106, Taiwan (R.O.C.); E-mail: jasohsu@tmu.edu.tw

2.3. Features

We identified features associated with the outcomes based on diagnosis, medication, and laboratory codes from outpatient and inpatient datasets. The features were collected, including (i) patient characteristics (i.e., age, sex), (ii) comorbidities (i.e., any diagnoses before the date of taking antidiabetic drugs), (iii) other medication uses, and (iv) laboratory exams (i.e., Glucose, HbA1C, etc.). Other medication uses, and laboratory exams were collected before the date of taking antidiabetics during 6-month periods.

2.4. Statistical analysis and Model development

The dataset was divided into training and testing to obtain robust models and mimic the sample selection bias. The training set, containing the data of Taipei Medical University Hospital and Wang Fang Hospital, was used to perform the learning processing. The testing set, including the data of Shuang Ho Hospital, was used to validate the models. The stratified 10-fold cross-validate was applied in the training set to assess different machine learning models' performance and general errors.

We used different ML techniques, such as Logistic Regression (LR), Linear Discriminant Analysis (DT), Gradient Boosting Machine (GBM) and Random Forest (RF), to develop the prediction models. The performance of the algorithms was measured by Area Under the Curve (AUC), sensitivity, specificity, and F1-score.

3. Result

3.1. Patient baseline and characteristic

In our study, we collected a large cohort of 9,279 T2DM patients whose training cohort is 4,697 patients and the testing cohort is 4,582 patients. The mean age of the Training cohort is 60.7 (±15), and the Testing cohort is 62.9 (±13) (Table 1).

Table 1. Patient baseline and characteristic

Feature	Training cohort (n=4697)	Testing cohort (n=4582)
Ischemic stroke, No. (%), patient	89 (1.9%)	128 (2.8%)
Age, Mean (SD), y	60.7 (15.0)	62.9 (13.0)
Gender, No. (%)		
Female	2182 (46.5%)	2123 (46.3%)
Male	2515 (53.5%)	2459 (53.7%)
Visit per patient, Mean (SD)	24.1 (0.34)	23.3 (0.35)
Laboratory test, Mean (SD)	6.03 (0.07)	6.7 (0.09)
Medication, Mean (SD)	23 (0.2)	27 (0.2)

3.2. Model performance

Table 2. Model performance evaluation

Model	AUC (CV)	AUC (Testing)	Sensitivity	Specificity	F1-score
Logistic Regression	0.88	0.85	0.819	0.748	0.16
Linear Discriminant Analysis	0.88	0.85	0.721	0.802	0.161
Gradient Boosting Machine	0.91	0.85	0.744	0.813	0.164
Random Forest	0.93	0.84	0.811	0.692	0.133

Table 2 shows our prediction models, including LR, LDA, GBM and RF outperformed with the AUCs from 0.84 (RF) and 0.85 (LR, LDA, GBM), respectively. Moreover, the Receiver Operating Characteristic (ROC) Curve visualized the performance of all three models on the testing cohort (Figure 1).

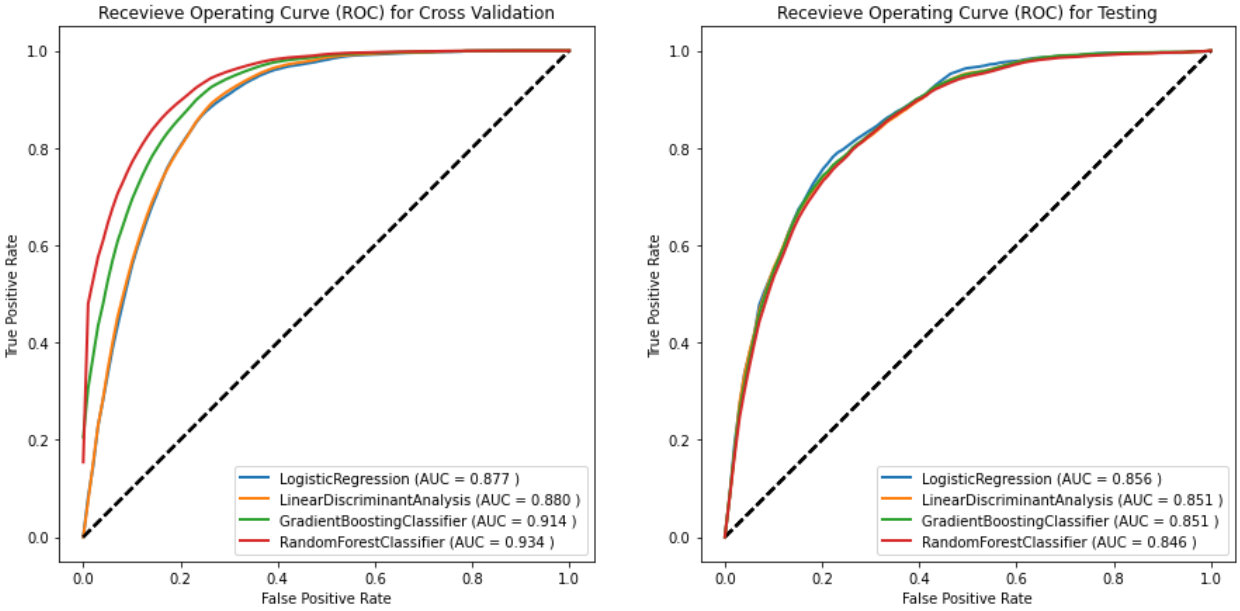

Figure 1. Receiver Operating Characteristic (ROC) Curve to evaluate the model performance

4. Discussion

In our study, we successfully developed machine learning models to predict the risk of ischemic stroke among T2DM. Our model performance improved from Random Forest to Gradient Boosting Machine. The top three important features executed from our best model are antiplatelet agent, age, and prior stroke.

The strong association of diabetes with stroke has long been appreciated (6). To the best of our knowledge, there are limited studies in classifying and predicting ischemic stroke in the T2DM cohort by developing machine learning-based models. Therefore, our findings were essential to improve the accuracy in early detection, diagnosis, and prognosis of ischemic stroke to manage the risk of diabetes complications.

Our study has some limitations. Firstly, our data source was collected from one center, covering a small portion of the Taipei population. It is uncertain whether these findings can be generalized to other location groups (7). However, it could implement in other datasets as a pretrain model. Secondly, an observational cohort study could investigate the risk occurring in a period since the risk factors may vary due to the individual condition and management.

5. Conclusion

The machine learning algorithms could predict the risk of ischemic stroke in T2DM patients with high accuracy and sensitivity. The significant clinical features that affect the model include antiplatelet medication, age, and prior stroke. The model can support the clinical diagnosis and management of the complication of diabetes patients.

References

1. Ellahham S. Artificial Intelligence: The Future for Diabetes Care [Internet]. Vol. 133, American Journal of Medicine. Elsevier Inc.; 2020 [cited 2020 Dec 23]. p. 895–900. Available from: https://doi.org/10.1016/j.amjmed.2020.03.033
2. Bloomgarden Z, Chilton R. Diabetes and stroke: An important complication. J Diabetes [Internet]. 2021 Mar 1 [cited 2022 Jan 5];13(3):184–90. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/1753-0407.13142
3. van Sloten TT, Sedaghat S, Carnethon MR, Launer LJ, Stehouwer CDA. Cerebral microvascular complications of type 2 diabetes: stroke, cognitive dysfunction, and depression. Lancet Diabetes Endocrinol [Internet]. 2020;8(4):325–36. Available from: http://dx.doi.org/10.1016/S2213-8587(19)30405-X
4. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting. BMC Med [Internet]. 2011 Sep 8 [cited 2020 Oct 29];9:103. Available from: /pmc/articles/PMC3180398/?report=abstract
5. Orso M, Cozzolino F, Amici S, De Giorgi M, Franchini D, Eusebi P, et al. Validity of cerebrovascular ICD-9-CM codes in healthcare administrative databases. The Umbria data-value project. PLoS One [Internet]. 2020 [cited 2022 Jan 7];15(1):1–15. Available from: https://doi.org/10.1371/journal.pone.0227653
6. American Diabetes Association. Cardiovascular disease and risk management: Standards of medical care in diabetes. Diabetes Care. 2021;44(January):S125–50.
7. Gu J, Pan J an, Fan Y qi, Zhang H li, Zhang J feng, Wang C qian. Prognostic impact of HbA1c variability on long-term outcomes in patients with heart failure and type 2 diabetes mellitus. Cardiovasc Diabetol [Internet]. 2018;17(1):1–11. Available from: https://doi.org/10.1186/s12933-018-0739-3