Use of machine learning techniques for phenotyping ischemic stroke instead of the rule-based methods: A nationwide population-based study

Kwon-Duk Seo, MD[1], Hyunsun Lim, PhD[2]*, Youngmin Park, MD, MPH[3], JH Hong[2], MS; Ki-Bong Yoo, Ph[34].*

*[1]Department of Neurology, National Health Insurance Service Ilsan Hospital, [2]Department of Research and Analysis, National Health Insurance Service Ilsan Hospital, [3]Department of Family Medicine, National Health Insurance Service Ilsan Hospital, [4]Division of Health Administration, Yonsei University, *Authors contributed equally*

Background. Many studies have evaluated stroke using claims data; most of these studies have defined ischemic stroke by using an operational definition following the rule-based method. Rule-based methods tend to overestimate the number of patients with ischemic stroke. We aimed to identify an appropriate algorithm for phenotyping stroke by applying machine learning (ML) techniques to analyze the claims data.

Methods. We obtained the data from the Korean National Health Insurance Service database, which is linked to the Ilsan Hospital database (n=30,897). The NHIS covers compulsory health insurance for all citizens in South Korea and provides cost-free annual or biennial health screening examinations for all insured individuals. Patients diagnosed with ischemic stroke were defined as those who were treated by a neurologist or identified through a review of the medical records of patients who visited Ilsan Hospital between 2015 and 2021. Suspected patients were defined as those who underwent at least one brain magnetic resonance imaging (MRI)/computed tomography (CT) scan, excluding those diagnosed with ischemic stroke. The control group consisted of patients with suspected and diagnosed ischemic stroke, matched by sex and age, and were selected at a 1:1 ratio. The model used 61 features: 1 rule-based operational definition, 5 personal information, 21 health examinations, 4 medical records, and 30 word-embedding variables. The embedding variables were based on the assumption that the codes frequently used in similar medical situations will have a higher probability of appearing, using a word-embedding technique to screen to a total of 2,692 codes (633 diagnosis codes, 1,841 procedure records, 100 procedure material codes, or 118 prescription records). The performance of prediction models (extreme gradient boosting [XGBoost] or long short-term memory [LSTM]) was evaluated using the area under the receiver operating characteristic curve (AUROC), the area under precision-recall curve (AUPRC), and calibration curve. The importance of model features was examined using the gain method.

Results. In total, 30,897 patients were enrolled in this study, 3,145 of whom (10.18%) had ischemic stroke. XGBoost, a tree-based ML technique, had the AUROC was 93.63% and AUPRC was 64.05%. LSTM showed results similar to those of the rule-based method. The F_1 score was 70.01%, while the AUROC was 97.10% and AUPRC was 85.70%, which was the highest.

Operational definitions	AUROC	AUPRC	F_1 score	Precision	Recall	Accuracy	Specificity
Rule-based method	91.47	73.39	68.68	54.80	91.98	91.47	91.41
XGBoost	93.63	64.05	62.50	61.46	63.56	92.15	95.43
LSTM	97.10	85.70	70.01	55.28	95.44	91.59	91.14

AUROC: The area under the receiver operating characteristic curve, AUPRC: The area under precision-recall curve, Rule-based method: individuals who are hospitalized with a diagnosis of I63 and have received anti-platelet therapy and anti-coagulant therapy within 30 days of diagnosis, LSTM: long short-term memory.

Age is the most important feature, followed by the rule-based method, death, sex, weight, highest blood pressure, height, income, and fasting blood glucose level. Income and total medical costs had a significant impact, and the medical utilization records, such as the number of days of care and hospitalization, were important variables that were explained.

Conclusion. This study found that the recurrent neural network based deep learning techniques can improve the predictability of operational definitions that have relied on rule-based methods employed in previous studies using claims data. This can provide processed and refined disease variables rather than primitive data, such as diagnosis codes, calculation special codes, medication, test and procedure codes, examination items, and qualification information when conducting studies based on claims data. Therefore, it can quickly and accurately derive the study results when conducting future studies using big data.