**Development and validation of an interpretable machine learning model for predicting post-stroke epilepsy**

1. **Supplementary questionnaire**

Have you experienced any form of seizures/epilepsies/convulsions before?

* 1. If yes,
     1. Did you have seizures/epilepsies/convulsions before the stroke? (If yes, exclude)
     2. Within the 2 years after your stroke, how many seizures have you had?
        1. If 1 seizure: How soon after the stroke did you experience the seizure?
        2. If >1 seizure, how soon after the stroke did you experience your first seizure? How soon after the first seizure did you experience the subsequent seizure(s)?
     3. Do you know what type of seizure you experienced?
        1. If no, has a seizure ever been witnessed by an observer?
           + If yes, can they explain what happened during the seizure?

How did you feel before the seizure? (e.g. normal, a feeling of déjà vu)

How did you feel after the seizure? (e.g. drowsy/confused, exhausted/sleepy, weak, muscle aches)

* + 1. What kind of medications are you on?
  1. If no, what kind of medications are you on?

Have you experienced any unexplained black outs or loss of consciousness? (If yes, how long did it last for? Was anyone else around when it happened? Was he/she able to describe what happened when you loss consciousness?)

**2. Supplementary tables**

Table S1 Detailed definitions of the variables included in this study.

|  |  |
| --- | --- |
| Variables | Definition |
| Sex | Sex was dichotomized as female and male and was extracted from electronic medical records. |
| Age | Age of onset refers to the age at which the ischemic stroke occurred and was extracted from electronic medical records. |
| Length of stay | Length of stay refers to the number of days from admission to hospital discharge and was extracted from electronic medical records. |
| Hypertension | The definition of hypertension was based on a self-reported diagnosis from a doctor or an average of two clinic blood pressure recordings ≥140 mmHg (systolic) and/or ≥90 mmHg (diastolic). |
| Diabetes | Diabetes was defined by a self-reported history of adult-onset of diabetes, or fasting glucose ≥7 mmol/L. |
| Hyperlipidemia | Hyperlipidemia was defined by a self-report of a physician diagnosis. |
| Atrial fibrillation | Atrial fibrillation was defined as either a self-report of history or the presence of fibrillatory waves on the baseline electrocardiogram. |
| Coronary heart disease | Coronary heart disease was identified by a self-report of heart disease, angina, or myocardial infarction. |
| Cancer | Cancer was defined as a self-reported history of any cancer. |
| History of ischemic stroke | History of ischemic stroke was defined by a self-report of physician’s diagnosis. |
| Smoking | Smoking was self-reported. |
| Alcohol-drinking | Alcohol-drinking was self-reported. |
| Systolic blood pressure (SBP) / diastolic blood pressure (DBP) | SBP and DBP were measured upon admission. |
| Stroke severity | The neurological severity of patients was assessed by an experienced neurologist according to the National Institutes of Health Stroke Scale (NIHSS) score on admission and was recorded in the medical record. |
| Stroke cause | The stroke cause of patients was determined by the clinical team responsible for the patient according to the Trial of ORG 10172 in Acute Stroke Treatment (TOAST) and was recorded by the neurology resident. |
| Laboratory variables | Blood samples were collected within 24 hours of admission and analyzed using routine methods. The laboratory data were extracted from the Hospital Information system (HIS). |
| Thrombectomy | Patients underwent endovascular mechanical thrombectomy during acute ischemic stroke. |
| Thrombolysis | Patients who arrived at the hospital within 4.5 h of stroke onset and underwent thrombolysis (IV-tPA) therapy. |
| Lung infection | Lung infection was identified based on clinical and laboratory indicators of respiratory tract infection, such as fever, new purulent sputum, cough, and bronchial breath sounds, with support from typical findings on chest radiographs. |
| Multiple lobes involvement | Neuroimaging revealed cerebral infarction involving two or more lobes. |
| Cortical involvement | Neuroimaging revealed cerebral infarction involved the cortex. |
| Territory of MAC | The cortical branches of the MCA supply the lateral surface of the hemisphere, except for the medial part of the frontal and the parietal lobe (anterior cerebral artery), and the inferior part of the temporal lobe (posterior cerebral artery). The deep penetrating LSA-branches includes most of the basal ganglia. Heubner's artery is the largest of the medial lenticulostriate arteries and supplies the anteromedial part of the head of the caudate and anteroinferior internal capsule. |
| Hemorrhagic transformation | After the first head CT/MRI following a cerebral infarction did not reveal any bleeding, a second head CT/MRI examination indicated the presence of intracranial hemorrhage, or confirmed the presence of hemorrhagic infarction as determined by the initial head CT/MRI. |
| Early seizure | Early seizures were defined as seizures occurring within the first 7 days after the acute ischemic stroke. |

Table S2 Missing data before imputation within the training set data (n=1383).

|  |  |  |
| --- | --- | --- |
| Variables | Missing | Percent missing (%) |
| Sex (male) | 0 | 0 |
| Age (years) | 0 | 0 |
| Length of stay(days) | 0 | 0 |
| **Vascular risk factors** |  |  |
| Hypertension | 0 | 0 |
| Diabetes | 0 | 0 |
| Hyperlipidemia | 0 | 0 |
| Atrial fibrillation | 0 | 0 |
| Coronary heart disease | 0 | 0 |
| Cancer | 0 | 0 |
| History of ischemic stroke | 0 | 0 |
| Smoking | 0 | 0 |
| Alcohol-drinking | 0 | 0 |
| **Vital signs** |  |  |
| SBP | 0 | 0 |
| DBP | 0 | 0 |
| **NIHSS at admission** | 5 | 0.4 |
| **Stroke cause** |  |  |
| Large-artery atherosclerosis | 5 | 0.4 |
| Cardioembolism | 5 | 0.4 |
| Small-vessel occlusion | 5 | 0.4 |
| Other determined cause | 5 | 0.4 |
| Undetermined cause | 5 | 0.4 |
| **Laboratory variables** |  |  |
| Fasting blood glucose | 7 | 0.5 |
| Total cholesterol | 4 | 0.3 |
| Triglycerides | 4 | 0.3 |
| Cholesterol LDL | 4 | 0.3 |
| D-dimer | 18 | 1.3 |
| **Reperfusion treatment** |  |  |
| Thrombectomy | 0 | 0 |
| Thrombolysis | 0 | 0 |
| **Lung infection** | 0 | 0 |
| **Neuroimaging markers** |  |  |
| Multiple lobes involvement | 14 | 1.0 |
| Cortical involvement | 26 | 1.9 |
| Territory of MAC | 26 | 1.9 |
| Hemorrhagic transformation | 0 | 0 |
| **Early seizure** | 0 | 0 |

Table S3 Baseline characteristics before and after imputation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Patient characteristics | Before imputation | After  imputation | | *P* | |
| **NIHSS at admission, median (IQR)** | 3.42（1,5） | | 3.46（1,5） | | 1.000 |
| **Laboratory variables, median (IQR)** |  | |  | |  |
| Fasting blood glucose, mmol/L | 6.30（4.71,7.10） | | 6.31（4.70,7.10） | | 0.990 |
| Total cholesterol, mmol/L | 4.38（3.62,5.04） | | 4.38（3.62,5.04） | | 0.950 |
| Triglycerides, mmol/L | 1.56（0.95,1.78） | | 1.56（0.95,1.78） | | 0.996 |
| Cholesterol LDL, mmol/L | 2.61（2.01,3.13） | | 2.61（2.0,3.12） | | 0.985 |
| D-dimer, ng/mL | 433（150,380） | | 436（150,380） | | 0.857 |

Table S4 The risk of PSE within the 2 years after stroke under the different SeLECT value.

|  |  |
| --- | --- |
| SeLECT value | Risk estimate (%) |
| 0 | 1.18443 |
| 1 | 2.03046 |
| 2 | 3.38409 |
| 3 | 5.58376 |
| 4 | 9.47547 |
| 5 | 15.5668 |
| 6 | 25.2115 |
| 7 | 39.5939 |
| 8 | 58.0372 |
| 9 | 77.6650 |

Table S5 Characteristics of patients from derivation cohort and validation cohort who dropped out versus those who did not.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Patient characteristics | Derivation cohort | | | Validation cohort | | |
| Included data  (n=1977) | Censored data  (n=1351) | *P* | Included data  (n=870) | Censored data  (n=712) | *P* |
| Sex (male), n (%) | 1266 (64.0%) | 879 (65.1%) | 0.54 | 588 (67.6%) | 455 (63.9%) | 0.12 |
| Age(years), median (IQR) | 63 (56, 72) | 65 (55, 75) | 0.09 | 66 (59, 75) | 65 (58, 78) | 0.25 |
| Length of stay(days), median (IQR) | 10 (7, 11) | 10 (7, 12) | 0.07 | 11 (9, 13) | 11 (9, 13) | 0.40 |
| **Vascular risk factors, n (%)** |  |  |  |  |  |  |
| Hypertension | 1429 (72.3%) | 963 (71.3%) | 0.40 | 678 (78.0%) | 548 (77%) | 0.65 |
| Diabetes | 661 (33.4%) | 447 (33.1%) | 0.47 | 328 (37.7%) | 287 (40.3%) | 0.30 |
| Hyperlipidemia | 197 (10.0%) | 149 (11%) | 0.29 | 160 (18.4%) | 126 (17.7%) | 0.72 |
| Atrial fibrillation | 168 (8.5%) | 126 (9.3%) | 0.34 | 86 (9.9%) | 79 (11.1%) | 0.43 |
| Coronary heart disease | 353 (17.9%) | 255 (18.9%) | 0.36 | 306 (35.2%) | 247 (34.7%) | 0.84 |
| Cancer | 71 (3.6%) | 61 (4.5%) | 0.39 | 42 (4.8%) | 42 (5.9%) | 0.34 |
| History of ischemic stroke | 296 (15.0%) | 221 (16.4%) | 0.27 | 176 (20.2%) | 153 (21.5%) | 0.54 |
| Smoking | 698 (35.3%) | 444 (32.9%) | 0.84 | 360 (41.4%) | 379 (39.2%) | 0.38 |
| Alcohol-drinking | 586 (29.6%) | 357 (26.4%) | 0.10 | 288 (33.1%) | 219 (30.8%) | 0.32 |
| **Vital signs, median (IQR)** |  |  |  |  |  |  |
| SBP, mmHg | 149 (136, 164) | 150 (137, 166) | 0.08 | 146 (135, 162) | 147 (135, 167) | 0.13 |
| DBP, mmHg | 83 (75, 92) | 83 (75, 93) | 0.94 | 81 (74, 90) | 83 (76, 92) | 0.23 |
| **NIHSS at admission, median (IQR)** | 2 (1, 5) | 2 (1, 7) | 0.07 | 3 (1, 5) | 3 (1, 5) | 0.13 |
| **Stroke cause, n(%)** |  |  | 0.20 |  |  | 0.17 |
| Large-artery atherosclerosis | 1102 (55.7%) | 758 (56.1%) |  | 568 (65.3%) | 442 (62.1%) |  |
| Cardioembolism | 143 (7.2%) | 105 (7.8%) |  | 70 (8.0%) | 83 (11.7%) |  |
| Small-vessel occlusion | 547 (27.7%) | 326 (24.1%) |  | 192 (22.1%) | 152 (21.3%) |  |
| Other determined cause | 64 (3.2%) | 40 (3.0%) |  | 20 (2.3%) | 15 (2.1%) |  |
| Undetermined cause | 121 (6.1%) | 88 (6.5%) |  | 20 (2.3%) | 20 (2.8%) |  |
| **Laboratory variables, median (IQR)** |  |  |  |  |  |  |
| Fasting blood glucose, mmol/L | 5.39 (4.71, 7.06) | 5.44 (4.75, 7.01) | 0.69 | 5.31 (4.62, 7.07) | 5.3 (4.62, 7.15) | 0.78 |
| Total cholesterol, mmol/L | 4.26 (3.60, 5.01) | 4.23 (3.51, 4.99) | 0.22 | 4.58 (3.85, 5.33) | 4.64 (3.90, 5.34) | 0.44 |
| Triglycerides, mmol/L | 1.26 (0.94, 1.76) | 1.23 (0.92, 1.69) | 0.10 | 1.66 (1.23, 2.50) | 1.22 (0.91, 1.66) | 0.62 |
| Cholesterol LDL, mmol/L | 2.52 (2.00, 3.12) | 2.50 (1.94, 3.13) | 0.36 | 2.65 (2.14, 3.13) | 2.55 (1.99, 3.17) | 0.11 |
| D-dimer, ng/mL | 230 (150, 370) | 240 (160, 390) | 0.06 | 400 (290, 610) | 380 (250, 700) | 0.10 |
| **Reperfusion treatment, n(%)** |  |  |  |  |  |  |
| Thrombectomy | 7 (0.3%) | 9 (0.7%) | 0.10 | 16 (1.8%) | 12 (1.7%) | 0.53 |
| Thrombolysis | 153 (7.7%) | 112 (8.3%) | 0.21 | 40 (4.6%) | 39 (5.5%) | 0.16 |
| Lung infection, n(%) | 130 (6.6%) | 95 (7.0%) | 0.08 | 116 (13.3%) | 111 (15.6%) | 0.07 |
| **Neuroimaging markers, n(%)** |  |  |  |  |  |  |
| Multiple lobes involvement | 496 (25.1%) | 358 (26.5%) | 0.52 | 216 (24.8%) | 172 (24.2%) | 0.42 |
| Cortical involvement | 548 (27.7%) | 416 (30.8%) | 0.08 | 304 (35.0%) | 227 (32.0%) | 0.17 |
| Territory of MAC | 1051 (53.2%) | 737 (54.6%) | 0.18 | 488 (56.1%) | 405 (56.9%) | 0.08 |
| Hemorrhagic transformation | 58 (2.9%) | 47 (3.5%) | 0.07 | 40 (4.6%) | 36 (5.1%) | 0.67 |
| **Early seizure, n(%)** | 24 (1.2%) | 15 (1.1%) | 0.11 | 10 (1.1%) | 7 (1.0%) | 0.75 |

Abbreviation: DBP, diastolic blood pressure; HDL, high-density lipoprotein; IQR, interquartile range; LDL, low-density lipoprotein; MAC, middle cerebral artery; NIHSS National Institutes of Health Stroke Scale; SBP, systolic blood pressure.

Table S6 AUC-ROC comparison of different ML models and SeLECT score on the internal and external validation sets a

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ML models | LR | NB | SVM | MLP | AdaBoost | GBDT | SeLECT |
| **Internal validation** | | | | | | | |
| LR | - | 0.483 | 0.533 | 0.079 | 0.052 | 0.011 | 0.671 |
| NB | 0.483 | - | 0.686 | 0.147 | 0.050 | 0.015 | 0.941 |
| SVM | 0.533 | 0.686 | - | 0.233 | 0.264 | 0.005 | 0.778 |
| MLP | 0.079 | 0.147 | 0.233 | - | 0.970 | 0.395 | 0.276 |
| AdaBoost | 0.052 | 0.050 | 0.264 | 0.970 | - | 0.370 | 0.063 |
| GBDT | 0.011 | 0.015 | 0.005 | 0.395 | 0.370 | - | 0.039 |
| SeLECT | 0.671 | 0.941 | 0.778 | 0.276 | 0.063 | 0.039 | - |
| **External validation** | | | | | | | |
| LR | - | 0.658 | 0.892 | 0.001 | 0.314 | ＜0.001 | 0.747 |
| NB | 0.658 | - | 0.935 | 0.001 | 0.015 | ＜0.001 | 0.611 |
| SVM | 0.892 | 0.935 | - | ＜0.001 | 0.038 | ＜0.001 | 0.624 |
| MLP | 0.001 | 0.001 | ＜0.001 | - | 0.628 | 0.407 | 0.008 |
| AdaBoost | 0.314 | 0.015 | 0.038 | 0.628 | - | 0.040 | 0.096 |
| GBDT | ＜0.001 | ＜0.001 | ＜0.001 | 0.407 | 0.040 | - | 0.002 |
| SeLECT | 0.747 | 0.611 | 0.624 | 0.008 | 0.096 | 0.002 | - |

a All values presented were P values

Abbreviation: AdaBoost, adaptive boost; GBDT, gradient boosting decision tree; LR, logistic regression; MLP, multilayer perceptron; NB, naive Bayes; SVM, support vector machine.