Supplementary methods 1. Selection of variables which were used as input for model

The input features for predictive models included only clinical variables measurable at admission. List of the variables and detailed counts of missing variables are listed in Supplementary Table 1. The missing values were substituted by Multivariate Imputation by Chained Equations (MICE).[1,2] In addition, predictors were excluded if they were found to have multicollinearity by a variance inflation factor before model development. Full lists of the variables which were used as input for model are as follows: age, sex, hypertension, diabetes, high risk of cardiac embolic sources, hyperlipidemia, current smoker, previous stroke including transient ischemic attack, initial systolic blood pressure (BP), initial diastolic BP, hematocrit, initial glucose, total cholesterol, high-density lipoprotein cholesterol, National Institutes of Health Stroke Scale (NIHSS) at admission, duration between onset and admission, body mass index, and trial of ORG 10172 in acute stroke treatment (TOAST) classification.

Supplementary methods 2. Developments of model

Predictive models were constructed using logistic regression (LR) and machine learning (ML) algorithms including deep learning (DL), radial-kernel support vector machine, random forest, and XGBoost. Multiple LR analyses were performed with stepwise model selection using the Akaike information criterion (AIC). The DL model used in this study had the structure of a deep feed-forward neural network, also known as the multilayer perceptron. The targeted encoding scheme was used to convert a categorical variable into binary features, and standardization was employed to normalize continuous variables when constructing ML models except for random forest and XGBoost.[3] As the performance of models derived from the same algorithms can vary according to the settings of the various hyperparameters, we tuned them by searching the best sets using 3-fold cross-validation and a random search strategy. Cross-entropy which is weighted with class frequency was used as a loss function. In a post-processing, temperature scaling and isotonic regression was applied to help the neural network and the other ML models to calibrate, respectively.[4,5] The models and strategies were implemented on Python 3.7.3 with the Scikit-learn and Skorch library.[6,7]

Supplementary methods 3. Evaluation of reliability and clinical benefit

We used expected calibration error (ECE) for quantitative assessment of calibration. ECE is the average of all gaps between the actual and predicted probabilities in each bin, as depicted in a reliability diagram.[8] More precisely,

$$\mathrm{ECE}=\sum_{m=1}^{M}\frac{|B_m|}{n}\,|\,\mathrm{actual}(B_m)-\mathrm{predicted}(B_m)|,$$

where actual(B_m) is the observed frequency (probability) of the favorable outcome in the mth bin and predicted(B_m) is the average of the predicted probabilities in the mth bin. If the predictive model is perfectly calibrated, predicted probability is equal to actual probability, resulting ECE value becomes zero. If the model is overconfident, the predicted probability will be out of the actual probability, resulting ECE value becomes large. In this study, each bin has the same number of samples; i.e., 10-quantile binning. We did not perform an additional statistical test, Hosmer-Lemeshow test, to assess agreement between actual and predicted probabilities as Moons et al.[9] were recommended.

We constructed decision curve analysis to assess the clinical utility of different decision tools, which shows net benefit across probability thresholds.[10] When none of the diagnosis or treatment strategy would apply (none-strategy), it has no benefit (e.g., early detection of disease) and no cost or harms (e.g., superfluous exposure to radiation in person without disease).[11] On the other hand, for instance, some discrimination allows the population to have more benefits than a case that all the patients are diagnosed with some disease (all-strategy).

Supplementary References

1. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* 2011;20:40-9.
2. van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate imputation by chained equations in R. *J Stat Softw* 2011;45:1-67.
3. Lundberg SM, Erion GG, Lee SI. Consistent individualized feature attribution for tree ensembles. arxiv. http://arxiv.org/abs/1802.03888. 2018. Accessed September 3, 2020.
4. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning (ICML 2017); 2017 Aug 6-11; Sydney, AU. Red Hook, NY: Curran Associates, 2017;2130-2143.

5. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. Proceedings of the 29th International Conference on Machine Learning (ICML 2005); 2005 Aug 7-11; Bonn, DE. New York, NY: ACM, 2005;625-632.
6. Varoquaux G, Buitinck L, Louppe G, Grisel O, Pedregosa F, Mueller A. Scikit-learn: machine learning in python. *GetMobile* 2015;19:29-33.
7. Skorch documentation. https://skorch.readthedocs.io/en/stable/index.html. 2017. Accessed September 3, 2020.
8. Naeini MP, Cooper GF, Hauskrecht M. Obtaining well calibrated probabilities using bayesian binning. Proceedings of the AAAI 29th Conference on Artificial Intelligence; 2015 Jan 25-29; Austin, TX. Palo Alto, CA: AAAI Press, 2015;2901-2907.
9. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1-W73.
10. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. *J Clin Oncol* 2016;34:2534-2540.
11. Localio AR, Goodman S. Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. *Ann Intern Med* 2012;157:294-295.

https://doi.org/10.5853/jos.2020.02537