frontiers in Neurology

TYPE Original Research
PUBLISHED 10 September 2026
DOI 10.3389/fneur.2026.1867290

Check for updates

OPEN ACCESS

EDITED BY
Theodoros Mavridis,
Tallaght Hospital, Ireland

REVIEWED BY
Ibrahim Abu Ella,
Delta University for Science and Technology, Egypt
Wei-Jia Luo,
Kaohsiung Medical University, Taiwan

*CORRESPONDENCE
Hongyan Li
lihongyan1019@126.com

RECEIVED 28 April 2026
REVISED 05 August 2026
ACCEPTED 12 August 2026
PUBLISHED 10 September 2026

CITATION
Zeng S, Yusufujiang A, Dang H, Zhu L, Yasheng A, Wang J, Yan W, Yu H, Mamattursun G, Aini M, Kang D and Li H (2026) External validation of machine learning models for predicting prehospital delay in acute ischemic stroke: a retrospective cohort study. *Front. Neurol.* 17:1867290. doi: 10.3389/fneur.2026.1867290

COPYRIGHT
© 2026 Zeng, Yusufujiang, Dang, Zhu, Yasheng, Wang, Yan, Yu, Mamattursun, Aini, Kang and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

External validation of machine learning models for predicting prehospital delay in acute ischemic stroke: a retrospective cohort study

Shan Zeng[1,2], Aishanjiang Yusufujiang[1,2], Hui Dang[1,2], Li Zhu[1,2], Abudukeyoumu Yasheng[3], Jingjing Wang[4], Wei Yan[4], Huijuan Yu[4], Gulijanat Mamattursun[4], Maierpu· Aini[3], Di Kang[3] and Hongyan Li[1,2]*

[1]People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi, Xinjiang, China, [2]Xinjiang Clinical Research Center for Stroke and Rare Neurological Disease, Urumqi, Xinjiang, China, [3]Kashi Prefecture Second People's Hospital, Kashi, China, [4]The First People's Hospital of Kashi Prefecture, Kashi, China

Background and purpose: Prehospital delay remains a major barrier to timely reperfusion therapy in acute ischemic stroke (AIS). We externally validated eight prediction algorithms and examined whether complex machine-learning models added value beyond logistic regression.

Methods: This retrospective cohort study included consecutive patients with AIS from two hospitals in Kashgar, China (January 2019–December 2024). Hospital 1 (n = 1,377) was the development cohort and Hospital 2 (n = 1,015) the independent validation cohort. Prehospital delay was defined as onset-to-door time >4.5 h. LASSO selection and tuning were restricted to Hospital 1. Sensitivity analyses used all 36 encoded candidate features and repeated all models without SMOTE. Evaluation included discrimination, calibration, post-hoc recalibration, and decision curves.

Results: Delay occurred in 80.9 and 83.2% of the two cohorts. In the prespecified LASSO-plus-SMOTE analysis, XGBoost had the highest validation AUC (0.771; 95% CI, 0.734–0.807), but did not significantly outperform logistic regression (difference, 0.015; 95% CI, −0.003 to 0.034; p = 0.107). Full-feature modeling produced no consistent gain. Omitting SMOTE improved calibration and often discrimination; logistic regression without SMOTE achieved AUC 0.774 and Brier score 0.122, versus 0.756 and 0.180 with SMOTE. The original SMOTE-trained logistic model underestimated absolute risk. Rural residence, non-emergency-channel presentation, referral, and non-ambulance transport were associated with absolute risk differences of 15.8, 21.5, 9.2, and 36.3 percentage points in Hospital 2.

Conclusion: Complex algorithms did not show a clear advantage over logistic regression. Because several predictors are available only after arrival, the framework is intended for retrospective health-system profiling and quality-improvement planning, not individual pre-arrival prediction. Universal interventions remain warranted given the high prevalence of delay.

KEYWORDS
acute ischemic stroke, external validation, machine learning, prediction model, prehospital delay, TRIPOD

Introduction

Acute ischemic stroke (AIS) is a leading cause of death and disability worldwide, with time-sensitive treatments such as intravenous thrombolysis and endovascular thrombectomy offering substantial benefits when administered within narrow therapeutic windows (1–3). However, prehospital delay—defined as prolonged onset-to-door time (ODT)—remains a critical barrier to timely reperfusion therapy (4–6).

Understanding factors associated with prehospital delay is important for improving stroke-care delivery (51). Most published prediction models rely on internal validation, which may overestimate performance and does not establish transportability to another clinical setting (7–12, 44, 48). The two-hospital, non-random split by setting used here corresponds to TRIPOD Type 2b under the original TRIPOD classification (13).

Machine-learning algorithms can represent nonlinear associations and interactions, but their added value over simpler methods should be demonstrated rather than assumed (14–17). This is particularly relevant when logistic LASSO is used for feature selection. Moreover, several candidate predictors in this study, including NIHSS and stroke-characteristic variables, are available only after hospital arrival. The intended use is therefore retrospective health-system surveillance, subgroup identification, and quality-improvement planning rather than emergency-dispatch use or individual pre-arrival prediction.

We externally validated eight algorithms using a development cohort from Hospital 1 and an independent validation cohort from Hospital 2. In response to the model-selection concern, we repeated all algorithms with the full candidate feature set and without SMOTE. We also evaluated calibration, *post-hoc* recalibration, decision curves, and subgroup absolute risks for modifiable care-pathway factors.

Methods

Study design and setting

This retrospective cohort study was conducted at two tertiary stroke centers in Kashgar Prefecture, Xinjiang, China: People's Hospital of Kashgar Prefecture (Hospital 1) and First People's Hospital of Kashgar City (Hospital 2). Both hospitals maintain prospective stroke registries and serve distinct catchment populations. Hospital 1 served as the development cohort and Hospital 2 as the independent validation cohort. The non-random split by hospital corresponds to TRIPOD Type 2b under the original TRIPOD classification (13).

The studies involving human participants were approved by the Clinical Research Ethics Committee of Xinjiang Uygur Autonomous Region People's Hospital (approval No. KY2224110701) and were conducted in accordance with local legislation and institutional requirements. The participants provided written informed consent to participate in the study. The study was reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement (47).

Study population

We included consecutive adult patients (≥18 years) with AIS admitted between January 1, 2019, and December 31, 2024. AIS was diagnosed according to World Health Organization criteria and confirmed by computed tomography or magnetic resonance imaging (18). Exclusion criteria were: (1) wake-up stroke with unknown symptom onset time; (2) missing ODT data; and (3) invalid ODT values (≤0 h). These criteria were applied identically to both hospitals to ensure comparability.

Outcome definition

The primary outcome was prehospital delay, defined as ODT >4.5 h, corresponding to the extended time window for IV tPA in selected patients (19). ODT was calculated as the time from symptom onset (or last known well time) to hospital arrival, recorded prospectively in the stroke registry.

Candidate predictors

We identified 36 candidate predictors based on literature review and clinical relevance, covering demographics and clinical history, stroke characteristics, presenting symptoms, prehospital care pathways, and temporal factors. Ethnicity was retained for descriptive reporting but excluded from modeling. Treatment-related variables, including intravenous thrombolysis and thrombectomy, were excluded because they occur after arrival and would create reverse-causality leakage (20, 21).

NIHSS and several stroke-characteristic variables require assessment after hospital arrival. Accordingly, the models were not designed for emergency-dispatch use, real-time pre-arrival triage, or individual prognostication.

Data preprocessing

All preprocessing parameters were fitted exclusively on Hospital 1 data and then applied to Hospital 2 data to prevent data leakage—a critical methodological safeguard for external validation (49). Missing continuous variables were imputed using median values, and missing categorical variables were imputed using the most frequent category. Continuous variables were standardized using z-score normalization. One-hot encoding was applied to nominal categorical variables with more than two categories.

Feature selection

We used LASSO logistic regression with 10-fold stratified cross-validation on Hospital 1 to reduce model complexity. One hundred inverse-regularization values from 10^{-4} to 10^{2} were evaluated, and the value maximizing cross-validated AUC was selected. Features with nonzero coefficients were retained for the prespecified analysis (22, 23).

Model development

Eight algorithms were evaluated: logistic regression, random forest, XGBoost, LightGBM, AdaBoost, support vector classifier, multilayer perceptron, and Gaussian naive Bayes (50). Hyperparameters were selected by grid search using identical

five-fold stratified splits in Hospital 1. The complete search grids and selected values are reported in Supplementary Table S2. Random state 42 was used.

The prespecified analysis applied SMOTE (sampling strategy "auto," five nearest neighbors, random state 42) inside each training fold only. Validation folds and Hospital 2 contained only observed records. After tuning, models were refitted on Hospital 1 and evaluated once in Hospital 2 (24).

Model evaluation

The primary endpoint was AUC in Hospital 2. Additional measures included Brier score, log loss, accuracy, sensitivity, specificity, PPV, NPV, and F1 score at a probability threshold of 0.50. AUC confidence intervals and paired differences between model predictions were estimated by stratified bootstrap resampling.

Calibration was assessed using calibration-in-the-large, logistic calibration intercept and slope, Brier score, log loss, and calibration curves. Post-hoc logistic and isotonic recalibration were evaluated. Recalibration performance was estimated with 10-fold cross-fitting within Hospital 2, while full-cohort coefficients were reported separately as apparent estimates.

Decision-curve analysis compared each model with reviewing all records and reviewing no records (45). The modeled action was retrospective selection of a patient record for detailed quality-improvement review of EMS access, referral routing, or transport processes. It was not an individual treatment decision. The analysis does not validate aggregation of patient-level scores for subgroup- or service-line resource allocation; such use would require a prespecified aggregation rule and prospective evaluation.

Sensitivity analyses

To assess whether LASSO selection favored additive models, all eight algorithms were repeated using all 36 non-constant encoded features. Development-cohort permutation importance was calculated for nonlinear models for interpretation. To assess the effect of SMOTE, every LASSO-selected and full-feature model was also trained without SMOTE. Thus, 32 configurations were compared using discrimination, calibration, Brier score, and decision curves.

Model interpretability

We applied SHAP analysis (25, 43) to the best-performing model using Hospital 2 data to assess external generalizability of feature importance. SHAP values quantify each feature's contribution to individual predictions based on game-theoretic principles. We generated summary plots showing global feature importance, beeswarm plots showing feature value distributions, separate individual SHAP contribution plots, and dependence plots for the top two features.

Statistical analysis

Continuous variables were reported as median (IQR) and compared using the Mann–Whitney U test. Categorical variables were reported as number (percentage) and compared using the chi-square or Fisher exact test. Multivariable logistic regression was performed in Hospital 1 (46). Variance inflation factors were calculated with an intercept, and Spearman correlations were examined for key predictors.

In Hospital 2, observed delay risks, absolute risk differences, and risk ratios were estimated for rural residence, non-emergency-channel presentation, referral, and non-ambulance transport, with confidence intervals from 1,000 bootstrap samples. These subgroup estimates are descriptive and not causal effects.

Analyses used Python 3.12 with pandas 3.0.3, NumPy 2.0.2, scikit-learn 1.9.0, statsmodels 0.14.6, imbalanced-learn 0.14.2, XGBoost 3.3.0, LightGBM 4.7.0, and SHAP 0.52.0. All tests were two-sided with $\alpha = 0.05$.

Results

Study population

Between January 2019 and December 2024, 1,590 patients were screened at Hospital 1 and 1,061 at Hospital 2. After excluding 213 and 46 patients with wake-up stroke, respectively, the final cohorts comprised 1,377 Hospital 1 patients and 1,015 Hospital 2 patients (Figure 1). No additional record had missing or invalid ODT after this exclusion.

Baseline characteristics

Baseline characteristics are summarized in Table 1. Prehospital delay occurred in 1,114 patients (80.9%) at Hospital 1 and 844 (83.2%) at Hospital 2. Median ODT was 16 h (IQR, 6–48) and 24 h (IQR, 8–40), respectively. The higher Hospital 2 median with a narrower IQR reflects greater concentration near 24 h and a heavier upper tail at Hospital 1; medians and IQR widths need not move in parallel. The corrected ODT distribution is shown in Supplementary Figure S1.

Feature selection

LASSO selected 19 of 36 encoded candidate features in Hospital 1 (Figure 2). All preprocessing, feature selection, and tuning were restricted to Hospital 1.

Multivariable logistic regression

Six factors were independently associated with delay in Hospital 1 (Table 2): non-emergency-channel presentation (aOR, 4.41; 95% CI, 2.53–7.67), absence of atrial fibrillation (aOR, 3.79; 95% CI, 1.95–7.36), non-ambulance transport (aOR, 3.22; 95% CI, 2.20–4.72), referral (aOR, 2.84; 95% CI, 1.94–4.18), absence of slurred speech (aOR, 2.25; 95% CI, 1.57–3.24), and rural residence (aOR, 2.16; 95% CI, 1.57–2.99). VIFs ranged from 1.02 to 1.18, indicating no important multicollinearity.

Machine learning model performance

In the prespecified LASSO-plus-SMOTE analysis, XGBoost had the highest Hospital 2 AUC (0.771; 95% CI, 0.734–0.807), followed by

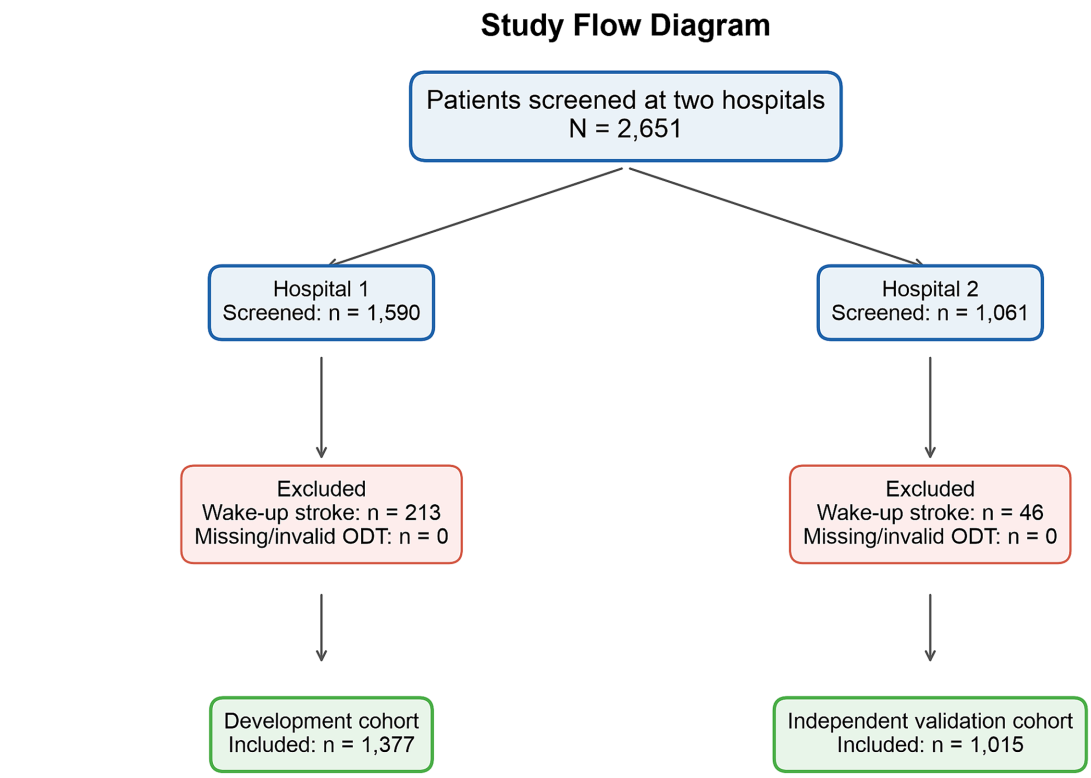

FIGURE 1
Patient flow diagram.

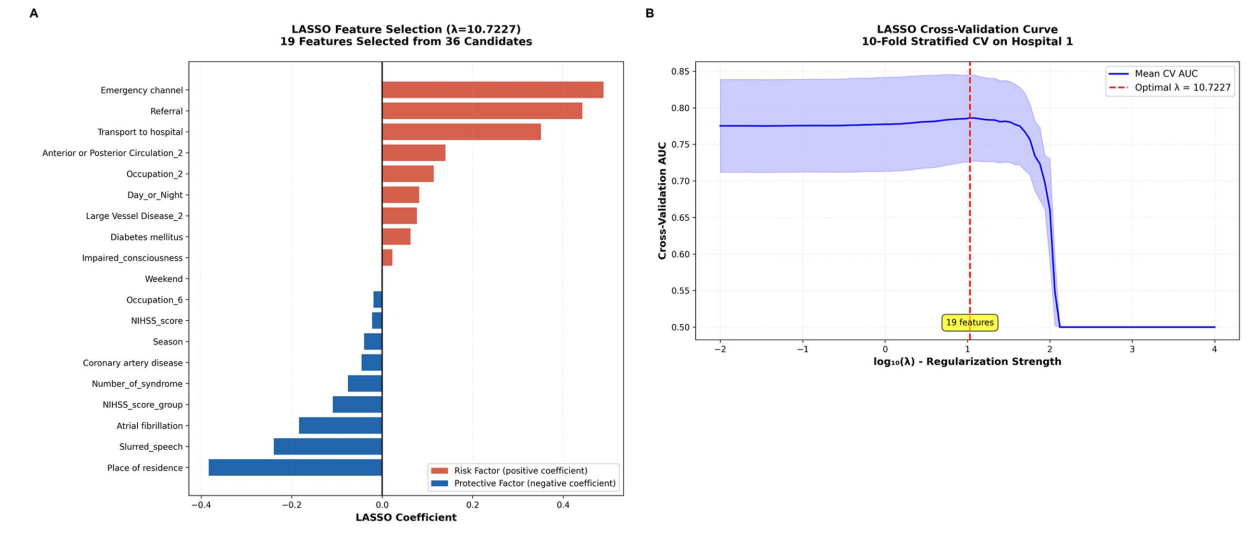

FIGURE 2
LASSO feature selection. **(A)** Horizontal bar plot of the 19 nonzero LASSO coefficients at the selected regularization value; red and blue bars denote positive and negative coefficients, respectively. **(B)** Cross-validation curve showing mean AUC versus log10 regularization strength (λ = 1/C), with the selected λ indicated by a vertical dashed line.

LightGBM (0.764), random forest (0.764), AdaBoost (0.759), and logistic regression (0.756) (Table 3; Figure 3A). XGBoost sensitivity was 0.764, specificity 0.596, PPV 0.903, NPV 0.339, and Brier score 0.162. The AUC difference between XGBoost and logistic regression was small and not significant (0.015; 95% CI, −0.003 to 0.034; $p = 0.107$). The XGBoost internal out-of-fold AUC was 0.776 (95% CI, 0.746–0.807).

Full-feature and no-SMOTE sensitivity analyses

Using all 36 features did not produce a consistent discrimination advantage for nonlinear models (Figure 3B; Supplementary Table S4). With SMOTE, the full-feature AUC changes relative to LASSO-selected models were −0.003 for logistic regression, −0.005 for

10.3389/fneur.2026.1867290

TABLE 1 Baseline characteristics of the development and validation cohorts.

Variable	Category	Hospital 1 (Training) (*n* = 1,377)	Hospital 2 (Validation) (*n* = 1,015)	*p* value	Statistical test
Age (years)	—	63.0 (55.0–72.0)	63.0 (54.0–73.0)	0.870	Mann–Whitney U
NIHSS score	—	3.0 (2.0–7.0)	4.0 (2.0–7.0)	0.002	Mann–Whitney U
Onset-to-door time (hours)	—	16.0 (6.0–48.0)	24.0 (8.0–40.0)	0.001	Mann–Whitney U
Number of risk factors	—	1.0 (1.0–2.0)	1.0 (1.0–2.0)	0.014	Mann–Whitney U
Number of symptoms	—	2.0 (1.0–2.0)	2.0 (1.0–2.0)	0.607	Mann–Whitney U
Sex	Male	846 (61.4%)	616 (60.7%)	—	—
	Female	527 (38.3%)	399 (39.3%)	—	—
Age group	<50 years	152 (11.0%)	125 (12.3%)	—	—
	50–59 years	285 (20.7%)	242 (23.8%)	—	—
	60–69 years	343 (24.9%)	239 (23.5%)	—	—
	70–79 years	257 (18.7%)	233 (23.0%)	—	—
	≥80 years	330 (24.0%)	171 (16.8%)	—	—
Educational level	Primary or below	13 (0.9%)	547 (53.9%)	—	—
	Secondary	2 (0.1%)	344 (33.9%)	—	—
	College or above	1 (0.1%)	91 (9.0%)	—	—
	Unknown	1,361 (98.8%)	33 (3.3%)	—	—
Ethnicity	Han	178 (12.9%)	152 (15.0%)	—	—
	Uyghur	1,188 (86.3%)	853 (84.0%)	—	—
	Other	11 (0.8%)	10 (1.0%)	—	—
Occupation	Retired	250 (18.2%)	185 (18.2%)	—	—
	Freelance	141 (10.2%)	46 (4.5%)	—	—
	Farmer	642 (46.6%)	599 (59.0%)	—	—
	Employee	84 (6.1%)	73 (7.2%)	—	—
	Civil servant	25 (1.8%)	19 (1.9%)	—	—
	Other	235 (17.1%)	93 (9.2%)	—	—
Residence	Urban	617 (44.8%)	469 (46.2%)	—	—
	Rural	760 (55.2%)	546 (53.8%)	—	—
Hypertension	No	463 (33.6%)	315 (31.0%)	—	—
	Yes	914 (66.4%)	700 (69.0%)	—	—
Diabetes mellitus	No	1,032 (74.9%)	715 (70.4%)	—	—
	Yes	345 (25.1%)	300 (29.6%)	—	—
Coronary Artery Disease	No	1,099 (79.8%)	744 (73.3%)	—	—
	Yes	278 (20.2%)	271 (26.7%)	—	—
Prior stroke	No	1,143 (83.0%)	879 (86.6%)	—	—
	Yes	234 (17.0%)	136 (13.4%)	—	—
Atrial fibrillation	No	1,329 (96.5%)	993 (97.8%)	—	—
	Yes	48 (3.5%)	22 (2.2%)	—	—
Emergency channel	Emergency department	1,078 (78.3%)	774 (76.3%)	—	—
	Outpatient	291 (21.1%)	226 (22.3%)	—	—
	Unknown	8 (0.6%)	15 (1.5%)	—	—

(Continued)

10.3389/fneur.2026.1867290

TABLE 1 (Continued)

Variable	Category	Hospital 1 (Training) (*n* = 1,377)	Hospital 2 (Validation) (*n* = 1,015)	*p* value	Statistical test
Transport to hospital	Ambulance (120)	165 (12.0%)	67 (6.6%)	—	—
	Private car	98 (7.1%)	20 (2.0%)	—	—
	Other	8 (0.6%)	4 (0.4%)	—	—
	Unknown	1,106 (80.3%)	924 (91.0%)	—	—
Inter-facility referral	No	877 (63.7%)	789 (77.7%)	—	—
	Yes	500 (36.3%)	226 (22.3%)	—	—
Onset time (night)	1	668 (48.5%)	473 (46.6%)	—	—
	2	708 (51.4%)	540 (53.2%)	—	—
Weekend onset	Weekday	1,017 (73.9%)	717 (70.6%)	—	—
	Weekend	359 (26.1%)	296 (29.2%)	—	—
Season	Spring	435 (31.6%)	242 (23.8%)	—	—
	Summer	336 (24.4%)	245 (24.1%)	—	—
	Autumn	321 (23.3%)	236 (23.3%)	—	—
	Winter	285 (20.7%)	292 (28.8%)	—	—
NIHSS group	Mild (≤5)	837 (60.8%)	637 (62.8%)	—	—
	Moderate (6–9)	166 (12.1%)	151 (14.9%)	—	—
	Moderate–severe (10–16)	180 (13.1%)	110 (10.8%)	—	—
	Severe (17–20)	16 (1.2%)	20 (2.0%)	—	—
	Very severe (>20)	18 (1.3%)	27 (2.7%)	—	—
IV tPA	Yes	147 (10.7%)	174 (17.1%)	—	—
	No	1,230 (89.3%)	841 (82.9%)	—	—
IA thrombectomy	Yes	126 (9.2%)	28 (2.8%)	—	—
	No	1,251 (90.8%)	987 (97.2%)	—	—
Stroke territory	Anterior	869 (63.1%)	553 (54.5%)	—	—
	Posterior	207 (15.0%)	238 (23.4%)	—	—
	Both	271 (19.7%)	151 (14.9%)	—	—
	Unknown	30 (2.2%)	73 (7.2%)	—	—
Large vessel disease	Yes	399 (29.0%)	251 (24.7%)	—	—
	No	832 (60.4%)	623 (61.4%)	—	—
	Unknown	146 (10.6%)	141 (13.9%)	—	—
Motor weakness	No	351 (25.5%)	326 (32.1%)	—	—
	Yes	1,026 (74.5%)	689 (67.9%)	—	—
Slurred speech	No	439 (31.9%)	414 (40.8%)	—	—
	Yes	938 (68.1%)	601 (59.2%)	—	—
Hemianopsia	No	1,315 (95.5%)	969 (95.5%)	—	—
	Yes	62 (4.5%)	46 (4.5%)	—	—
Impaired consciousness	No	1,290 (93.7%)	966 (95.2%)	—	—
	Yes	87 (6.3%)	49 (4.8%)	—	—
Dizziness	No	1,170 (85.0%)	797 (78.5%)	—	—
	Yes	207 (15.0%)	218 (21.5%)	—	—
Impaired sensitivity	No	1,266 (91.9%)	890 (87.7%)	—	—
	Yes	111 (8.1%)	125 (12.3%)	—	—

(Continued)

TABLE 1 (Continued)

Variable	Category	Hospital 1 (Training) (*n* = 1,377)	Hospital 2 (Validation) (*n* = 1,015)	*p* value	Statistical test
Facial palsy	No	1,220 (88.6%)	894 (88.1%)	—	—
	Yes	156 (11.3%)	121 (11.9%)		
	Missing	1 (0.1%)	0 (0.0%)		

Percentages are based on available observations. Missingness is reported in Supplementary Table S1A. Educational-level category 4 denotes unknown information and is markedly more frequent in Hospital 1.

TABLE 2 Interpretable six-factor logistic regression in Hospital 1.

Predictor	Adjusted OR	95% CI	*p* value
Rural residence	2.16	1.57–2.99	<0.001
Non-emergency channel	4.41	2.53–7.67	<0.001
Inter-facility referral	2.84	1.94–4.18	<0.001
Non-ambulance transport	3.22	2.20–4.72	<0.001
Absence of atrial fibrillation	3.79	1.95–7.36	<0.001
Absence of slurred speech	2.25	1.57–3.24	<0.001

random forest, −0.018 for XGBoost, −0.021 for LightGBM, and −0.003 for AdaBoost. Full-feature SMOTE logistic regression had similar AUC but better calibration than LASSO-plus-SMOTE logistic regression (Brier score, 0.125 versus 0.180).

Omitting SMOTE frequently improved discrimination and probability accuracy. With LASSO-selected features, no-SMOTE AUCs were 0.791 for AdaBoost, 0.789 for random forest, 0.781 for XGBoost, and 0.774 for logistic regression. The logistic-regression Brier score improved from 0.180 with SMOTE to 0.122 without SMOTE.

Calibration and recalibration

The original SMOTE-trained logistic model underestimated the observed delay risk (Figure 3C). Calibration-in-the-large was 1.101, the calibration intercept was 1.377, the slope was 0.793, and the Brier score was 0.180. Logistic regression without SMOTE had calibration-in-the-large 0.276, intercept 0.103, slope 1.197, and Brier score 0.122. Post-hoc logistic recalibration reduced the 10-fold cross-fitted Brier score to 0.123 and calibration-in-the-large to 0.002. Because recalibration used Hospital 2 outcomes, it requires confirmation in a third cohort.

Decision-curve analysis

The original SMOTE-trained logistic model exceeded both review-all and review-none strategies only at thresholds approximately 0.77–0.95; the no-SMOTE logistic model did so at approximately 0.40–0.95 (Figure 3D). At low thresholds, the high prevalence of delay favored reviewing all records. These curves concern retrospective selection of records for quality-improvement review, not treatment decisions.

Subgroup absolute risks

In Hospital 2, observed delay risk was 90.5% in rural versus 74.6% in urban residents (absolute difference, 15.8 percentage points; 95% CI, 10.9–20.3). Absolute differences were 21.5 points (95% CI, 18.4–24.5) for non-emergency versus emergency-channel presentation, 9.2 points (95% CI, 4.4–13.6) for referral versus direct presentation, and 36.3 points (95% CI, 23.8–48.7) for non-ambulance versus ambulance transport (Table 4).

Model interpretability

SHAP analysis was performed for XGBoost, which had the highest AUC in the prespecified pipeline (Figure 4). Mean absolute SHAP values are unitless magnitudes of average contribution on the model-output scale; they are not probabilities, odds ratios, or causal effects. The individual case plots were reformatted with separated panels and non-overlapping labels.

Discussion

In this two-hospital retrospective cohort study, we externally validated eight algorithms for prehospital delay in AIS. XGBoost had the highest AUC in the prespecified LASSO-plus-SMOTE analysis, but its difference from logistic regression was small and not statistically significant. Full-feature analyses did not reveal nonlinear gains masked by LASSO selection, and omitting SMOTE generally improved calibration and often discrimination. The principal contribution is therefore the external validation and interpretation of care-pathway factors associated with delay, rather than evidence that complex machine-learning algorithms are superior.

Comparison with previous studies

Previous studies have examined prehospital delay and related acute stroke care processes (7–9, 26–28). One recent prediction model used internal validation, which may overestimate performance (9). External validation can reveal lower discrimination than development studies, underscoring the importance of evaluation in an independent setting (11, 12). In the present study, the small differences among algorithms did not support a claim that complex models outperform logistic regression.

Our findings align with previous studies identifying rural residence (29, 30), non-ambulance transport (31, 32), and inter-facility referral (33) as factors associated with prehospital delay. We extend this

10.3389/fneur.2026.1867290

TABLE 3 Model performance across feature-selection and sampling configurations.

Feature set	Sampling	Model	External AUC (95% CI)	Brier	Sensitivity	Specificity	PPV	NPV
LASSO-selected	SMOTE	Logistic Regression	0.756 (0.718–0.797)	0.180	0.732	0.649	0.912	0.329
LASSO-selected	SMOTE	Random Forest	0.764 (0.726–0.800)	0.184	0.686	0.696	0.918	0.310
LASSO-selected	SMOTE	XGBoost	0.771 (0.734–0.807)	0.162	0.764	0.596	0.903	0.339
LASSO-selected	SMOTE	LightGBM	0.764 (0.726–0.801)	0.150	0.834	0.468	0.886	0.364
LASSO-selected	SMOTE	AdaBoost	0.759 (0.725–0.796)	0.213	0.723	0.626	0.905	0.314
LASSO-selected	SMOTE	SVC	0.740 (0.703–0.780)	0.185	0.724	0.655	0.912	0.325
LASSO-selected	SMOTE	MLP	0.725 (0.684–0.766)	0.198	0.706	0.614	0.900	0.297
LASSO-selected	SMOTE	Gaussian NB	0.741 (0.705–0.778)	0.260	0.636	0.702	0.913	0.281
Full candidate set	SMOTE	Logistic Regression	0.753 (0.716–0.793)	0.125	0.962	0.234	0.861	0.556
Full candidate set	SMOTE	Random Forest	0.759 (0.725–0.795)	0.176	0.762	0.550	0.893	0.319
Full candidate set	SMOTE	XGBoost	0.753 (0.714–0.792)	0.147	0.854	0.444	0.884	0.382
Full candidate set	SMOTE	LightGBM	0.744 (0.705–0.784)	0.137	0.897	0.310	0.865	0.379
Full candidate set	SMOTE	AdaBoost	0.756 (0.719–0.792)	0.210	0.845	0.439	0.881	0.364
Full candidate set	SMOTE	SVC	0.728 (0.686–0.769)	0.130	0.957	0.211	0.857	0.500
Full candidate set	SMOTE	MLP	0.706 (0.660–0.747)	0.132	0.976	0.094	0.842	0.444
Full candidate set	SMOTE	Gaussian NB	0.534 (0.508–0.564)	0.172	0.987	0.035	0.835	0.353
LASSO-selected	No SMOTE	Logistic Regression	0.774 (0.734–0.814)	0.122	0.979	0.146	0.850	0.581
LASSO-selected	No SMOTE	Random Forest	0.789 (0.751–0.824)	0.119	0.993	0.082	0.842	0.700
LASSO-selected	No SMOTE	XGBoost	0.781 (0.744–0.817)	0.121	0.970	0.199	0.857	0.576
LASSO-selected	No SMOTE	LightGBM	0.748 (0.703–0.790)	0.124	0.962	0.193	0.855	0.508
LASSO-selected	No SMOTE	AdaBoost	0.791 (0.754–0.826)	0.168	0.987	0.140	0.850	0.686
LASSO-selected	No SMOTE	SVC	0.716 (0.673–0.758)	0.130	0.983	0.105	0.844	0.562
LASSO-selected	No SMOTE	MLP	0.710 (0.667–0.751)	0.128	0.993	0.041	0.836	0.538

(Continued)

TABLE 3 (Continued)

Feature set	Sampling	Model	External AUC (95% CI)	Brier	Sensitivity	Specificity	PPV	NPV
LASSO-selected	No SMOTE	Gaussian NB	0.765 (0.730–0.797)	0.163	0.873	0.392	0.876	0.385
Full candidate set	No SMOTE	Logistic Regression	0.768 (0.729–0.806)	0.130	0.955	0.251	0.863	0.531
Full candidate set	No SMOTE	Random Forest	0.783 (0.744–0.819)	0.122	0.995	0.018	0.833	0.429
Full candidate set	No SMOTE	XGBoost	0.784 (0.749–0.819)	0.121	0.968	0.199	0.856	0.557
Full candidate set	No SMOTE	LightGBM	0.764 (0.722–0.804)	0.121	0.976	0.228	0.862	0.661
Full candidate set	No SMOTE	AdaBoost	0.785 (0.750–0.821)	0.175	0.980	0.205	0.859	0.673
Full candidate set	No SMOTE	SVC	0.690 (0.647–0.734)	0.131	0.994	0.082	0.842	0.737
Full candidate set	No SMOTE	MLP	0.659 (0.613–0.701)	0.167	0.903	0.269	0.859	0.359
Full candidate set	No SMOTE	Gaussian NB	0.678 (0.634–0.724)	0.168	0.981	0.053	0.836	0.360

Classification metrics use a probability threshold of 0.50. Calibration and DCA for all 32 configurations are reported in Supplementary Table S4.

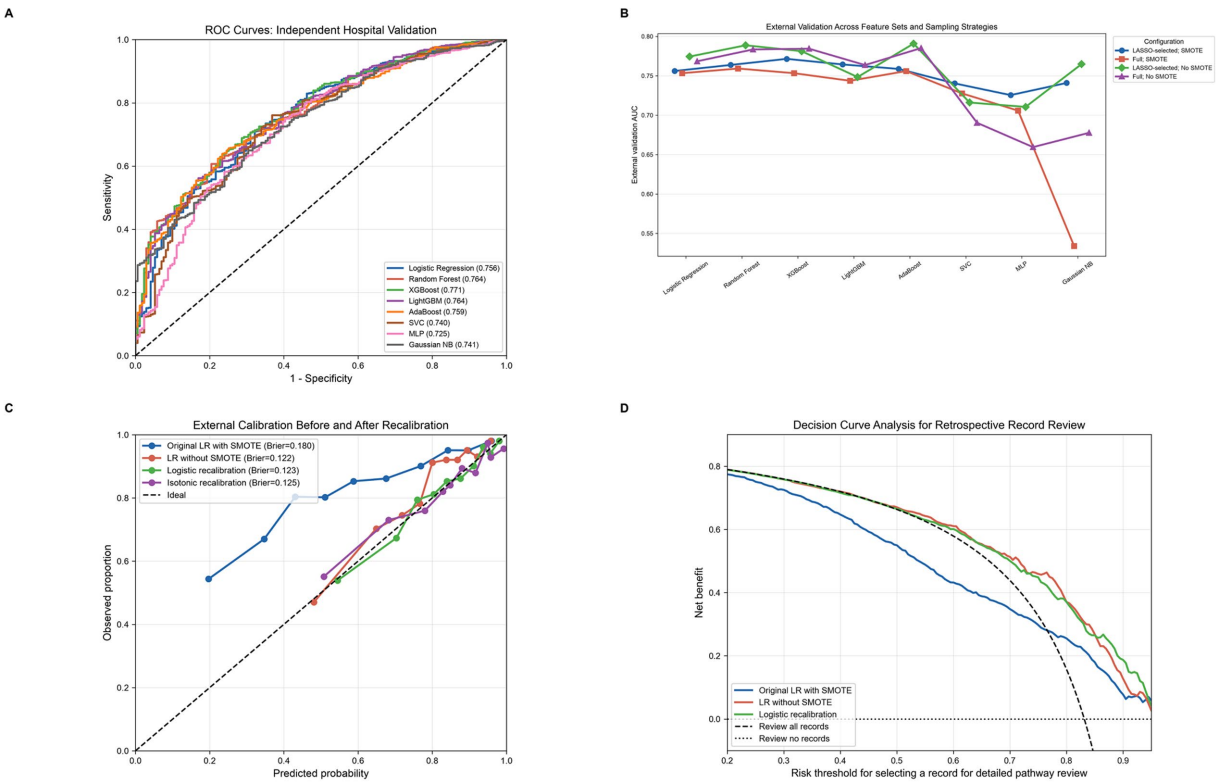

FIGURE 3
Model performance and sensitivity analyses in Hospital 2. **(A)** ROC curves for the eight prespecified LASSO-plus-SMOTE models. **(B)** External AUCs across LASSO-selected versus full candidate feature sets and SMOTE versus no-SMOTE training. **(C)** External calibration of the original SMOTE-trained logistic model, the no-SMOTE logistic model, and post-hoc logistic and isotonic recalibration. **(D)** Decision-curve analysis for retrospective selection of records for detailed quality-improvement review, compared with reviewing all records and reviewing no records.

TABLE 4 Observed subgroup absolute risks in Hospital 2.

Factor	Exposed, *n*/ delayed	Risk, %	Reference, *n*/ delayed	Risk, %	Absolute risk difference, percentage points (95% CI)	Risk ratio (95% CI)
Rural residence	546/494	90.5	469/350	74.6	15.8 (10.9–20.3)	1.21 (1.14–1.28)
Non-emergency channel	241/240	99.6	774/604	78.0	21.5 (18.4–24.5)	1.28 (1.23–1.32)
Inter-facility referral	226/204	90.3	789/640	81.1	9.2 (4.4–13.6)	1.11 (1.05–1.17)
Non-ambulance transport	948/811	85.5	67/33	49.3	36.3 (23.8–48.7)	1.74 (1.39–2.30)

Estimates are descriptive observed associations and should not be interpreted as causal effects.

literature by evaluating these factors in an independent hospital and by reporting subgroup absolute risks. Importantly, treatment-related variables were excluded to avoid reverse causality. The low VIFs and negligible correlation between referral and non-emergency-channel presentation reduce concern that these associations reflect multicollinearity, although the estimates remain observational rather than causal.

External validation and model transportability

A key strength of our study is the use of validation in an independent hospital. The non-random split by setting corresponds to TRIPOD Type 2b (13). The internal out-of-fold XGBoost AUC was 0.776 (95% CI, 0.746–0.807), and the external estimate of 0.771 lay within this interval. This finding supports limited transportability between the two participating hospitals, not broader geographic generalizability.

Calibration was suboptimal in the original SMOTE-trained logistic model, which underestimated the high observed risk. Although restricting SMOTE to training folds prevented leakage, oversampling changed the effective training prevalence and affected probability estimation. No-SMOTE models were better calibrated. Post-hoc recalibration improved probability agreement, but it was exploratory because it used Hospital 2 outcomes and requires validation in another cohort (34).

Clinical implications

Our analysis identifies several modifiable system-level factors that can inform quality-improvement priorities. Patients presenting through non-emergency channels had substantially higher observed delay risk, supporting public education that emphasizes activation of emergency services for stroke symptoms (35). Referred patients also had higher delay risk, highlighting opportunities to streamline transfer pathways and direct suspected stroke to capable centers (36). Rural residence and non-ambulance transport were associated with large absolute risk differences, supporting improved EMS access, telemedicine-enabled referral support, and targeted education in rural communities (37–39).

Because more than 80% of patients experienced delay, universal measures—multilingual stroke education, promotion of EMS activation, and streamlined referral protocols—remain appropriate. A model should not be used to withhold these measures. Its narrower potential value is to describe pathways with especially high observed risk when intensive resources are limited. The 36.3-point absolute difference associated with non-ambulance transport and the 21.5-point difference associated with non-emergency-channel presentation may help prioritize service review, EMS planning, and targeted education.

It is important to emphasize that several predictors, including NIHSS and stroke-characteristic variables, are recorded only after arrival. The framework cannot predict a delay before it occurs and should not be used for emergency dispatch, real-time triage, or individual bedside prognostication. The decision curves therefore represent retrospective selection of records for detailed quality-improvement review, not treatment of individual patients.

Model interpretability and clinical adoption

A major barrier to clinical adoption of machine-learning models is difficulty understanding how predictions are generated (40). SHAP analysis provides global and individual descriptions of model output. Mean absolute SHAP values are unitless average magnitudes of contribution on the model-output scale; they are not probabilities, odds ratios, or causal effects. Given the similar discrimination of complex and logistic models, transparent regression estimates and observed subgroup risks may be more useful for health-system communication than individual algorithmic explanations.

Methodological considerations

All preprocessing parameters were fitted exclusively on Hospital 1 and then applied to Hospital 2, preventing information leakage (41). SMOTE was restricted to training folds, never applied to validation folds or Hospital 2 (42). Feature selection and tuning were also restricted to Hospital 1. To address the concern that logistic LASSO selection could favor additive models, we repeated all eight algorithms with the full feature set and calculated permutation importance for nonlinear models. Full-feature modeling did not reveal an incremental discrimination advantage for nonlinear algorithms. The no-SMOTE analyses showed that correct within-fold implementation prevented leakage but did not prevent altered class prevalence from affecting calibration.

Limitations

Several limitations warrant consideration. First, both hospitals are located in Kashgar Prefecture, which limits generalizability to regions with different healthcare systems, stroke awareness, and socioeconomic conditions. Validation in other geographic settings is needed.

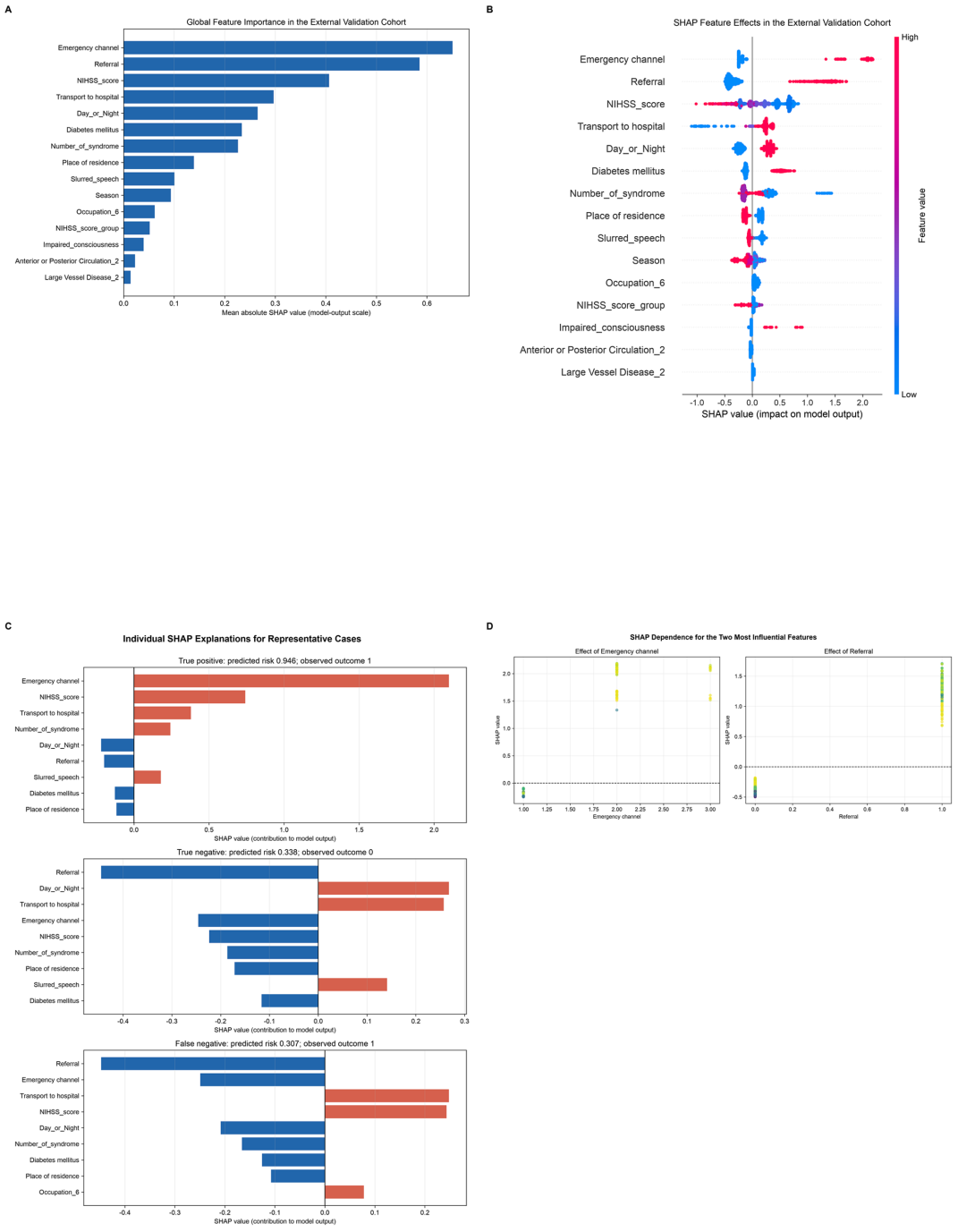

FIGURE 4
SHAP interpretation of XGBoost, which had the highest external AUC in the prespecified LASSO-plus-SMOTE pipeline, in Hospital 2 **(A)** Global mean absolute SHAP importance. **(B)** Beeswarm feature-effect distribution. **(C)** Separate individual SHAP contribution plots for a true-positive, true-negative, and false-negative case, with predicted risk and observed outcome. **(D)** Dependence plots for emergency-channel presentation and referral, the two highest-ranked features. Mean absolute SHAP values are unitless average magnitudes of contribution on the model-output scale.

Second, the high delay rates in both hospitals limit NPV and generalizability to settings with better prehospital systems. XGBoost NPV was 0.339 and the original SMOTE logistic NPV was 0.329. Thus, a low model score cannot reliably identify patients who will not experience delay, and universal interventions must not be withheld from a purported low-risk group.

Third, the framework requires information available only after hospital arrival, precluding individual pre-arrival prediction or real-time triage. It is limited to retrospective health-system profiling and quality-improvement planning.

Fourth, wake-up stroke patients were excluded because symptom-onset time could not be determined reliably. Fifth, potentially important variables were unavailable, including distance to hospital, socioeconomic status, health literacy, household decision-making, language concordance, and initial symptom recognition. Sixth, retrospective registry data may contain recording errors. Finally, post-hoc recalibration used validation outcomes and requires independent confirmation; subgroup differences are associational rather than causal, and small differences among the many model configurations should be interpreted cautiously.

Future directions

Prospective validation in other geographic and healthcare settings is needed. Implementation studies should evaluate whether targeted education, EMS optimization, and referral-pathway redesign reduce delay. Future pre-arrival models would need variables available to emergency dispatchers, such as caller-reported symptoms, location, and time of day. Qualitative studies may identify additional barriers not captured in registry data.

Conclusion

Prehospital delay affected more than four in five patients at both Kashgar hospitals. Complex algorithms did not provide a clear advantage over logistic regression, and full-feature analyses did not reveal nonlinear gains masked by LASSO selection. SMOTE worsened probability calibration and was unnecessary for useful discrimination. The framework should be used only for retrospective health-system profiling. Universal quality-improvement measures remain indicated, while subgroup differences in transport mode, emergency-channel access, rural residence, and referral may help prioritize intensive interventions. Prospective validation in another setting is required before probability-based implementation.

Data availability statement

Deidentified data and analysis code will be made available upon reasonable request to the corresponding author, subject to institutional review board approval and data-sharing agreements.' Patient-level data are not fully contained in the article or Supplementary Material.

Ethics statement

The studies involving human participants were approved by the Clinical Research Ethics Committee of Xinjiang Uygur Autonomous Region People's Hospital (approval No. KY2224110701) and were conducted in accordance with local legislation and institutional requirements. The participants provided written informed consent to participate in the study.

Author contributions

SZ: Writing – original draft, Investigation, Conceptualization, Data curation. AYu: Project administration, Writing – original draft, Methodology. HD: Project administration, Validation, Writing – review & editing, Investigation. LZ: Writing – review & editing, Data curation. AYa: Writing – review & editing, Resources. JW: Writing – review & editing, Resources. WY: Resources, Writing – review & editing. HY: Resources, Writing – review & editing. GM: Writing – review & editing, Resources. MA: Writing – review & editing, Resources. DK: Resources, Writing – review & editing. HL: Resources, Writing – review & editing, Funding acquisition, Validation, Project administration, Supervision.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0527802), the Central Government Guidance Fund for Local Sci-Tech Development (ZYYD2025QY21).

Acknowledgments

We thank the medical staff at all participating hospitals for their dedication to stroke care and data collection. We acknowledge the contributions of the Xinjiang Clinical Research Center for Stroke and Rare Neurological Disease.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that generative AI tools were used to assist with language editing, document-consistency checking, and code-assisted verification of analyses. All data, analytical outputs, interpretations, and text were critically reviewed and approved by the authors, who take full responsibility for the content. Generative AI was not used to fabricate data.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fneur.2026.1867290/full#supplementary-material

SUPPLEMENTARY TABLE S1A
Missing-data analysis.

SUPPLEMENTARY TABLE S1B
Participant flow.

SUPPLEMENTARY TABLE S2
Hyperparameter search spaces and selected values.

SUPPLEMENTARY TABLE S3
External calibration and recalibration.

SUPPLEMENTARY TABLE S4A
Calibration and DCA across all 32 configurations.

SUPPLEMENTARY TABLE S4B
Paired AUC comparisons.

SUPPLEMENTARY TABLE S4C
DCA interpretation.

SUPPLEMENTARY TABLE S5A
Six-factor logistic regression.

SUPPLEMENTARY TABLE S5B
Variance inflation factors.

SUPPLEMENTARY TABLE S5C
Spearman correlations.

SUPPLEMENTARY TABLE S6
Model-specific permutation importance.

SUPPLEMENTARY TABLE S7
Pairwise external AUC comparisons.

SUPPLEMENTARY TABLE S8
Apparent logistic recalibration.

SUPPLEMENTARY TABLE S9
Decision-curve threshold summaries.

SUPPLEMENTARY TABLE S10
Mean absolute SHAP importance.

SUPPLEMENTARY FIGURE S1
ODT distribution by hospital.

References

1. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, et al. Guidelines for the early Management of Patients with Acute Ischemic Stroke: 2019 update to the 2018 guidelines for the early Management of Acute Ischemic Stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*. (2019) 50:e344–418. doi: 10.1161/STR.0000000000000211

2. Emberson J, Lees KR, Lyden P, Blackwell L, Albers G, Bluhmki E, et al. Effect of treatment delay, age, and stroke severity on the effects of intravenous thrombolysis with alteplase for acute ischaemic stroke: a meta-analysis of individual patient data from randomised trials. *Lancet*. (2014) 384:1929–35. doi: 10.1016/S0140-6736(14)60584-5

3. Goyal M, Menon BK, van Zwam WH, Dippel DW, Mitchell PJ, Demchuk AM, et al. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. *Lancet*. (2016) 387:1723–31. doi: 10.1016/S0140-6736(16)00163-X

4. Evenson KR, Foraker RE, Morris DL, Rosamond WD. A comprehensive review of prehospital and in-hospital delay times in acute stroke care. *Int J Stroke*. (2009) 4:187–99. doi: 10.1111/j.1747-4949.2009.00276.x

5. Fassbender K, Balucani C, Walter S, Levine SR, Haass A, Grotta J. Streamlining of prehospital stroke management: the golden hour. *Lancet Neurol*. (2013) 12:585–96. doi: 10.1016/S1474-4422(13)70100-5

6. Lacy CR, Suh DC, Bueno M, Kostis JB. Delay in presentation and evaluation for acute stroke: stroke time registry for outcomes knowledge and epidemiology (STROKE). *Stroke*. (2001) 32:63–9. doi: 10.1161/01.str.32.1.63

7. Heo JH, Kim YD, Nam HS, Hong KS, Ahn SH, Cho HJ, et al. A computerized in-hospital alert system for thrombolysis in acute stroke. *Stroke*. (2010) 41:1978–83. doi: 10.1161/STROKEAHA.110.583591

8. Kesinger MR, Sequeira DJ, Buffalini S, Guyette FX. Comparing National Institutes of Health stroke scale among a stroke team and helicopter emergency medical service providers. *Stroke*. (2015) 46:575–8. doi: 10.1161/STROKEAHA.114.007850

9. Zhao J, Liu R. Prehospital and in-hospital delays in ischemic stroke care. *J Stroke Cerebrovasc Dis*. (2017) 26:2763–8. doi: 10.1016/j.jstrokecerebrovasdis.2017.06.053

10. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. (2016) 69:245–7. doi: 10.1016/j.jclinepi.2015.04.005

11. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. (2014) 14:40. doi: 10.1186/1471-2288-14-40

12. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol*. (2015) 68:25–34. doi: 10.1016/j.jclinepi.2014.09.007

13. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. (2015) 350:g7594–4. doi: 10.1136/bmj.g7594

14. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med*. (2016) 375:1216–9. doi: 10.1056/NEJMp1606181

15. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. (2019) 380:1347–58. doi: 10.1056/NEJMra1814259

16. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. (2018) 319:1317–8. doi: 10.1001/jama.2017.18391

17. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. (2019) 1:206–15. doi: 10.1038/s42256-019-0048-x

18. Hatano S. Experience from a multicentre stroke register: a preliminary report. *Bull World Health Organ*. (1976) 54:541–53.

19. Thomalla G, Simonsen CZ, Boutitie F, Andersen G, Berthezene Y, Cheng B, et al. MRI-guided thrombolysis for stroke with unknown time of onset. *N Engl J Med*. (2018) 379:611–22. doi: 10.1056/NEJMoa1804355

20. Addo J, Ayerbe L, Mohan KM, Crichton S, Sheldenkar A, Chen R, et al. Socioeconomic status and stroke: an updated review. *Stroke*. (2012) 43:1186–91. doi: 10.1161/STROKEAHA.111.639732

21. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. *ACM Trans Knowl Discov Data*. (2012) 6:1–21. doi: 10.1145/2382577.2382579

22. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B Stat Methodol*. (1996) 58:267–88. doi: 10.1111/j.2517-6161.1996.tb02080.x

23. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. (2010) 33:1–22. doi: 10.18637/jss.v033.i01

24. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority oversampling technique. *J Artif Intell Res*. (2002) 16:321–57. doi: 10.1613/jair.953

25. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. (2020) 2:56–67. doi: 10.1038/s42256-019-0138-9

26. Jin H, Zhu S, Wei JW, Wang J, Liu M, Wu Y, et al. Factors associated with prehospital delays in the presentation of acute stroke in urban China. *Stroke*. (2012) 43:362–70. doi: 10.1161/STROKEAHA.111.623512

27. Moser DK, Kimble LP, Alberts MJ, Alonzo A, Croft JB, Dracup K, et al. Reducing delay in seeking treatment by patients with acute coronary syndrome and stroke: a scientific statement from the American Heart Association Council on cardiovascular nursing and stroke council. *Circulation*. (2006) 114:168–82. doi: 10.1161/CIRCULATIONAHA.106.176040

28. Bae HJ, Kim DH, Yoo NT, Choi JH, Huh JT, Cha JK, et al. Prehospital notification from the emergency medical service reduces the transfer and intra-hospital processing times for acute stroke patients. *J Clin Neurol*. (2010) 6:138–42. doi: 10.3988/jcn.2010.6.3.138

29. Lichtman JH, Watanabe E, Allen NB, Jones SB, Dostal J, Goldstein LB. Hospital arrival time and intravenous t-PA use in US academic medical centers, 2001-2004. *Stroke*. (2009) 40:3845–50. doi: 10.1161/STROKEAHA.109.562660

30. Meurer WJ, Majersik JJ, Frederiksen SM, Kade AM, Sandretto AM, Scott PA. Provider perceptions of barriers to the emergency use of tPA for acute ischemic stroke: a qualitative study. *BMC Emerg Med*. (2011) 11:5. doi: 10.1186/1471-227X-11-5

31. Kleindorfer D, Kissela B, Schneider A, Woo D, Khoury J, Miller R, et al. Eligibility for recombinant tissue plasminogen activator in acute ischemic stroke: a population-based study. *Stroke*. (2004) 35:e27–9. doi: 10.1161/01.STR.0000109767.11426.17

32. Fonarow GC, Zhao X, Smith EE, Saver JL, Reeves MJ, Bhatt DL, et al. Door-to-needle times for tissue plasminogen activator administration and clinical outcomes in acute ischemic stroke before and after a quality improvement initiative. *JAMA*. (2014) 311:1632–40. doi: 10.1001/jama.2014.3203

33. Menon SC, Pandey DK, Morgenstern LB. Critical factors determining access to acute stroke care. *Neurology*. (1998) 51:427–32. doi: 10.1212/wnl.51.2.427

34. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. (2019) 17:230. doi: 10.1186/s12916-019-1466-7

35. Morgenstern LB, Staub L, Chan W, Wein TH, Bartholomew LK, King M, et al. Improving delivery of acute stroke therapy: the TLL Temple foundation stroke project. *Stroke*. (2002) 33:160–6. doi: 10.1161/hs0102.101990

36. Prabhakaran S, Ward E, John S, Lopes DK, Chen M, Temes RE, et al. Transfer delay is a major factor limiting the use of intra-arterial treatment in acute ischemic stroke. *Stroke*. (2011) 42:1626–30. doi: 10.1161/STROKEAHA.110.609750

37. Ebinger M, Winter B, Wendt M, Weber JE, Waldschmidt C, Rozanski M, et al. Effect of the use of ambulance-based thrombolysis on time to thrombolysis in acute ischemic stroke: a randomized clinical trial. *JAMA*. (2014) 311:1622–31. doi: 10.1001/jama.2014.2850

38. Schwamm LH, Holloway RG, Amarenco P, Audebert HJ, Bakas T, Chumbler NR, et al. A review of the evidence for the use of telemedicine within stroke systems of care: a scientific statement from the American Heart Association/American Stroke Association. *Stroke*. (2009) 40:2616–34. doi: 10.1161/STROKEAHA.109.192360

39. Patel MD, Rose KM, O'Brien EC, Rosamond WD. Prehospital notification by emergency medical services reduces delays in stroke evaluation: findings from the North Carolina stroke care collaborative. *Stroke*. (2011) 42:2263–8. doi: 10.1161/STROKEAHA.110.605857

40. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. *Proc Mach Learn Res*. (2019) 106:359–80.

41. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns (N Y)*. (2023) 4:100804. doi: 10.1016/j.patter.2023.100804

42. Santos MS, Soares JP, Abreu PH, Araujo HJ, Santos JAM. Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches. *IEEE Comput Intell Mag*. (2018) 13:59–76. doi: 10.1109/MCI.2018.2866730

43. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. (2017) 30:4765–74.

44. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*. (2015) 68:279–89. doi: 10.1016/j.jclinepi.2014.06.018

45. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak*. (2006) 26:565–74. doi: 10.1177/0272989X06295361

46. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. (1996) 49:1373–9. doi: 10.1016/s0895-4356(96)00236-3

47. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet*. (2007) 370:1453–7. doi: 10.1016/S0140-6736(07)61602-X

48. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. (2013) 10:e1001381. doi: 10.1371/journal.pmed.1001381

49. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA*. (2002) 99:6562–6. doi: 10.1073/pnas.102102699

50. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer (2009).

51. Pandian JD, Gall SL, Kate MP, Silva GS, Akinyemi RO, Ovbiagele BI, et al. Prevention of stroke: a global perspective. *Lancet*. (2018) 392:1269–78. doi: 10.1016/S0140-6736(18)31269-8