Biomolecules & Biomedicine

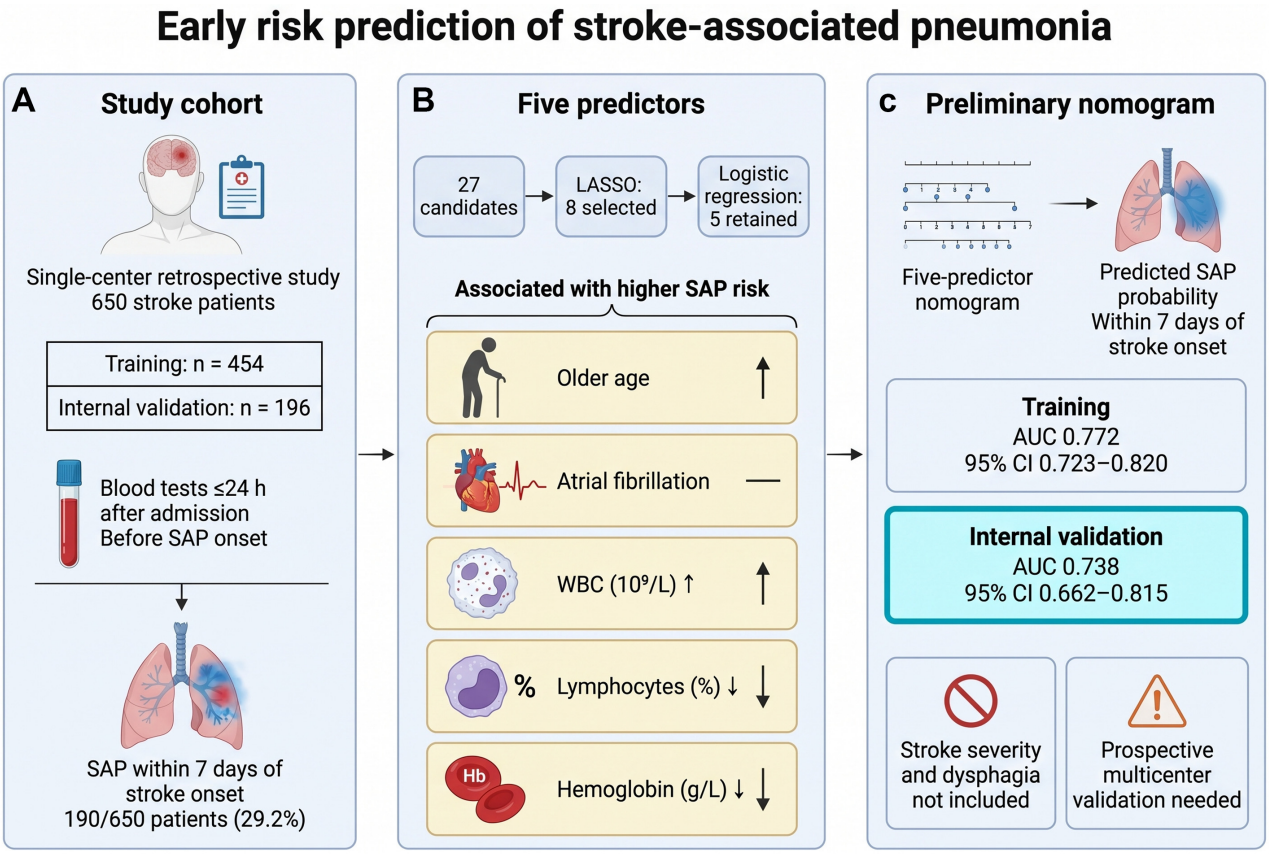
RESEARCH ARTICLE

Stroke-associated pneumonia risk prediction: Development and internal validation of a preliminary nomogram

Baichen Wang*

*Correspondence to Baichen Wang: WBC8060@126.com

Graphical Abstract



Created in BioRender. Mlaćo, Š (2026) https://BioRender.com/s2kon5d

Highlights

- Older age and atrial fibrillation are associated with pneumonia after stroke.
- Higher WBC and lower lymphocyte percentage and hemoglobin indicate higher risk.
- A five-predictor nomogram achieves an internal validation AUC of 0.738.
- Prospective multicenter validation is needed before clinical implementation.

In brief

Wang develops a preliminary nomogram to estimate pneumonia risk within seven days of stroke onset using retrospective data from 650 patients. Age, atrial fibrillation, white blood cell count, lymphocyte percentage, and hemoglobin provide moderate discrimination in internal validation. The model shows potential for early risk stratification, but omission of stroke severity and dysphagia, alongside the single-center design, necessitates prospective multicenter validation.

DOI: 10.17305/bb.2026.14918

Biomolecules & Biomedicine

RESEARCH ARTICLE

Stroke-associated pneumonia risk prediction: Development and internal validation of a preliminary nomogram

Baichen Wang*

Stroke-associated pneumonia (SAP) is a common complication of acute stroke, highlighting the need for early risk identification. This study aimed to develop and internally validate a preliminary SAP nomogram using routinely available clinical and hematological parameters. This single-center retrospective study included 650 patients randomly divided into training ($n = 454$) and internal validation ($n = 196$) cohorts. Baseline blood samples were obtained within 24 hours of admission and before SAP onset; the outcome was SAP within 7 days of stroke onset. Least absolute shrinkage and selection operator (LASSO) regression followed by multivariable logistic regression was used to select predictors and construct the nomogram. SAP occurred in 190 patients (29.2%). Older age, atrial fibrillation, higher white blood cell count, lower lymphocyte percentage, and lower hemoglobin were independently associated with SAP. The area under the receiver operating characteristic curve (AUC) was 0.772 in the training cohort and 0.738 in the internal validation cohort, with respective 95% confidence intervals of 0.723–0.820 and 0.662–0.815. The optimism-corrected training AUC was 0.762. The validation Brier score was 0.181, and decision curve analysis suggested potential clinical net benefit. This preliminary nomogram may support early SAP risk stratification. However, the single-center retrospective design and omission of key neurological predictors, including stroke severity and dysphagia, necessitate prospective multicenter validation before clinical implementation.

Keywords: Stroke-associated pneumonia, nomogram, prediction model, predictors.

Introduction

Stroke is the second leading cause of death worldwide. Stroke-associated pneumonia (SAP) is a common and severe complication following a stroke [1, 2]. Epidemiological studies report an SAP incidence ranging from 5% to 30% [3]. SAP not only exacerbates neurological deficits and delays functional recovery but also prolongs hospital stays, increases healthcare costs, and triples the 30-day mortality rate [4, 5]. Therefore, early and accurate identification of patients at high risk for SAP is crucial for improving clinical outcomes.

The pathophysiology of SAP involves multiple interacting factors. Neurological impairments, such as altered consciousness and dysphagia, significantly increase the risk of aspiration, creating a favorable environment for pulmonary infection [6]. Additionally, severe central nervous system injury triggers stroke-induced immunosuppression syndrome, which compromises systemic immunity, disrupts inflammatory homeostasis, and increases susceptibility to infection [7]. Currently, there is no unified consensus on the key predictors of SAP. Critically, early investigations lacked a clear diagnostic time window for SAP. This omission could introduce grouping bias and weaken the robustness of the findings. To address this, the revised Pneumonia in Stroke Consensus (PISCES) and the Centers for Disease Control and Prevention (CDC) recommend standardized criteria [8, 9]. They classify pneumonia diagnosed within 7 days of stroke onset as SAP; pneumonia occurring after 7 days is categorized as conventional hospital-acquired pneumonia. This standardized framework provides essential methodological support for developing reliable SAP prediction models.

Advanced diagnostic technologies in nanomedicine have shown high accuracy in complex diseases [10, 11]. However, using cutting-edge molecular biomarkers is often time-consuming and impractical in acute stroke units. Currently, several clinical tools assess SAP risk, including the age, atrial fibrillation, dysphagia, sex, and stroke severity (A2DS2), ISAN, preventive Antibacterial Therapy in Acute Ischemic Stroke (PANTHERIS), and acute ischemic stroke-associated pneumonia score (AIS-APS) scores [12–15]. While useful for risk stratification, their broad real-world implementation remains limited. Some scores require too many variables or rely on data that are slow to obtain, making them unsuitable for acute emergency settings. Furthermore, conventional scoring systems often convert continuous variables into categorical ones using manual cutoff points. This can lead to data loss and reduced predictive accuracy. Traditional multivariable logistic regression is statistically

Department of Neurology, Tongde Hospital of Zhejiang Province, Hangzhou, Zhejiang, China*.

*Correspondence to Baichen Wang: WBC8060@126.com

DOI: 10.17305/bb.2026.14918

© 2026 Wang. This article is available under a Creative Commons License (Attribution 4.0 International, as described at https://creativecommons.org/licenses/by/4.0/).

rigorous, but its abstract mathematical formulation makes it difficult to use rapidly at the bedside. Nomograms address this limitation by converting complex regression equations into intuitive visual graphs [16, 17]. By combining clinical characteristics and routine laboratory results, nomograms can help estimate individualized SAP probabilities [18]. Although previous studies have developed SAP nomograms, most have notable limitations. Many include redundant predictors, ignore updated SAP diagnostic criteria, or rely on specialized tests not routinely available in clinical practice.

This study identified independent predictors associated with SAP and constructed an individualized nomogram model. It aims to propose a promising and preliminary decision-support tool. The model could eventually assist clinicians in the early identification of SAP and provide objective evidence for individualized preventive interventions.

Materials and methods

Study design and population

This was a single-center, retrospective observational study. The study systematically collected clinical data from stroke patients admitted to Tongde Hospital of Zhejiang Province between January 2021 and December 2025. Inclusion criteria were: (1) Confirmed diagnosis of stroke according to current clinical guidelines [19]; (2) Time from symptom onset to hospital admission $\leq$ 48 h. Exclusion criteria were: (1) Confirmed pulmonary infection on admission or a history of chronic specific/nonspecific pulmonary inflammation; (2) Primary or secondary brain tumors documented by computed tomography (CT)/magnetic resonance imaging (MRI); (3) Severe hepatic or renal dysfunction; (4) End-stage disease or severe systemic comorbidities; (5) Missing core data for outcome adjudication (e.g., undetermined SAP status, missing stroke onset time); (6) Missing more than 20% of baseline clinical variables (145 patients were excluded for this reason); (7) Pneumonia onset occurring more than 7 days after stroke.

Diagnostic criteria and grouping for SAP

The primary outcome was the development of SAP within 7 days of stroke onset. SAP was diagnosed according to the clinical criteria established by the PISCES/CDC [8, 9]. Patients were diagnosed with SAP if they met all three of the following criteria (A, B, and C) within the first 7 days after stroke onset: Criterion A: At least one of the following: (1) Unexplained fever; (2) Leukopenia or leukocytosis (assessed via subsequent laboratory tests, distinct from baseline); (3) Altered mental status without an identifiable cause in patients aged $\geq$70 years. Criterion B: At least two of the following: (1) New purulent sputum, changes in sputum character within 24 h, increased respiratory secretions, or increased suctioning requirements; (2) New or worsening cough, dyspnea, or tachypnea; (3) Rales, crackles, or bronchial breath sounds; (4) Worsening gas exchange. Criterion C: At least one chest radiograph or CT scan showing new or progressive pulmonary infiltrates. Patients meeting these criteria formed the SAP group. Individuals who developed pneumonia after this 7-day window were excluded to prevent confounding bias. All remaining patients comprised the control group. To confirm the diagnosis, two independent neurologists, blinded to the study design, retrospectively reviewed all imaging and clinical data. Any discrepancies were resolved through consensus with another senior neurologist.

Data collection

Baseline clinical data were extracted from the electronic medical record system. Demographic characteristics included age and sex. Comorbidities included diabetes mellitus, hypertension, hyperlipidemia, atrial fibrillation, heart failure, coronary heart disease, chronic obstructive pulmonary disease (COPD), Alzheimer's disease, renal insufficiency, and organic mental disorders. Organic mental disorders were defined according to ICD-10 criteria (codes F00–F09). This category includes psychiatric or cognitive syndromes (e.g., dementia, delirium) attributed to brain dysfunction such as prior stroke, trauma, or neurodegeneration.

The first peripheral blood test results obtained within 24 h of admission were collected. Notably, the initial blood draws for all included patients were performed before the clinical onset of SAP. Patients who developed SAP before their initial blood draw were excluded. Analyzed hematological parameters included: red blood cell distribution width (RDW), white blood cell (WBC) count, neutrophil count and percentage (Neu, Neu%), lymphocyte count and percentage (Lym, Lym%), monocyte count and percentage (Mono, Mono%), platelet count (PLT), hemoglobin (HGB), and high-sensitivity C-reactive protein (hs-CRP). To evaluate systemic immune and inflammatory status, the study calculated five composite indices: Neutrophil-to-lymphocyte ratio (NLR); Platelet-to-lymphocyte ratio (PLR); Monocyte-to-lymphocyte ratio (MLR); Systemic inflammation response index (SIRI): (Neu $\times$ Mono)/Lym; Systemic immune-inflammation index (SII): (PLT $\times$ Neu)/Lym.

Ethical statement

This retrospective study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Tongde Hospital of Zhejiang Province (Approval No.: 2026-080-JY). Given the anonymous nature of the retrospective data collection, the requirement for informed consent was waived by the ethics committee.

Statistical analysis

Statistical analyses were performed using Statistical Package for the Social Sciences (SPSS) 25.0 (IBM Corp, Armonk, NY, USA) and R software (version 3.5.3, R Core Team, Vienna, Austria). A two-sided $P < 0.05$ was considered statistically significant. Patients were randomly assigned to a training cohort and an internal validation cohort at a 7:3 ratio. Missing values were present in RDW (9.5%) and hs-CRP (2.4%). To prevent data leakage, multiple imputation was performed after dataset splitting. The imputation was conducted independently within the training and validation cohorts, using the multiple imputation by chained equations (MICE) algorithm to generate 5 imputed datasets.

Model development (feature scaling, variable selection, and regression fitting) was performed exclusively on the training cohort. The internal validation cohort was only used for final performance evaluation. Normally distributed continuous variables (presented as mean ± standard deviation [SD]) were compared using independent-samples *t*-tests. Non-normally distributed data (presented as median and interquartile range [IQR]) were analyzed using nonparametric tests. Categorical variables (frequencies) were compared using Pearson's χ^2 test. Collinearity was assessed using the variance inflation factor (VIF). Variables with a tolerance ≤ 0.2 or VIF ≥ 5 were considered to have severe multicollinearity.

For model development, sample size was justified based on the events per variable (EPV) criterion. Initially, 27 candidate predictors were considered. These variables were subjected to Least Absolute Shrinkage and Selection Operator (LASSO) regression for feature selection. The optimal penalization parameter (λ) was determined using 10-fold cross-validation. The value applying the 1-standard error rule (λ_{1se}) was selected to ensure a parsimonious model. The selected variables were then entered into a multivariable logistic regression analysis. A backward stepwise elimination process, based on a P value threshold of < 0.05, was subsequently performed to establish the final individualized predictive model.

The model's discriminative performance was assessed using receiver operating characteristic (ROC) curves and the area under the curve (AUC). Calibration curves were generated to visualize the agreement between predicted and observed probabilities. To correct for over-optimism, the bootstrap method (1,000 resamples) was applied to the training cohort. Specifically, the bootstrap procedure was directly applied to the final fixed model to evaluate calibration and optimism-adjusted discrimination, rather than repeating the entire variable selection process. Standard grouped observation plotting was applied to the validation cohort. Goodness-of-fit was assessed using the Hosmer–Lemeshow test and Spiegelhalter's *Z*-test. Finally, decision curve analysis (DCA) was performed to evaluate the model's potential clinical net benefit.

Results

Baseline characteristics of study subjects

A total of 650 stroke patients were enrolled and divided into the SAP group ($n = 190$) and the control group ($n = 460$). The training cohort ($n = 454$) experienced 131 SAP events (28.9%), while the internal validation cohort ($n = 196$) had 59 SAP events (30.1%). As detailed in Table 1, sex distribution did not significantly differ between the groups ($P = 0.766$). However, SAP group patients were significantly older ($P < 0.001$). Conventional inflammatory and hematological markers (RDW, WBC count, Neu, Neu%, Mono, PLT, and hs-CRP) were all significantly elevated in the SAP group ($P \leq 0.001$). Similarly, composite immune-inflammatory indices (NLR, PLR, MLR, SIRI, and SII) were markedly higher in the SAP group ($P < 0.001$). Conversely, Lym ($P = 0.009$), Lym% ($P < 0.001$), and HGB ($P < 0.001$) were significantly lower in the SAP group. Furthermore, the SAP group exhibited higher rates of atrial fibrillation ($P < 0.001$), heart failure ($P < 0.001$), Alzheimer's disease ($P = 0.002$), renal insufficiency ($P = 0.007$), and organic mental disorders ($P = 0.008$). No significant between-group differences were observed for diabetes mellitus, hypertension, hyperlipidemia, and COPD prevalence.

Feature selection

To identify predictive variables and address multicollinearity, all 27 demographic and clinical variables from Table 1 were entered into the LASSO regression model within the training cohort. Figure 1A illustrates how the regression coefficients of individual variables changed with the penalty parameter, $-\log(\lambda)$. As penalty strength increased, coefficients gradually shrank, though some remained non-zero. Figure 1B presents the parameter selection process using 10-fold cross-validation, with binomial deviance as the evaluation metric. The two vertical dashed lines indicate the minimum binomial deviance (λ_{min}) and (λ_{1se}). Using the λ_{1se} criterion, eight variables with non-zero coefficients were selected as candidate predictors: age, RDW, WBC count, Lym%, Mono, HGB, atrial fibrillation, and organic mental disorders.

Multivariable logistic regression

The eight candidate predictors identified by LASSO regression were subsequently analyzed using multivariable logistic regression. Following a backward stepwise elimination process, five variables were ultimately retained and confirmed as independent predictors of SAP (Table 2). Age (odds ratio [OR] = 1.031, 95% confidence interval [CI]: 1.014–1.048, $P < 0.001$), WBC count (OR = 1.248, 95% CI: 1.120–1.392, $P < 0.001$), and atrial fibrillation (OR = 2.221, 95% CI: 1.218–4.051, $P = 0.009$) had ORs greater than 1, indicating that higher values or the presence of these factors predict an increased probability of SAP. In contrast, Lym% (OR = 0.966, 95% CI: 0.939–0.994, $P = 0.018$) and HGB (OR = 0.980, 95% CI: 0.969–0.992, $P < 0.001$) had ORs less than 1. This inverse relationship signifies that lower levels of Lym% and HGB are significant predictors for a higher probability of developing SAP. Finally, collinearity diagnostics confirmed the absence of multicollinearity, with VIF values for these five predictors ranging from 1.066 to 1.335.

Development of the individualized nomogram

Based on these five independent predictors, the study constructed a nomogram to estimate the probability of SAP in stroke patients (Figure 2). In this nomogram, each predictor is assigned a specific point value. A patient's total score is calculated by summing the points from all variables. The predicted probability of SAP is then determined by projecting the total score vertically downward to the "Risk of SAP" axis. The model intercept (β_0) was −1.643. Full regression coefficients (β) for each predictor are detailed in Table 2. The exact probability of developing SAP for an individual patient can be calculated using the following equation:

$$P = \frac{1}{1+e^{-(-1.643+0.030\times \text{Age}+0.222\times \text{WBC}-0.035\times Lym\%-0.020\times \text{HGB}+0.798\times \text{Atrial Fibrillation})}}$$

Table 1. Baseline demographic and clinical characteristics of patients with stroke

Variable	SAP ($n = 190$)	Control ($n = 460$)	t/χ2/z	*P* value
Age, years	78.2 ± 13.3	70.7 ± 16.1	5.67	**<0.001***
Male sex, *n* (%)	122 (64.2%)	301 (65.4%)	0.089[a]	0.766
RDW, %	13.2 [12.6, 14]	12.8 [12.4, 13.4]	−4.753	**<0.001***
WBC, 10^9/L	7.3 [5.8, 9.3]	6.1 [5.1, 7.5]	−5.541	**<0.001***
Neu, 10^9/L	5.0 [3.9, 7]	4.0 [3.1, 5.2]	−6.341	**<0.001***
Neu%	70.4 ± 11.2	65.5 ± 9.8	5.656	**<0.001***
Lym, 10^9/L	1.4 ± 0.6	1.5 ± 0.6	−2.638	**0.009***
Lym%	19.0 ± 9.0	24.4 ± 8.9	−7.027	**<0.001***
Mono, 10^9/L	0.5 [0.4, 0.6]	0.4 [0.3, 0.5]	−5.289	**<0.001***
Mono%	6.8 ± 2.3	6.7 ± 1.8	0.851	0.395
PLT, 10^9/L	224.5 [168.8, 287]	203 [165, 245.8]	−3.37	**0.001***
hs-CRP	7.2 [2.2, 26.8]	3.6 [1.2, 13.9]	−3.829	**<0.001***
HGB, g/L	116.8 ± 22.1	128.3 ± 20.0	−6.463	**<0.001***
NLR	4.1 [2.8, 6.3]	2.7 [1.9, 3.9]	−6.69	**<0.001***
PLR	184.2 [136.9, 235.9]	137.5 [105.4, 183.5]	−5.832	**<0.001***
MLR	0.4 [0.3, 0.5]	0.3 [0.2, 0.4]	−7.313	**<0.001***
SIRI	1.9 [1.2, 3.4]	1.1 [0.7, 1.8]	−7.882	**<0.001***
SII	910.1 [562, 1564.2]	535.5 [360.6, 851.1]	−7.353	**<0.001***
Diabetes, *n* (%)	76 (40.0%)	182 (39.6%)	0.011[a]	0.918
Hypertension, *n* (%)	141 (74.2%)	370 (80.4%)	3.099[a]	0.078
Hyperlipidemia, *n* (%)	20 (10.5%)	56 (12.2%)	0.354[a]	0.552
Atrial fibrillation, *n* (%)	49 (25.8%)	39 (8.5%)	34.424[a]	**<0.001***
Heart failure, *n* (%)	20 (10.5%)	11 (2.4%)	19.592[a]	**<0.001***
COPD, *n* (%)	8 (4.2%)	8 (1.7%)	3.421[a]	0.064
Alzheimer's disease, *n* (%)	17 (8.9%)	15 (3.3%)	9.289[a]	**0.002**
Renal insufficiency, *n* (%)	36 (18.9%)	51 (11.1%)	7.166[a]	**0.007**
Organic mental disorders, *n* (%)	12 (6.3%)	10 (2.2%)	7.054[a]	**0.008**

Mean ± SD for continuously distributed data; median (IQR) for non-normally distributed data; frequencies (*n*) for categorical variables. Differences between the two groups were compared using Student's *t*-test (t), the Mann-Whitney *U* test (z), or the chi-square test (χ^2). [a]Indicates the chi-square (χ^2) value. *Indicates statistical significance ($P < 0.05$). **Abbreviations:** COPD, chronic obstructive pulmonary disease; HGB, hemoglobin; hs-CRP, high-sensitivity C-reactive protein; IQR, interquartile range; Lym, lymphocyte count; Lym%, lymphocyte percentage; MLR, monocyte-to-lymphocyte ratio; Mono, monocyte count; Mono%, monocyte percentage; Neu, neutrophil count; Neu%, neutrophil percentage; NLR, neutrophil-to-lymphocyte ratio; PLR, platelet-to-lymphocyte ratio; PLT, platelet count; RDW, red blood cell distribution width; SAP, stroke-associated pneumonia; SD, standard deviation; SII, systemic immune-inflammation index; SIRI, systemic inflammation response index; WBC, white blood cell count.

Table 2. Multivariable logistic regression analysis in the training cohort

Variable	β coefficient	OR	95% CI	*P* value	VIF
Age (years)	0.030	1.031	1.014–1.048	<0.001*	1.140
WBC (10^9/L)	0.222	1.248	1.120–1.392	<0.001*	1.261
Lym%	−0.035	0.966	0.939–0.994	0.018*	1.335
HGB (g/L)	−0.020	0.980	0.969–0.992	<0.001*	1.139
Atrial fibrillation	0.798	2.221	1.218–4.051	0.009*	1.066

*Indicates statistical significance ($P < 0.05$). **Abbreviations:** CI, confidence interval; HGB, hemoglobin; Lym%, lymphocyte percentage; OR, odds ratio; VIF, variance inflation factor; WBC, white blood cell count.

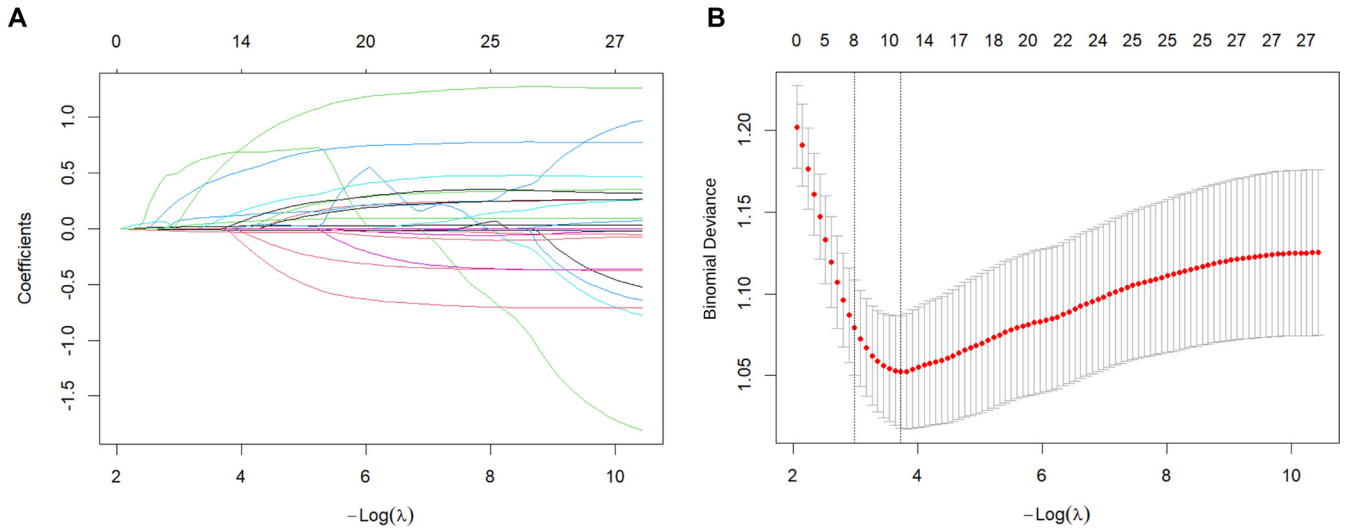


Figure 1. Selection of candidate predictors of stroke-associated pneumonia using LASSO regression. (A) Coefficient profiles of the 27 candidate predictors in the training cohort, plotted against −log(λ). Each colored curve represents an individual predictor; increasing the penalty parameter λ, corresponding to movement toward the left, shrinks coefficients toward zero. (B) Selection of λ using 10-fold cross-validation. Red points represent the mean cross-validated binomial deviance, and gray error bars indicate variability around these estimates. The vertical dotted lines mark λ_{1SE} (left) and λmin (right). The λmin value minimizes the mean binomial deviance, whereas λ_{1SE} is the largest λ yielding a deviance within one SE of the minimum. The λ_1SE criterion was applied, retaining eight predictors with non-zero coefficients for subsequent multivariable logistic regression. Numbers along the upper axes indicate the number of predictors with non-zero coefficients. **Abbreviations**: LASSO, least absolute shrinkage and selection operator; SE, standard error.

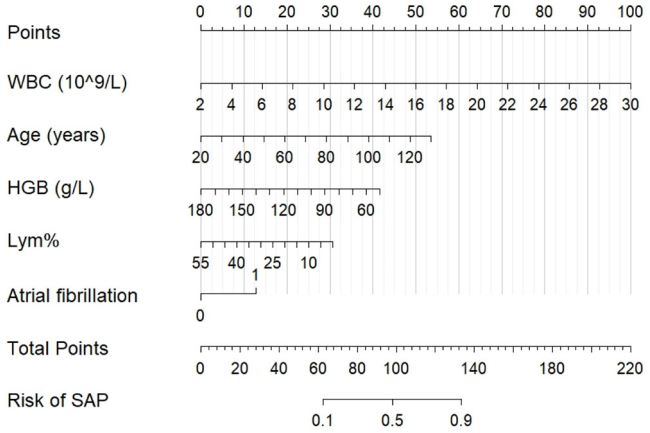


Figure 2. Nomogram for predicting the risk of SAP. The nomogram estimates the probability of SAP within 7 days of stroke onset using five predictors retained in the final multivariable logistic regression model: WBC, age, HGB, Lym%, and atrial fibrillation. WBC is expressed in 10^9/L, age in years, and HGB in g/L; atrial fibrillation is coded as 0 (absent) or 1 (present). To use the nomogram, locate the patient's value on each predictor axis and project vertically upward to the "Points" scale. Sum the points assigned to all five predictors, locate the sum on the "Total Points" axis, and project vertically downward to the "Risk of SAP" axis to obtain the predicted probability. **Abbreviations**: HGB, hemoglobin; Lym%, lymphocyte percentage; SAP, stroke-associated pneumonia; WBC, white blood cell count.

Performance evaluation of the predictive model

The nomogram's performance was initially evaluated in the training cohort. ROC curve analysis revealed an AUC of 0.772 (95% CI: 0.723–0.820) (Figure 3A). Utilizing an optimal probability cutoff of 0.268 (determined by the maximum Youden Index), the model demonstrated a positive predictive value (PPV) of 48.9%, a negative predictive value (NPV) of 85.3%, and an overall diagnostic accuracy of 70.3%. To assess model stability and mitigate potential overfitting, internal validation was conducted via bootstrap resampling. The bootstrap-calculated optimism bias was minimal (0.010), resulting in an optimism-corrected AUC of 0.762. The calibration curve, generated using bootstrap resampling ($B = 1{,}000$), closely aligned with the ideal diagonal line, showing a mean absolute error (MAE) of 0.011 (Figure 3B). Furthermore, DCA indicated that within a threshold probability range of 0.05–0.70, the model's net benefit surpassed both the "treat-all" and "treat-none" strategies (Figure 3C).

The model underwent further evaluation in an internal validation cohort ($n = 196$). ROC curve analysis yielded an AUC of 0.738 (95% CI: 0.662–0.815) (Figure 3D). The calibration curve demonstrated strong agreement between the nomogram-predicted probabilities and observed outcomes (Figure 3E). The model achieved a Brier score of 0.181, a calibration intercept of −0.221, and a calibration slope of 0.724. DCA confirmed that within a threshold range of 0.06–0.68, the nomogram's net benefit was superior to both "treat-all" and "treat-none" strategies (Figure 3F). Overall goodness-of-fit was confirmed in both the training cohort (Hosmer-Lemeshow test, $P = 0.344$) and the validation cohort (Spiegelhalter's Z-test, $P = 0.151$).

Discussion

This study identified five independent predictors and developed a preliminary nomogram model for predicting SAP. The model incorporates age, WBC count, Lym%, HGB, and atrial fibrillation. In the training cohort, the model achieved an AUC

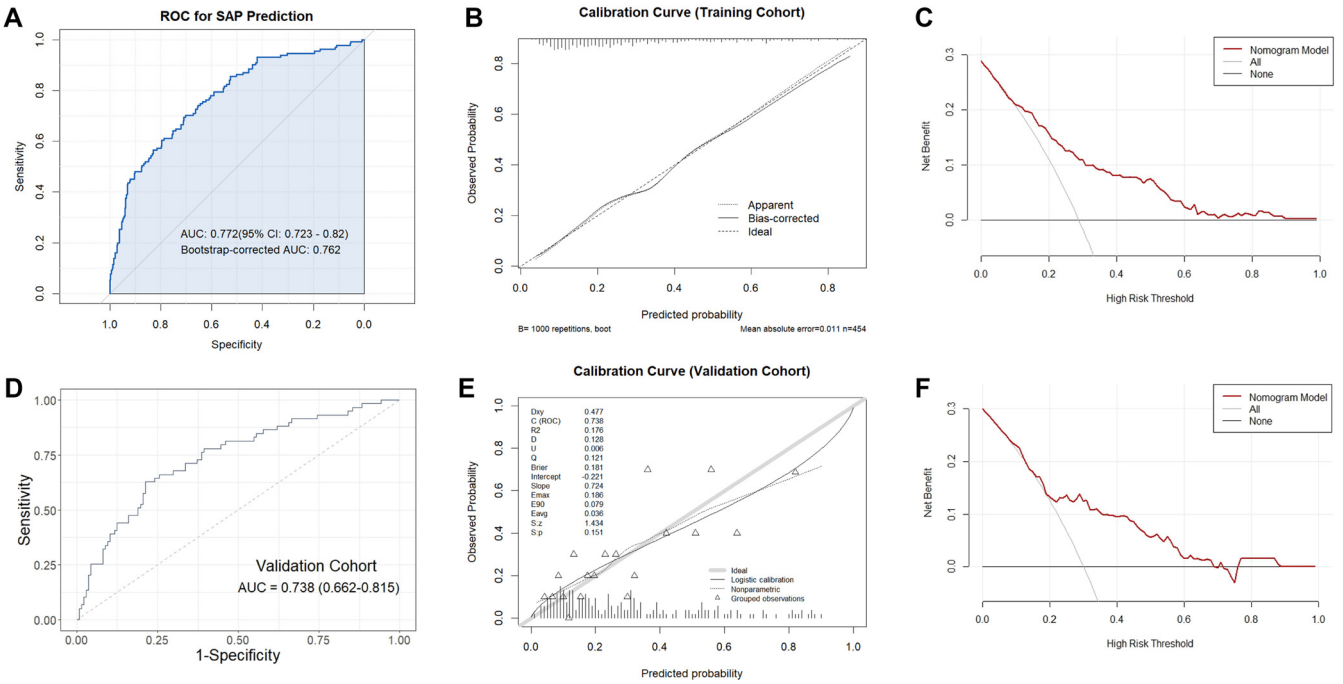


Figure 3. Performance of the nomogram for predicting SAP in the training and internal validation cohorts. (A) ROC curve in the training cohort. The apparent AUC was 0.772 (95% CI, 0.723–0.820), and the bootstrap optimism-corrected AUC was 0.762. (B) Calibration plot in the training cohort, showing the apparent and bias-corrected curves relative to the ideal diagonal. Bias correction used 1,000 bootstrap resamples of the final model, yielding an MAE of 0.011. (C) DCA in the training cohort. (D) ROC curve in the internal validation cohort, with an AUC of 0.738 (95% CI, 0.662–0.815). (E) Calibration plot in the internal validation cohort, showing logistic and nonparametric calibration curves and grouped observations (triangles). The Brier score was 0.181, with a calibration intercept of −0.221 and a calibration slope of 0.724. In both calibration plots, the ideal diagonal represents perfect agreement between predicted and observed probabilities. (F) DCA in the internal validation cohort. In both DCA plots, the red line represents the nomogram, the descending gray line represents the "treat-all" strategy, and the horizontal line at zero represents the "treat-none" strategy. DCA indicated greater net benefit for the nomogram than for both reference strategies across threshold probabilities of 0.05–0.70 in the training cohort and 0.06–0.68 in the internal validation cohort. **Abbreviations**: AUC, area under the curve; CI, confidence interval; DCA, decision curve analysis; MAE, mean absolute error; ROC, receiver operating characteristic; SAP, stroke-associated pneumonia.

of 0.772 (95% CI: 0.723–0.820), and in the validation cohort, the AUC was 0.738 (95% CI: 0.662–0.815). These results indicate good discriminative ability in identifying patients at high risk of SAP. Calibration curves demonstrated strong agreement between predicted and observed probabilities, signifying stable performance. Moreover, DCA verified the model's potential clinical net benefit across a threshold range of 0.05–0.70 in the training cohort and 0.06–0.68 in the validation cohort, highlighting its promise for application in emergency departments and stroke units.

Well-established tools like the A2DS2 or ISAN scores often necessitate specific neurological functional assessments or formal swallowing evaluations. In busy emergency departments or resource-limited settings, such detailed assessments may face delays or be subject to inter-rater variability. In contrast, this nomogram is constructed exclusively from routine demographic and objective laboratory parameters obtained immediately upon admission. By relying solely on universally accessible data, this model offers a promising and practical alternative for early SAP risk stratification. Additionally, this study applied standardized diagnostic criteria consistent with PISCES and CDC recommendations, which effectively reduced interference from hospital-acquired pneumonia and enhanced the reliability of our findings.

The study identified advanced age (OR = 1.031) and atrial fibrillation (OR = 2.221) as independent predictors of SAP, aligning with previous reports [20–24]. Recent clinical evidence underscores that acute stroke patients aged 85 years or older, particularly females, represent a highly vulnerable subpopulation facing exceptionally high in-hospital mortality driven by medical complications, including respiratory events [25]. Elderly patients frequently exhibit immunosenescence, characterized by degenerative changes in both innate and adaptive immunity. Their physiological defense mechanisms, such as cough and swallowing reflexes, are typically weakened. Poor oral hygiene and nutritional status further increase oropharyngeal bacterial colonization and aspiration risk. Given that pneumonia in elderly patients often presents atypically and is challenging to detect early, including age in the predictive model aids in identifying these vulnerable individuals proactively [26]. Furthermore, atrial fibrillation significantly elevates the risk of cardiogenic cerebral embolism, often leading to large vessel occlusion and extensive cerebral infarction with repeated ischemic involvement. This results in severe neurological deficits, indirectly increasing the risk of SAP [27–29].

Although composite inflammatory indices showed significant differences in univariate analysis, WBC count and Lym% demonstrated more stable predictive performance after

adjusting for multicollinearity. Elevated WBC count (OR = 1.248) and reduced Lym% (OR = 0.966) were significantly associated with an increased probability of SAP. An elevated WBC count may directly reflect enhanced systemic inflammation triggered by infection. Conversely, a lower Lym% indicates post-stroke immunosuppression, predisposing patients to secondary pulmonary infections [30–32]. Beyond the localized inflammatory response of pneumonia, the systemic inflammatory status of acute stroke patients warrants close attention. Stroke patients with concomitant polyvascular atherothrombotic disease exhibit notably higher rates of mortality [33]. Therefore, the inflammatory indicators in the nomogram may not only reflect susceptibility to pulmonary infection but may also serve as surrogate markers for the underlying systemic atherothrombotic burden, collectively contributing to adverse outcomes. Additionally, hepcidin upregulation mediated by IL-6 inhibits HGB production [34]. Thus, low HGB partially reflects the systemic inflammatory burden. A 10-year cross-sectional study using a national database also confirmed anemia as a major predictor of SAP [35]. Consistent with this, the study showed that reduced HGB (OR = 0.980) predicts a higher probability of SAP. This reduction may be linked to insufficient nutritional intake and chronic blood loss, which worsen overall frailty and create favorable conditions for pulmonary infection.

This study has several limitations. First, due to its retrospective nature, several clinically crucial predictors of SAP were unavailable or had substantial missing data in the electronic medical records system. These missing variables include stroke severity (admission National Institutes of Health Stroke Scale [NIHSS] scores), the presence and severity of dysphagia, level of consciousness (Glasgow Coma Scale), specific infarct locations, formalized aspiration risk assessments, and airway protection interventions (nasogastric tube placement and mechanical ventilation). Because swallowing dysfunction, bulbar palsy, and altered consciousness are fundamental mechanical pathways leading to aspiration pneumonia, the absence of these core neurological variables constitutes a major limitation. Consequently, the study could not directly compare the predictive performance (AUC and DCA) of our nomogram with established models like A2DS2 and ISAN within the same cohort. The model predominantly reflects systemic inflammatory, hematological, and demographic susceptibility, rather than a complete pathophysiological profile. Second, the single-center, retrospective design inherently introduces potential selection and information biases. There may also be variability in the exact timing of blood draws relative to stroke onset, as well as differences in clinical management across different physicians. Therefore, this model must be regarded as preliminary. Future large-scale, multicenter prospective studies are imperative.

Conclusion

In conclusion, this study preliminarily identified five readily available clinical variables as independent predictors of SAP in stroke patients: age, atrial fibrillation, WBC count, Lym%, and HGB. Based on these predictors, a promising nomogram was developed for early risk stratification. However, given the retrospective and single-center design, this model requires prospective, multicenter validation before broad clinical implementation. Future research should explore the integration of dynamic biomarker monitoring and advanced machine learning algorithms.

Conflicts of interest: Author declares no conflicts of interest.

Funding: Author received no specific funding for this work.

Data availability: The data presented in this study are available on request from the corresponding author.

Submitted: 20 August 2026
Accepted: 09 September 2026
Published online: 11 September 2026

References

[1] Szylińska A, Bott-Olejnik M, Wańkowicz P, Karoń D, Rotter I, Kotfis K. A novel index in the prediction of pneumonia following acute ischemic stroke. Int J Environ Res Public Health 2022;19(22):15306. https://doi.org/10.3390/ijerph192215306.

[2] GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Neurol 2021;20(10):795–820. https://doi.org/10.1016/S1474-4422(21)00252-0.

[3] Westendorp WF, Nederkoorn PJ, Vermeij JD, Dijkgraaf MG, van de Beek D. Post-stroke infection: a systematic review and meta-analysis. BMC Neurol 2011;11:110. https://doi.org/10.1186/1471-2377-11-110

[4] Ali AN, Howe J, Majid A, Redgrave J, Pownall S, Abdelhafiz AH. The economic cost of stroke-associated pneumonia in a UK setting. Top Stroke Rehabil 2018;25(3):214–23. https://doi.org/10.1080/10749357.2017.1398482.

[5] Wang J, Yang C, Zhang R, et al. Development and validation of a predictive model for stroke associated pneumonia in patients after thrombectomy for acute ischemic stroke. Front Med (Lausanne) 2024;11:1370986. https://doi.org/10.3389/fmed.2024.1370986.

[6] Zhang L, Wang Q, Li Y, Fang Q, Tang X. Individualized prediction of stroke-associated pneumonia for patients with acute ischemic stroke. Front Neurol 2025;16:1505270. https://doi.org/10.3389/fneur.2025.1505270.

[7] Yan D, Dai C, Xu R, Huang Q, Ren W. Predictive ability of systemic inflammation response index for the risk of pneumonia in patients with acute ischemic stroke. Gerontology 2023;69(2):181–8. https://doi.org/10.1159/000524759.

[8] Smith CJ, Kishore AK, Vail A, et al. Diagnosis of stroke-associated pneumonia: recommendations from the Pneumonia in Stroke Consensus Group. Stroke 2015;46(8):2335–40. https://doi.org/10.1161/STROKEAHA.115.009617

[9] Horan TC, Andrus M, Dudeck MA. CDC/NHSN surveillance definition of health care-associated infection and criteria for specific types of infections in the acute care setting. Am J Infect Control 2008;36(5):309–32. https://doi.org/10.1016/j.ajic.2008.03.002.

[10] Zhang X, Kong X, He X, et al. Nanoengineered bacterial extracellular vesicles for the photo-chemo programmed therapy to treat melanoma. Chem Eng J 2025;522:167807. https://doi.org/10.1016/j.cej.2025.167807.

[11] Zhang X, Tan Y, He X, et al. The diagnostic accuracy of exosomes for glioma: a meta-analysis. Biomol Biomed 2025;25(3):541–52. https://doi.org/10.17305/bb.2024.11268.

[12] Ji R, Shen H, Pan Y, et al. Novel risk score to predict pneumonia after acute ischemic stroke. Stroke 2013;44(5):1303–9. https://doi.org/10.1161/STROKEAHA.111.000598

[13] Smith CJ, Bray BD, Hoffman A, et al. Can a novel clinical risk score improve pneumonia prediction in acute stroke care? A UK multicenter cohort study. J Am Heart Assoc 2015;4(1):e001307. https://doi.org/10.1161/JAHA.114.001307.

[14] Harms H, Grittner U, Dröge H, Meisel A. Predicting post-stroke pneumonia: the PANTHERIS score. Acta Neurol Scand 2013;128(3):178–84. https://doi.org/10.1111/ane.12095.
[15] Hoffmann S, Harms H, Ulm L, et al. Stroke-induced immunodepression and dysphagia independently predict stroke-associated pneumonia – the PREDICT study. J Cereb Blood Flow Metab 2017;37(12):3671–82. https://doi.org/10.1177/0271678X16671964.
[16] Yu Z, Pan Y, Wang B. A prediction model for the risk of gastrointestinal bleeding associated with antiplatelet therapy in patients with ischemic stroke. Neurol Asia 2025;30(3):681–90. https://doi.org/10.54029/2025hmv.
[17] Wang X, Lu J, Song Z, Zhou Y, Liu T, Zhang D. From past to future: bibliometric analysis of global research productivity on nomogram (2000–2021). Front Public Health 2022;10:997713. https://doi.org/10.3389/fpubh.2022.997713.
[18] Kishore AK, Vail A, Bray BD, et al. Clinical risk scores for predicting stroke-associated pneumonia: a systematic review. Eur Stroke J 2016;1(2):76–84. https://doi.org/10.1177/2396987316651759.
[19] Liu L, Li Z, Zhou H, et al. Chinese Stroke Association guidelines for clinical management of ischaemic cerebrovascular diseases: executive summary and 2023 update. Stroke Vasc Neurol 2023;8(6):e3. https://doi.org/10.1136/svn-2023-002998
[20] Xie M, Liu Z, Dai F, Cao Z, Wang X. Predicting stroke-associated pneumonia in acute ischemic stroke: a machine learning model development and validation study with CBC-derived inflammatory indices. Int J Gen Med 2025;18:3117–28. https://doi.org/10.2147/IJGM.S524450.
[21] Li D, Liu Y, Jia Y, et al. Association between malnutrition and stroke-associated pneumonia in patients with ischemic stroke. BMC Neurol 2023;23(1):290. https://doi.org/10.1186/s12883-023-03340-1.
[22] Wu B, Luo H, Li J, et al. The relationship between the Barthel Index and stroke-associated pneumonia in elderly patients and factors of SAP. BMC Geriatr 2024;24(1):829. https://doi.org/10.1186/s12877-024-05400-8
[23] Salminen A. Activation of immunosuppressive network in the aging process. Ageing Res Rev 2020;57:100998. https://doi.org/10.1016/j.arr.2019.100998.
[24] Komiya K, Ishii H, Kadota J. Healthcare-associated pneumonia and aspiration pneumonia. Aging Dis 2014;6(1):27–37. https://doi.org/10.14336/AD.2014.0127.
[25] Torres-Riera S, Arboix A, Parra O, García-Eroles L, Sánchez-López MJ. Predictive clinical factors of in-hospital mortality in women aged 85 years or more with acute ischemic stroke. Cerebrovasc Dis 2025;54(1):11–9. https://doi.org/10.1159/000536436.
[26] Grief SN, Loza JK. Guidelines for the evaluation and treatment of pneumonia. Prim Care 2018;45(3):485–503. https://doi.org/10.1016/j.pop.2018.04.001.
[27] Yuan MZ, Li F, Tian X, et al. Risk factors for lung infection in stroke patients: a meta-analysis of observational studies. Expert Rev Anti Infect Ther 2015;13(10):1289–98. https://doi.org/10.1586/14787210.2015.1085302.
[28] Uhm JS, Kim J, Yu HT, et al. Stroke and systemic embolism in patients with atrial fibrillation and heart failure according to heart failure type. ESC Heart Fail 2021;8(2):1582–9. https://doi.org/10.1002/ehf2.13264.
[29] Arboix A, García-Eroles L, Massons J, Oliveres M. Predictive clinical factors of in-hospital mortality in 231 consecutive patients with cardioembolic cerebral infarction. Cerebrovasc Dis 1998;8(1):8–13. https://doi.org/10.1159/000015809.
[30] Tuz AA, Ghosh S, Karsch L, et al. Stroke and myocardial infarction induce neutrophil extracellular trap release disrupting lymphoid organ structure and immunoglobulin secretion. Nat Cardiovasc Res 2024;3(5):525–40. https://doi.org/10.1038/s44161-024-00462-8.
[31] Clark SR, Ma AC, Tavener SA, et al. Platelet TLR4 activates neutrophil extracellular traps to ensnare bacteria in septic blood. Nat Med 2007;13(4):463–9. https://doi.org/10.1038/nm1565.
[32] Urra X, Cervera A, Villamor N, Planas AM, Chamorro A. Harms and benefits of lymphocyte subpopulations in patients with acute stroke. Neuroscience 2009;158(3):1174–83. https://doi.org/10.1016/j.neuroscience.2008.06.014.
[33] Blanco M, Sobrino T, Montaner J, et al. Stroke with polyvascular atherothrombotic disease. Atherosclerosis 2010;208(2):587–92. https://doi.org/10.1016/j.atherosclerosis.2009.07.041.
[34] Liu D, Yan J, Luo T, Yang L. Association between C-reactive protein and hemoglobin in US rheumatoid arthritis patients based on NHANES data analysis. Sci Rep 2025;15(1):8905. https://doi.org/10.1038/s41598-025-93720-z.
[35] Patel UK, Kodumuri N, Dave M, et al. Stroke-associated pneumonia: a retrospective study of risk factors and outcomes. Neurologist 2020;25(3):39–48. https://doi.org/10.1097/NRL.0000000000000269.