frontiers | Frontiers in Physiology

TYPE Original Research
PUBLISHED 24 March 2025
DOI 10.3389/fphys.2025.1528910

OPEN ACCESS

EDITED BY
Feng Gao,
The Sixth Affiliated Hospital of Sun Yat-sen University, China

REVIEWED BY
Yang Yang,
First Affiliated Hospital of Zhengzhou University, China
Qingan Fu,
Nanchang University, China

*CORRESPONDENCE
Junkai Wen,
12023203@shutcm.edu.cn
Jiwei Cheng,
chengjiwei1@126.com

†These authors have contributed equally to this work

RECEIVED 15 November 2024
ACCEPTED 13 March 2025
PUBLISHED 24 March 2025

CITATION
Ding C, Yuan M, Cheng J and Wen J (2025) Cross-sectional study on smoking types and stroke risk: development of a predictive model for identifying stroke risk. *Front. Physiol.* 16:1528910. doi: 10.3389/fphys.2025.1528910

Cross-sectional study on smoking types and stroke risk: development of a predictive model for identifying stroke risk

Chao Ding[1†], Minjia Yuan[2†], Jiwei Cheng[1*] and Junkai Wen[1*]

[1]Putuo Hospital, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [2]Aviation Health Department, Spring Airlines Co.,Ltd, Shanghai, China

Background: Stroke, a major global health concern, is responsible for high mortality and long-term disabilities. With the aging population and increasing prevalence of risk factors, its incidence is on the rise. Existing risk assessment tools have limitations, and there is a pressing need for more accurate and personalized stroke risk prediction models. Smoking, a significant modifiable risk factor, has not been comprehensively examined in current models regarding different smoking types.

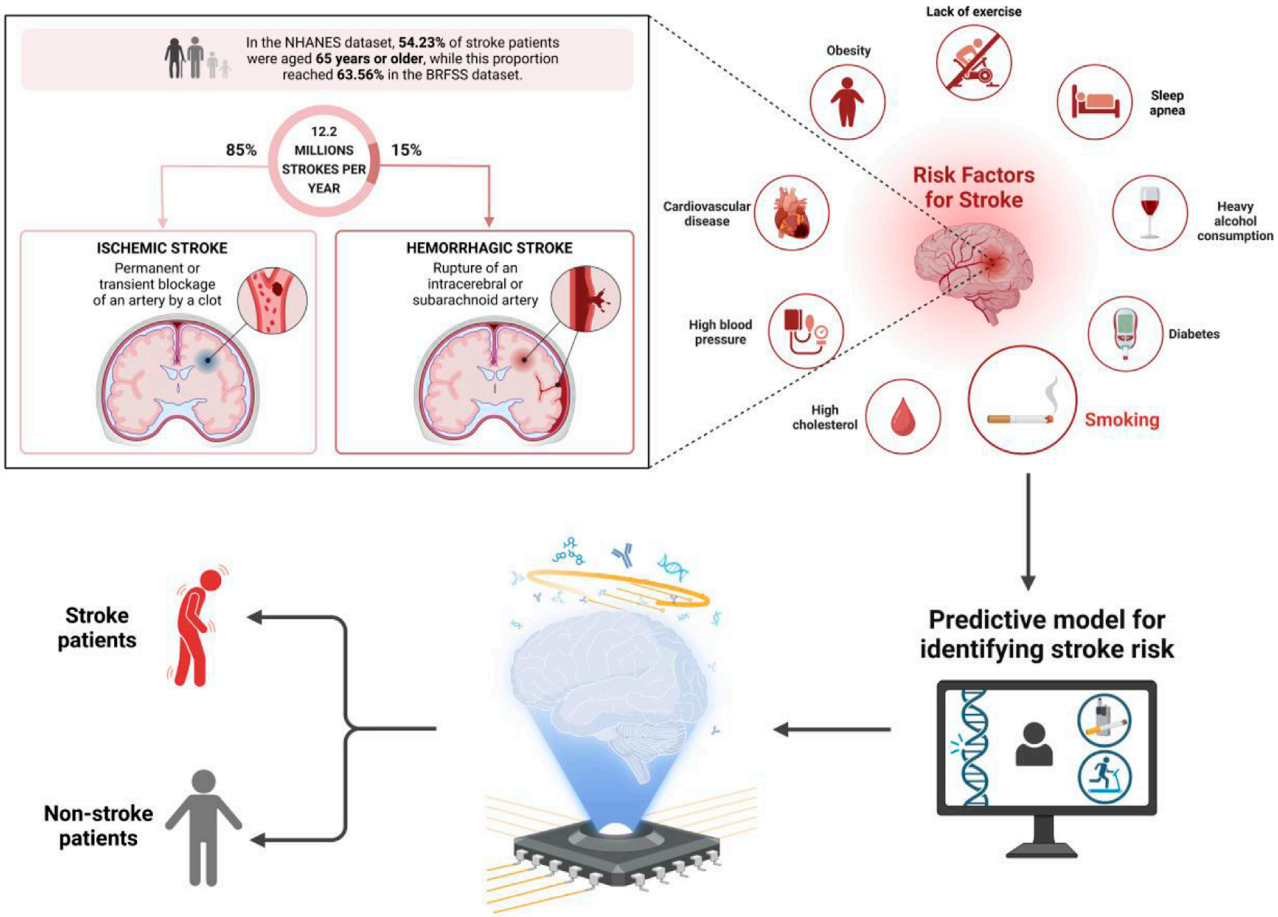
Methods: Data were sourced from the 2015–2018 National Health and Nutrition Examination Survey (NHANES) and the 2020–2021 Behavioral Risk Factor Surveillance System (BRFSS). Tobacco use (including combustible cigarettes and e-cigarettes) and stroke history were obtained through questionnaires. Participants were divided into four subgroups: non-smokers, exclusive combustible cigarette users, exclusive e-cigarette users, and dual users. Covariates such as age, sex, race, education, and health conditions were also collected. Multivariate logistic regression was used to analyze the relationship between smoking and stroke. Four machine-learning models (XGBoost, logistic regression, Random Forest, and Gaussian Naive Bayes) were evaluated using the area under the receiver-operating characteristic curve (AUC), and Shapley's additive interpretation method was applied for feature importance ranking and model interpretation.

Results: A total of 273,028 individuals were included in the study. Exclusive combustible cigarette users had an elevated stroke risk (β: 1.36, 95% CI: 1.26–1.47, $P < 0.0001$). Among the four machine-learning models, the XGBoost model showed the best discriminative ability with an AUC of 0.794 (95% CI = 0.787–0.802).

Conclusion: This study reveals a significant association between smoking types and stroke risk. An XGBoost-based stroke prediction model was established, which has the potential to improve the accuracy of stroke risk assessment and contribute to personalized interventions for stroke prevention, thus alleviating the healthcare burden related to stroke.

KEYWORDS

stroke, machine learning, prediction model, Shap, XGBoost



GRAPHICAL ABSTRACT

1 Introduction

Stroke represents a critical global health challenge, defined by the World Health Organization as the sudden onset of clinical symptoms indicative of focal or global cerebral dysfunction, typically lasting more than 24 h or leading to death, with no apparent cause other than vascular origin (Hilkens et al., 2024). This condition manifests primarily in two forms: ischemic stroke, accounting for approximately 85% of cases, and hemorrhagic stroke, comprising the remaining 15% (Maida et al., 2024; Tu et al., 2023a).

.5The pathogenesis of ischemic stroke involves the obstruction of cerebral blood flow, predominantly due to atherosclerosis or embolism. In contrast, hemorrhagic stroke results from the rupture of intracranial or peri-cerebral blood vessels, causing hemorrhage and subsequent compression of brain tissue. As a leading cause of mortality and long-term disability worldwide, stroke accounted for over 12.2 million incident cases and 6.55 million deaths in 2020 (Tu et al., 2023a; Ananth et al., 2023; Tu et al., 2023b). The global burden of stroke is projected to escalate further, driven by demographic aging and the increasing prevalence of modifiable risk factors such as hypertension and obesity (Tsao et al., 2023; Martin et al., 2024). This epidemiological trend highlights the critical need for enhanced predictive capabilities to identify high-risk populations and implement preventive measures effectively.

The consequences of stroke extend beyond acute mortality, often resulting in persistent neurological deficits that impose substantial personal and societal burdens, including extensive requirements for long-term care and rehabilitation (Feigin et al., 2023). While multiple risk factors contribute to stroke susceptibility, they can be broadly categorized into modifiable and non-modifiable determinants. Modifiable factors include metabolic disorders (diabetes, dyslipidemia), lifestyle behaviors (smoking, physical inactivity), and dietary patterns, whereas non-modifiable factors encompass genetic predisposition, age, and sex (Tu et al., 2023a; Ekker et al., 2023). Among these, smoking emerges as a particularly potent risk factor, with its impact on stroke risk varying according to smoking type, intensity, and duration

Abbreviations: ML, machine learning; NHANES, National Health and Nutrition Examination Survey; BRFSS, Behavioral Risk Factor Surveillance System; ORs, electronic cigarettes, e-cigarettes; odds ratios; CIs, confidence intervals; RF, random forest; AUC, area under the subject-operating characteristic curve; AI, artificial intelligence; DL, deep learning; LASSO, Least Absolute Shrinkage and Selection Operator; BMI, body mass index; LightGBM, Light Gradient Boosting Machine; DT, Decision Tree; KNN, K-Nearest Neighbors; CatBoost, Categorical Boosting; SVM, Support Vector Machine; MLP, Multi-Layer Perceptron.

(Martin et al., 2024; Wang Y. et al., 2024). Despite the identification of these risk factors, conventional risk assessment tools demonstrate limited predictive accuracy and generalizability across diverse populations.

The advent of artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL) technologies, has revolutionized stroke risk prediction through their capacity to discern complex patterns within extensive datasets (Singh et al., 2024). These advanced computational approaches have demonstrated superior performance in risk stratification compared to traditional methods, enabling more precise and individualized risk assessments. The integration of AI with emerging data sources, including electronic health records and wearable devices, facilitates continuous monitoring and early detection of stroke risk factors (Papadopoulou et al., 2024; Liu et al., 2024; Huang et al., 2023). This technological synergy offers promising solutions to address the growing stroke burden through timely intervention and personalized prevention strategies.

This study aims to develop an advanced predictive model that specifically addresses the identification of individuals at elevated stroke risk, with particular emphasis on the differential impacts of various smoking behaviors. By leveraging state-of-the-art machine learning algorithms, the research seeks to enhance existing risk assessment frameworks through a more comprehensive evaluation of stroke risk determinants. The inclusion of smoking typology as a critical predictive variable represents a novel contribution to stroke prediction models, as current approaches have not sufficiently addressed the heterogeneous effects of different smoking patterns on cerebrovascular health.

The anticipated outcomes of this research are twofold: first, to provide novel insights into the relationship between smoking behaviors and stroke risk; second, to demonstrate the transformative potential of machine learning applications in clinical risk assessment. The proposed model is expected to significantly improve the accuracy of stroke risk prediction while simultaneously supporting the development of targeted prevention strategies. Ultimately, this research aims to contribute to the reduction of stroke incidence and the mitigation of its associated healthcare and socioeconomic burdens through advanced predictive analytics and personalized intervention approaches.

2 Materials and methods

2.1 Source of data

We utilized data from two major surveys: the Behavioral Risk Factor Surveillance System (BRFSS) for 2020–2021 and the National Health and Nutrition Examination Survey (NHANES).

The BRFSS, a comprehensive and nationally representative telephone survey, is jointly administered by the CDC and all U.S. states, along with participating territories. It focuses on gathering data related to behavioral risk factors.

NHANES, on the other hand, is a periodic cross-sectional survey in the US. Since the early 1960 s, the National Center for Health Statistics and the CDC have conducted it. Starting from 1999, it has been a biennial program, interviewing over 5,000 individuals per iteration. Using a complex multi - stage probability sampling method, NHANES generates nationally representative statistics for the civilian (non-institutionalized) household population. It collects a wide range of health and nutrition data, covering demographics, diet, examinations, laboratory results, and questionnaire responses. The data collection for NHANES was approved by the National Center for Health Statistics Research Ethics Review Board, with all participants' parents or guardians providing written informed consent. Our report adheres to the Strengthening the Reporting of Observational Research in Epidemiology guidelines for presenting cross-sectional studies.

2.2 Study population

NHANES is a cross-sectional study designed to collect data on the health and nutritional status of the U.S. population. Data is obtained through structured home interviews, physical assessments at mobile screening sites, and laboratory analyses utilizing a multistage probability sampling technique. Initially, 19,225 individuals were identified from the NHANES 2015–2018 dataset. Participants without data on tobacco use (n = 7,378) were eliminated from the study. Additionally, individuals lacking stroke status information (n = 574) were eliminated. A total of 11,273 participants were included in the final analysis. All data employed in this study are publicly available (https://www.cdc.gov/nchs/nhanes/) and have been adjusted for demographic factors for additional analysis. Furthermore, we included 401,958 participants from the 2020–2021 BRFSS. After excluding absent smoking-related data (n = 139,521) and stroke-related data (n = 682), a total of 261,755 participants were finally included (Figure 1).

2.3 Tobacco use assessment

The NHANES database, specifically the Smoking Cigarette Use dataset (SMQ), gathered data on cigarette intake, current use, 30-day smoking prevalence, quantity, and other smoking-related details. Participants were asked via SMQ020 if they had smoked at least 100 cigarettes in their lives and about current habits (SMQ040), and via SMQ900 if they had ever used an e-cigarette. Smoking combustible cigarettes was defined as having smoked at least 100 cigarettes in a lifetime or currently smoking daily/occasionally, while e-cigarette use was defined as any single-time use (Okafor et al., 2022).

In the 2020 BRFSS, smoking inquiries paralleled NHANES, collecting data on cigarette and e-cigarette use. BRFSS used SAS variables SMOKE100 (Column 202) to ask about lifetime smoking of at least 100 cigarettes, SMOKDAY2 (Column 203) for current smoking behavior, and ECIGARET (Column 310) for e-cigarette use. Based on these, participants were divided into four subgroups: non-smokers, exclusive combustible cigarette smokers, exclusive e-cigarette users, and dual users.

2.4 Stroke diagnosis

In NHANES, stroke diagnosis relied on a self-reported questionnaire (MCQ160f). Participants were asked "Has a physician or other healthcare provider ever informed you about a stroke?" with

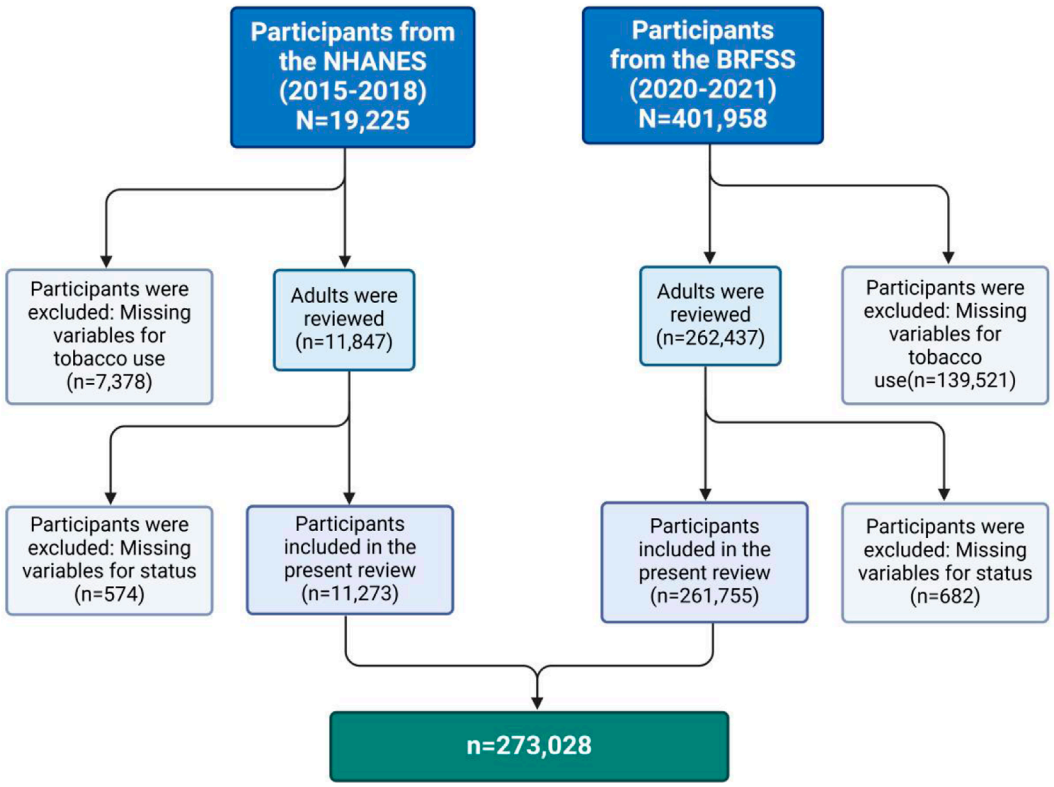


FIGURE 1
Flowchart for inclusion of study populations according to the purpose of the study.

"yes" or "no" responses. A "yes" answer indicated a confirmed stroke diagnosis.

In BRFSS, stroke was diagnosed through a self-reported questionnaire (SAS variable: CVDSTRK3, Column 117). The question "Have you ever been informed that you had a stroke?" was part of the Chronic Health Conditions segment. A positive response meant the participant had experienced a stroke.

2.5 Covariates

To ensure consistency between NHANES and BRFSS, several covariates were considered. Age was grouped as 18–65 and 65+; sex as female or male; race as Hispanic, non-Hispanic white, non-Hispanic black, or other; education as high-school graduated or not; marital status in multiple categories; income in three brackets; exercise as yes or no. Additionally, BMI, weight, and health conditions like diabetes and total heart disease were included.

Exercise data came from different questions in BRFSS (EXERANY2: "In the last 30 days, have you engaged in") and NHANES (PAQ746: "How frequently do you attend."). A positive response for exercise was marked as "YES", negative as "NO". Diabetes was determined by the self-reported DIABETE4 question in both surveys, and total heart disease was based on self-reported SP data (CVDINFR4, CVDCRHD4) asking about heart attack, angina, or coronary artery disease.

2.6 Statistical analysis

Statistical analyses were conducted using R software (version 4.1.6). Owing to the complex sampling designs utilized by the NHANES and BRFSS surveys, we integrated sample weights from many study eras in our analytical approaches to accurately estimate health-related data. Through multivariate logistic regression analysis, we derived β values and 95% confidence intervals for the association between types of tobacco use and stroke incidence. The key reason for choosing multivariate logistic regression is its interpretability. Logistic regression produces clear and interpretable coefficients that represent the chances ratios associated with each predictor variable. This transparency is particularly vital in clinical and epidemiological studies, where understanding the relationship between variables and outcomes is imperative. Model 2 included adjustments for gender, age, and racial demographics, whereas Model 1 was uncorrected. A third model (Model 3) was constructed with extensive covariate adjustment for the input variables of income, marital status, education, exercise status, weight, body mass index, diabetes, and heart disease.

2.7 Predictive modeling and assessment

2.7.1 Deterministic feature selection

Employing the Least Absolute Shrinkage and Selection Operator (LASSO) regression model, we discerned the primary

predictors of stroke while accounting for the covariation among the covariates. LASSO enhances the predictive accuracy and interpretability of statistical models by integrating a penalty into the regression equation, proportional to the sum of the absolute values of the coefficients. This software effectively eliminates variables with zero coefficients using the five-fold cross-validation method with LassoCV (Python version: sklearn 0.22.1).

2.7.2 ML modeling and development

We evaluated ten widely-used machine learning methods to construct and compare predictive models: XGBoost, Logistic Regression, Random Forest (RF), Gaussian Naive Bayes, LightGBM, Decision Tree (DT), k-Nearest Neighbors (KNN), CatBoost, Support Vector Machine (SVM), and Multilayer Perceptron (MLP). All models were executed using Python 3.7 and R 4.4.2. XGBoost, LightGBM, and CatBoost utilized the 'xgboost 1.2.1′, 'lightgbm 3.3.2′, and 'catboost 1.0.6′ packages, respectively. Other models, including Logistic Regression, RF, Gaussian Naive Bayes, DT, KNN, SVM, and MLP, were implemented through the 'scikit-learn 0.22.1′ package. Patients were randomly assigned to training and testing groups in an 80:20 ratio to ensure comprehensive model evaluation. For all models, hyperparameter optimization was performed via grid search combined with five-fold cross-validation to ensure fair comparison and mitigate overfitting.

2.7.3 Model optimization and evaluation

Five-fold cross-validation was utilized to evaluate the model's predictive performance and confirm its stability. To account for potential class imbalance in stroke prediction, stratified cross-validation was applied during both hyperparameter tuning and evaluation phases. The training dataset was randomly divided into five groups. During each iteration of the five-fold cross-validation, four subsets were randomly assigned as the training set, while the remaining subset functioned as the validation set. In each training phase of the model, 20% of the dataset was randomly selected from the training set to assess the model's performance. The importance of features was assessed using the Shapley additive explanation (Shap). Features with greater absolute Shap values substantially impacted the model prediction scores. Additionally, the distribution of feature values and their relationship with model predictions were evaluated. Model performance was evaluated using seven standard metrics: Area Under the ROC Curve (AUC), Accuracy, Sensitivity (Recall), Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), and F1-score. The validation dataset was utilized to evaluate and compare the effectiveness of each model. The models' ability to predict stroke was assessed using the area under the receiver operating characteristic curve for the individuals.

2.7.4 Web deployment tool based on the streamlit framework

The final prediction model is included into a web application using the Streamlit Python framework to facilitate its use in a clinical setting. Upon obtaining the values of the pertinent attributes in the final model, the program can deliver the likelihood of stroke together with a graph illustrating the variations of the particular sub-items.

3 Results

3.1 Participants' characteristics

Our combined analysis integrated data from NHANES and BRFSS, with 11,273 and 261,755 participants respectively. Key demographic and health-related differences between non-stroke and stroke groups are presented in Table 1, 2.

Stroke patients were significantly older in both datasets. In NHANES, 54.23% of stroke participants were 65 or older, and in BRFSS, this percentage was 63.56% ($P < 0.0001$). Although females were more common in both stroke groups (54.11% in NHANES and 54.80% in BRFSS), the difference was not statistically significant. Income disparities were evident, with a higher proportion of lower-income stroke patients in BRFSS (43.96% vs. 34.41% in NHANES, $P < 0.0001$). Widowed individuals were more prevalent among stroke patients (19.04% in NHANES, 24.22% in BRFSS, $P < 0.0001$). Non-Hispanic Black participants were more common in stroke groups compared to non - stroke groups (16.92% in NHANES and 10.16% in BRFSS, $P < 0.0001$). Non-stroke groups had higher educational attainment, with over 85% of stroke patients having at least a high-school education in both datasets ($P < 0.0001$).

Regarding health and lifestyle, stroke patients had a higher BMI (30.96 ± 6.83 in NHANES and 28.95 ± 6.65 in BRFSS, $P < 0.0001$). Diabetes and heart disease were more common in stroke patients (33.71% and 35.58% in NHANES, 31.08% and 36.45% in BRFSS, $P < 0.001$). Stroke patients exercised less (58.86% in NHANES, 59.28% in BRFSS, $P < 0.001$) and were more likely to be current combustible cigarette smokers (42.14% in NHANES, 10.32% in BRFSS, $P < 0.0001$).

3.2 The association between cigarette use and stroke

Figure 2 demonstrates the association between tobacco use patterns and stroke risk in individuals aged 20 and older, analyzed using three statistical models. Model 1 is unadjusted, Model 2 adjusts for demographic factors, and Model 3 fully adjusts for confounders such as age, sex, race, and comorbidities.

In Model 3, exclusive use of combustible cigarettes (β: 1.34–1.36, 95% CI: 1.26–1.47, $P < 0.0001$) and dual use of combustible cigarettes with e-cigarettes (β: 1.34, 95% CI: 1.23–1.46, $P < 0.0001$) were strongly associated with increased stroke risk. These results indicate that combustible cigarette use, whether alone or combined with e-cigarettes, significantly elevates stroke risk after adjusting for confounders.

In contrast, exclusive e-cigarette use showed no significant association with stroke risk in Model 3 ($P > 0.05$), suggesting a distinct risk profile compared to combustible cigarettes. However, in Model 2, all smoking modalities, including e-cigarette use, were significantly associated with stroke risk, highlighting the influence of demographic factors on these associations.

TABLE 1 Attributes of the NHANES research cohort.

Characteristic	Non-stroker (n = 10,791)	stroker (n = 482)	P-value
Characteristic			
Age (%)			<0.0001
18 ≤ age < 65	81.54	45.77	
≥65	18.46	54.23	
Sex (%)			0.4109
Male	48.16	45.89	
Female	51.84	54.11	
Income (%)			<0.0001
less than $25,000	15.25	34.41	
$25,000 to less than $50,000	25.23	28.68	
$50,000 or more	59.52	36.84	
Marital Status (%)			<0.0001
Married	53.85	49.97	
Widowed	5.47	19.04	
Divorced	9.76	14.56	
Separated	2.61	2.16	
Never married	18.73	8.40	
living with partner	9.58	5.87	
Race (%)			<0.0001
Hispanic	15.73	8.64	
Non-Hispanic White	63.01	63.83	
Non-Hispanic Black	11.24	16.92	
Other Races	10.03	10.61	
Education level (%)			<0.0001
<High school	8.53%	12.89%	
≥High school	91.47%	87.11%	
Alcohol use (%)			0.0518
Yes	63.97	56.95	
No	36.03	43.05	
BMI	29.57 ± 7.13	30.96 ± 6.83	<0.0001
Weight	83.80 ± 22.38	84.45 ± 22.14	0.6125
Diabetes (%)			<0.0001

(Continued on the following page)

TABLE 1 (*Continued*) Attributes of the NHANES research cohort.

Characteristic	Non-stroker (n = 10,791)	stroker (n = 482)	P-value
Yes	10.58	33.71	
No	89.42	66.29	
Total heart disease (%)			<0.001
Yes	5.86	35.58	
No	94.14	64.42	
Exercise (%)			<0.001
Yes	75.96%	58.86%	
No	24.04%	41.14%	
Type of smoking			<0.0001
Non-smokers	52.92	41.06	
Only combustible cigarettes	25.25	42.14	
Only e-cigarettes	4.67	1.79	
Both combustible cigarettes and e-cigarettes	17.16	15.01	

Mean + SD, for continuous variables: the P value was calculated by the weighted linear regression model; (%) for categorical variables: the P value was calculated by the weighted chi-square test.
Abbreviations: BMI, body mass index.

3.3 A prediction model to evaluate the stroke risk

The selection of variables included in the LASSO regression model was based on a combination of clinical relevance, prior literature, and statistical considerations. Specifically, we conducted an extensive literature review on known risk factors for stroke and included variables that have been widely recognized as relevant predictors (e.g., age, BMI, smoking status, hypertension, diabetes, and other cardiovascular risk factors) (Fu et al., 2024; Dutta et al., 2024; Notley et al., 2025; Hjort et al., 2024; Jumbe et al., 2025). Additionally, we ensured that all selected variables were available in both datasets (NHANES and BRFSS) to maintain consistency across analyses. LASSO regression was employed due to its ability to perform automatic variable selection by shrinking less important coefficients to zero, thereby reducing model complexity and improving interpretability. By applying 5-fold cross-validation, we optimized the regularization parameter (λ) to achieve a balance between model performance and feature sparsity.

We used baseline characteristics for stroke risk prediction. Through the LASSO method for feature selection, six key predictors were identified: 'age', 'income', 'exercise', 'alcohol consumption', 'diabetes', and 'total heart disease'. These were determined to be the most valuable for building the predictive model using LASSO regularization and five-fold cross-validation. Figure 3 shows the coefficients for each variable in the LASSO model.

3.4 Model explanation

We compared XGBoost, logistic regression, RF, and Gaussian Naive Bayes (Figure 4). While the initial visualization focuses on four representative models (linear vs nonlinear, parametric vs non-parametric), Supplementary Figure S1 in the Supplement provides complete ROC comparisons across all ten evaluated models. The results show that XGBoost has the best predictive ability (AUC = 0.794) among the primary candidates, though LightGBM achieved comparable performance (Validation AUC = 0.793) as detailed in Supplementary Table S1. Notably, simpler models like Logistic Regression (AUC = 0.785) demonstrated adequate discrimination, whereas KNN showed significant overfitting with a 25.3% AUC drop from training (0.930) to validation (0.742), likely due to local noise sensitivity in high-dimensional space. Following this, we utilized XGBoost to build a clinical prediction model given its optimal balance between performance and computational efficiency.

We compared the difference in AUC using Delong's nonparametric method with MedCalc version 19.6 (https://www.medcalc.org), progressively reducing the features of the selected ML model until a significant decrease in AUC was observed. We illustrated the predictive influence of the variables on the outcomes using SHAP plots. The impact of variables on outcomes can be visually assessed through the amplitude of SHAP values (shown by color variations) and the trend along the horizontal axis (likelihood of an adverse event). In the image on the right, older individuals (shown in red) have a higher likelihood of bad prognosis compared

TABLE 2 Attributes of the BRFSS study cohort.

Characteristic	Non-stroker (n = 250,890)	stroker (n = 10,865)	P-value
Characteristic			
Age (%)			<0.0001
18 ≤ age < 65	64.64%	36.44%	
≥65	35.36%	63.56%	
Sex (%)			0.4109
Male	45.14%	45.20%	
Female	54.86%	54.80%	
Income (%)			<0.0001
less than $25,000	22.41%	43.96%	
$25,000 to less than $50,000	23.77%	27.04%	
$50,000 or more	53.82%	29.00%	
Marital Status (%)			<0.0001
Married	53.13%	41.55%	
Widowed	11.09%	24.22%	
Divorced	12.94%	20.08%	
Separated	1.88%	2.85%	
Never married	17.37%	9.73%	
living with partner	53.13%	41.55%	
Race (%)			<0.0001
Hispanic	7.40%	4.01%	
Non-Hispanic White	77.44%	77.68%	
Non-Hispanic Black	7.37%	10.16%	
Other Races	7.80%	8.15%	
Education level (%)			<0.0001
<High school	6.31%	12.22%	
≥High school	93.69%	87.78%	
Alcohol use (%)			<0.0001
Yes	50.64	33.07	
No	49.36	66.93	
BMI	28.40 ± 6.40	28.95 ± 6.65	<0.0001
Weight	82.56 ± 21.07	82.93 ± 21.41	0.039
Diabetes (%)			<0.0001

(Continued on the following page)

TABLE 2 (*Continued*) Attributes of the BRFSS study cohort.

Characteristic	Non-stroker (n = 250,890)	stroker (n = 10,865)	P-value
Yes	12.75	31.08	
No	87.35	68.92	
Total heart disease (%)			<0.001
Yes	7.83	36.45	
No	92.17	63.55	
Exercise (%)			<0.001
Yes	76.70%	59.28%	
No	23.30%	40.72%	
Type of smoking			<0.0001
Non-smokers	74.52%	73.66%	
Only combustible cigarettes	6.06%	10.32%	
Only e-cigarettes	11.56%	6.47%	
Both combustible cigarettes and e-cigarettes	7.86%	9.55%	

Mean +SD, for continuous variables: the P value was calculated by the weighted linear regression model; (%) for categorical variables: the P value was calculated by the weighted chi-square test.
Abbreviations: BMI, body mass index.

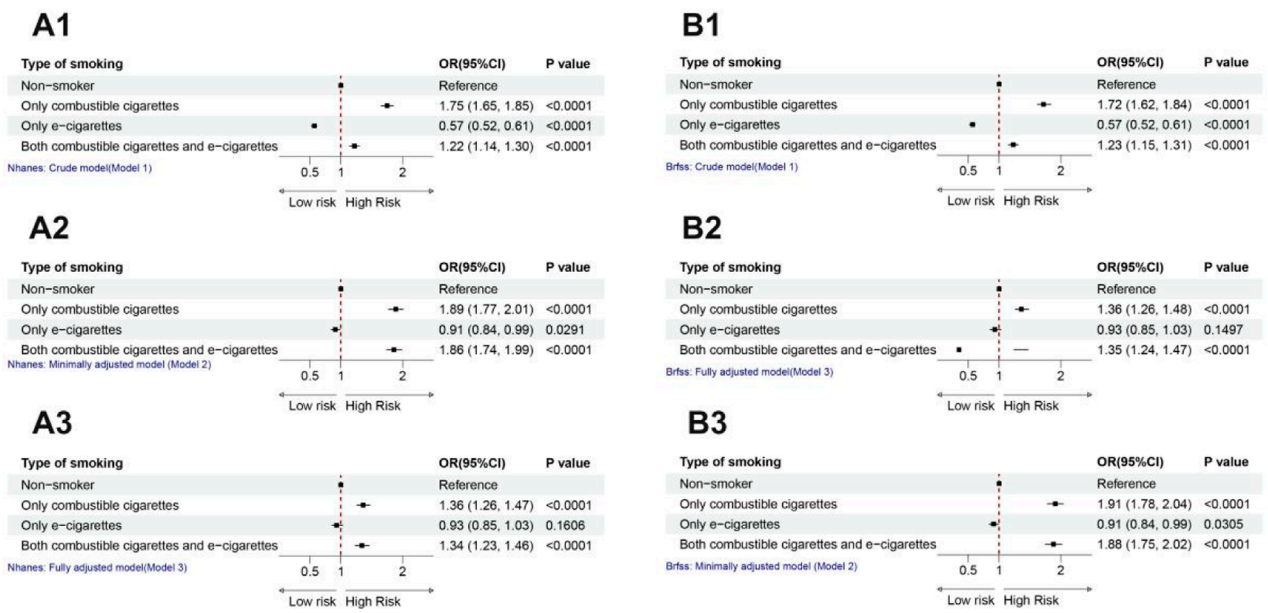


FIGURE 2
Forest plot of the relationship between cigarette use and stroke.

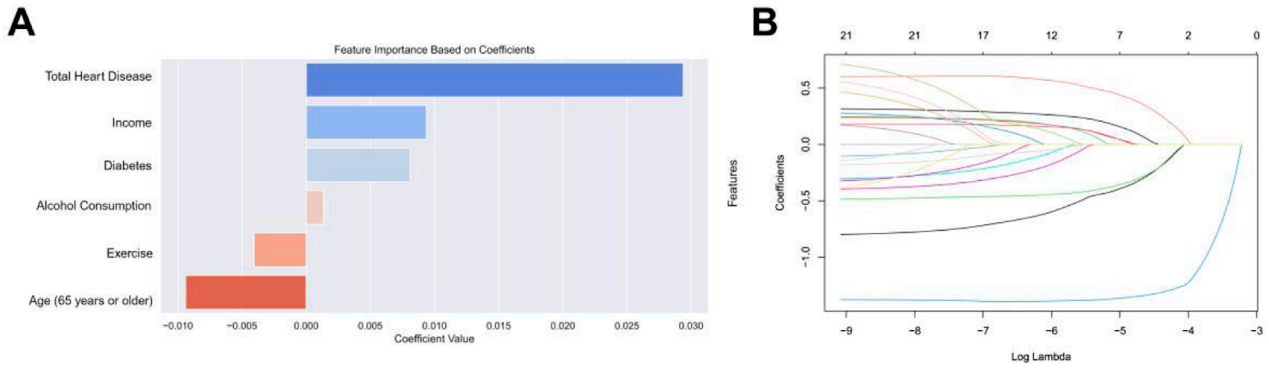


FIGURE 3
(A) Selection of features with non-zero coefficients and their coefficients using the LASSO regression method. **(B)** The impact of the penalty coefficient λ on the weight coefficients of each independent variable is represented on the horizontal axis as λ and on the vertical axis as the weight coefficients, with distinct colors indicating the weight coefficients of individual independent variables.

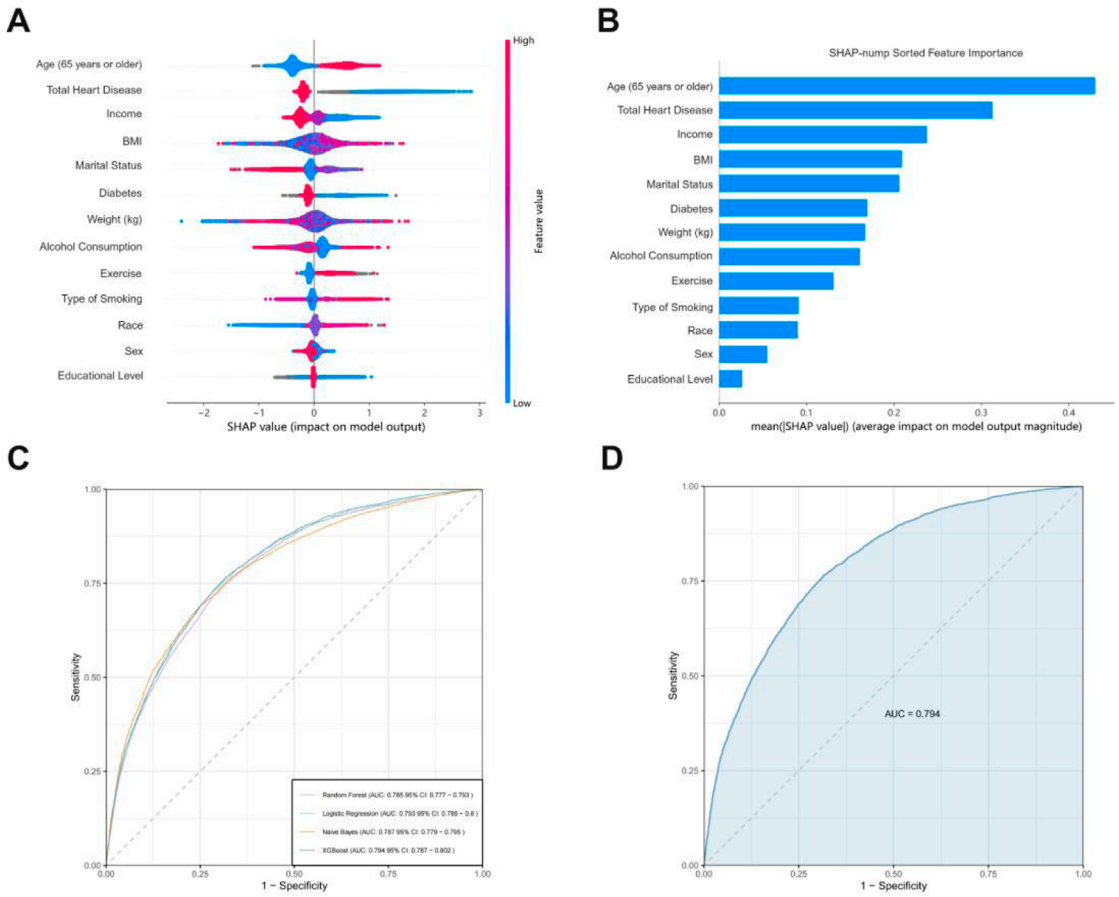


FIGURE 4
(A) The horizontal coordinates indicate the magnitude of the SHAP value, with positive values representing the positive contribution of the variable to a positive stroke outcome and negative values the opposite; the color ranges from blue to red to characterize the low to high values of the variable in order. **(B)** SHAP evaluations of the XGBoost algorithm for forecasting adverse outcomes in stroke patients. **(C)** The mean AUC performance of four machine learning models evaluated using five-fold external cross-validation. **(D)** ROC curve examination of the XGBoost algorithms for predicting stroke risk in the external test set.

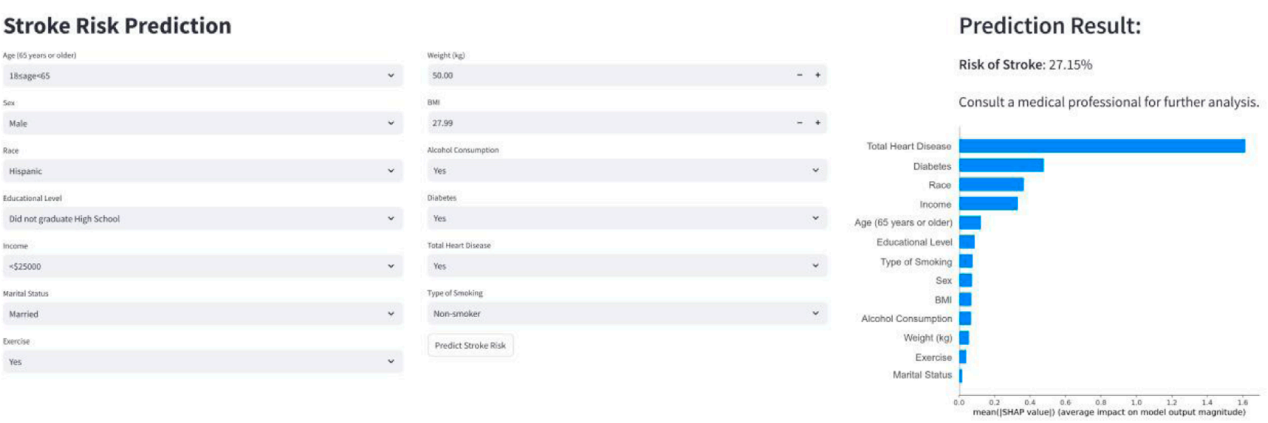


FIGURE 5
Web-based calculator predicts stroke risk.

to younger individuals (represented in blue). Individuals with a high BMI (red) are prone to experience a poor outcome as stroke patients (right panel).

3.5 Convenient application for clinical utility

Figure 5 illustrates the integration of the final prediction model into a web application to improve its practical utility in a clinical environment. In addition to risk prediction, SHAP summary plots were generated to illustrate the contribution of each feature to the decision-making process. Features highlighted in red indicate higher risk, while features highlighted in blue indicate lower risk. The web application can be accessed at the URL: http://10.50.1.58:8502.

4 Discussion

This study demonstrates a strong association between tobacco consumption and the risk of stroke. Epidemiological studies consistently demonstrate a substantial association between smoking behaviors and increased stroke incidence, with smokers experiencing a much higher risk than non-smokers (Wang X. et al., 2024; Feigin et al., 2024; Shi et al., 2023; Elser et al., 2023). Our machine learning framework extended these observations by identifying nonlinear risk thresholds. Meanwhile, this risk extends to both active smokers and those exposed to secondhand smoke (Lin et al., 2016; Lu et al., 2024), emphasizing the need for targeted public health interventions to reduce smoking rates. The intricate relationship between smoking and stroke highlights the importance of addressing tobacco use as a critical component of stroke prevention strategies.

The molecular mechanisms through which smoking contributes to stroke risk involve a variety of harmful constituents present in tobacco smoke, many of which are toxic and carcinogenic (Adler et al., 2023; Jang et al., 2024; Liang et al., 2022). Key components such as carbon monoxide and nicotine induce endothelial dysfunction and promote a pro-inflammatory state (Rezk-Hanna et al., 2019; Wang et al., 2023; Belkin et al., 2023), leading to vascular damage (Whitehead et al., 2021). Additionally, oxidative stress resulting from the inhalation of free radicals damages vascular endothelium, exacerbating atherosclerotic processes and increasing thrombogenic potential (Qin et al., 2020; Higashi, 2023). Notably, the accumulation of these detrimental substances impairs the healing of vascular injuries, further predisposing individuals to the development of cerebrovascular diseases (Siegel et al., 2022). These mechanisms align with our findings of a significantly increased stroke risk in traditional cigarette users. E-cigarettes became commercially available in the US in 2007 and have since gained widespread popularity among both adults and adolescents as they are generally considered to produce fewer toxins than traditional cigarettes (Sood et al., 2018; Goniewicz et al., 2017). However, emerging evidence suggests that they may exert cerebrovascular toxicity via nicotine-induced oxidative stress and inflammatory cytokine release (Crotty Alexander et al., 2018; Benowitz and Fraiman, 2017),and adversely affect the circulatory system through mechanisms such as increased heart rate and blood pressure, endothelial dysfunction, and accelerated platelet aggregation (Qasim et al., 2017; Goniewicz et al., 2014). Preclinical evidence shows that chronic e-cigarette exposure compromises blood-brain barrier (BBB) integrity and exacerbates ischemic injury, comparable to traditional smoking (Kaisar et al., 2017).

To enhance clinical utility, we developed a predictive model utilizing baseline characteristics as potential predictors of stroke risk. LASSO identified six linearly associated predictors, whereas SHAP revealed BMI as a critical nonlinear factor indicating that obesity elevated stroke risk through different mechanisms. While age consistently emerged as the strongest predictor reflecting cumulative vascular damage, BMI's absence in LASSO contrasted with its SHAP prominence, highlighting threshold effects captured by XGBoost. This divergence underscores the complementary strengths of LASSO and SHAP, advocating for their joint use in heterogeneous risk profiling.

The strengths of our research are rooted in its comprehensive approach, utilizing multiple machine learning algorithms to determine the most effective model for stroke prediction. Notably, the severe class imbalance (4.16% stroke cases) was addressed through XGBoost's scale_pos_weight parameter and weighted logistic regression, which improved model sensitivity while maintaining specificity. Additionally, we developed an online predictive platform, enhancing accessibility for clinicians. However, conventional accuracy metrics may overestimate clinical utility given this imbalance, as evidenced by the discrepancy between training set performance (XGBoost F1 = 0.808) and validation metrics (F1 = 0.837). Subsequent research should pursue multi-center validations incorporating advanced sampling techniques and composite metrics like AUC-PR to better handle skewed distributions.

However, we are also acutely aware of several limitations in our study. First, the extreme class imbalance (95.84% non-stroke cases) and potential biases from self-reported data may affect model accuracy and generalizability, particularly for low-prevalence predictors like e-cigarette use. Second, the cross-sectional design precludes causal inference and cannot account for reverse causality. Third, our analysis revealed no significant association between exclusive e-cigarette use and stroke risk in fully adjusted models, which may reflect methodological limitations such as insufficient capture of long-term cumulative exposure, underreporting of dual use with combustible products, or lack of granularity on e-cigarette consumption patterns. Additionally, the biological latency of cerebrovascular damage from e-cigarettes may exceed our observational timeframe, suggesting potential long-term harm that warrants further investigation using available long-term clinical and animal data. These findings underscore the need for comprehensive tobacco cessation strategies prioritizing complete nicotine abstinence over e-cigarette use alone.

In terms of future research directions, multi-center validations across diverse countries and ethnic groups are essential to confirm the generalizability of our model and findings. Large-scale prospective studies incorporating repeated exposure measurements are necessary to establish temporal precedence between e-cigarette use and stroke outcomes, thereby clarifying causal relationships. Long-term follow-up of participants will provide more precise insights into the temporal progression of stroke development.

In summary, our predictive model represents a significant advancement in stroke risk assessment, providing healthcare professionals with a robust tool for informed decision-making. The potential for widespread application in clinical settings holds promise for improving stroke prevention strategies, ultimately leading to better health outcomes for individuals at risk. We look forward to the continued exploration of this model's utility across diverse clinical scenarios and its role in enhancing patient care.

5 Conclusion

This study establishes that distinct tobacco consumption patterns differentially elevate stroke risk, with combustible tobacco products demonstrating the strongest association. Through machine learning-driven feature selection, we developed and validated a clinical prediction tool achieving in stroke risk stratification. The integration of this model into an open-access web platform enables real-time, individualized stroke risk assessment, offering clinicians a practical tool for targeted intervention strategies. These findings underscore the importance of smoking cessation in stroke prevention and provide a scalable solution for risk stratification in high-risk populations. Future studies should focus on multi-center validation and further exploration of dose-dependent effects of tobacco use to enhance the generalizability and precision of the model.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participant's legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

CD: Conceptualization, Data curation, Formal Analysis, Visualization, Writing–original draft, Writing–review and editing. MY: Conceptualization, Data curation, Formal Analysis, Visualization, Writing–review and editing. JC: Conceptualization, Funding acquisition, Supervision, Writing–review and editing. JW: Conceptualization, Data curation, Formal Analysis, Visualization, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The work was funded by Shanghai Putuo District Health System Clinical Characteristic Special Disease Construction Project (No. 2023tszb04).

Acknowledgments

Participants in the NHANES and BRFSS databases.

Conflict of interest

Author MY was employed by Spring Airlines Co,.Ltd.

The remaining authors declare that the research 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 author(s) declare that no Generative AI was used in the creation of this manuscript.

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/fphys.2025.1528910/full#supplementary-material

References

Adler, N., Bahcheli, A. T., Cheng, K. C. L., Al-Zahrani, K. N., Slobodyanyuk, M., Pellegrina, D., et al. (2023). Mutational processes of tobacco smoking and APOBEC activity generate protein-truncating mutations in cancer genomes. *Sci. Adv.* 9 (44), eadh3083. doi:10.1126/sciadv.adh3083

Ananth, C. V., Brandt, J. S., Keyes, K. M., Graham, H. L., Kostis, J. B., and Kostis, W. J. (2023). Epidemiology and trends in stroke mortality in the USA, 1975-2019. *Int. J. Epidemiol.* 52 (3), 858–866. doi:10.1093/ije/dyac210

Belkin, S., Benthien, J., Axt, P. N., Mohr, T., Mortensen, K., Weckmann, M., et al. (2023). Impact of heated tobacco products, E-cigarettes, and cigarettes on inflammation and endothelial dysfunction. *Int. J. Mol. Sci.* 24 (11), 9432. doi:10.3390/ijms24119432

Benowitz, N. L., and Fraiman, J. B. (2017). Cardiovascular effects of electronic cigarettes. *Nat. Rev. Cardiol.* 14 (8), 447–456. doi:10.1038/nrcardio.2017.36

Crotty Alexander, L. E., Drummond, C. A., Hepokoski, M., Mathew, D., Moshensky, A., Willeford, A., et al. (2018). Chronic inhalation of e-cigarette vapor containing nicotine disrupts airway barrier function and induces systemic inflammation and multiorgan fibrosis in mice. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 314 (6), R834-R847–r847. doi:10.1152/ajpregu.00270.2017

Dutta, J., Singh, S., Greeshma, M. V., Mahesh, P. A., and Mabalirajan, U. (2024). Diagnostic challenges and pathogenetic differences in biomass-smoke-induced versus tobacco-smoke-induced copd: a comparative review. *Diagn. Basel, Switz.* 14 (19), 2154. doi:10.3390/diagnostics14192154

Ekker, M. S., Verhoeven, J. I., Schellekens, M. M. I., Boot, E. M., van Alebeek, M. E., Brouwers, P., et al. (2023). Risk factors and causes of ischemic stroke in 1322 young adults. *Stroke* 54 (2), 439–447. doi:10.1161/STROKEAHA.122.040524

Elser, H., Vijayaraghavan, M., and Kasner, S. E. (2023). E-cigarettes and stroke risk-present uncertainties and future directions. *JAMA Neurol.* 80 (9), 883–884. doi:10.1001/jamaneurol.2023.2050

Feigin, V. L., Abate, M. D., Abate, Y. H., Abd ElHafeez, S., Abd-Allah, F., Abdelalim, A., et al. (2024). Global, regional, and national burden of stroke and its risk factors, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. *Lancet Neurol.* 23 (10), 973–1003. doi:10.1016/s1474-4422(24)00369-7

Feigin, V. L., Owolabi, M. O., and World Stroke Organization–Lancet Neurology Commission Stroke Collaboration Group (2023). Pragmatic solutions to reduce the global burden of stroke: a World stroke organization-lancet neurology commission. *Lancet Neurol.* 22 (12), 1160–1206. doi:10.1016/S1474-4422(23)00277-6

Fu, M., Mei, A., Min, X., Yang, H., Wu, W., Zhong, J., et al. (2024). Advancements in cardiovascular disease research affected by smoking. *Rev. Cardiovasc. Med.* 25 (8), 298. doi:10.31083/j.rcm2508298

Goniewicz, M. L., Gawron, M., Smith, D. M., Peng, M., Jacob, P., and Benowitz, N. L. (2017). Exposure to nicotine and selected toxicants in cigarette smokers who switched to electronic cigarettes: a longitudinal within-subjects observational study. *Nicotine Tob. Res.* 19 (2), 160–167. doi:10.1093/ntr/ntw160

Goniewicz, M. L., Knysak, J., Gawron, M., Kosmider, L., Sobczak, A., Kurek, J., et al. (2014). Levels of selected carcinogens and toxicants in vapour from electronic cigarettes. *Tob. Control* 23 (2), 133–139. doi:10.1136/tobaccocontrol-2012-050859

Higashi, Y. (2023). Smoking cessation and vascular endothelial function. *Hypertens. Res.* 46 (12), 2670–2678. doi:10.1038/s41440-023-01455-z

Hilkens, N. A., Casolla, B., and Leung, T. W. (2024). de Leeuw FE: **Stroke**. *Lancet* 403 (10446), 2820–2836. doi:10.1016/S0140-6736(24)00642-1

Hjort, G., Schwarz, C. W., Skov, L., and Loft, N. (2024). Clinical characteristics associated with response to biologics in the treatment of psoriasis: a meta-analysis. *JAMA dermatol.* 160 (8), 830–837. doi:10.1001/jamadermatol.2024.1677

Huang, Q., Lan, X., Chen, H., Li, H., Sun, Y., Ren, C., et al. (2023). Association between genetic predisposition and disease burden of stroke in China: a genetic epidemiological study. *Lancet Reg. Health West Pac* 36, 100779. doi:10.1016/j.lanwpc.2023.100779

Jang, H. J., Min, H. Y., Kang, Y. P., Boo, H. J., Kim, J., Ahn, J. H., et al. (2024). Tobacco-induced hyperglycemia promotes lung cancer progression via cancer cell-macrophage interaction through paracrine IGF2/IR/NPM1-driven PD-L1 expression. *Nat. Commun.* 15 (1), 4909. doi:10.1038/s41467-024-49199-9

Jumbe, S., Mwenda Kamninga, T., Kalu, U. G., Nyali, J., Saleh, L., Newby, C., et al. (2025). A systematic review and meta-analysis of factors associated with adolescent substance use in Africa, 2000 to 2020. *Addict. Abingdon, Engl.* doi:10.1111/add.70023

Kaisar, M. A., Villalba, H., Prasad, S., Liles, T., Sifat, A. E., Sajja, R. K., et al. (2017). Offsetting the impact of smoking and e-cigarette vaping on the cerebrovascular system and stroke injury: is Metformin a viable countermeasure? *Redox Biol.* 13, 353–362. doi:10.1016/j.redox.2017.06.006

Liang, F., Wang, G. Z., Wang, Y., Yang, Y. N., Wen, Z. S., Chen, D. N., et al. (2022). Tobacco carcinogen induces tryptophan metabolism and immune suppression via induction of indoleamine 2,3-dioxygenase 1. *Signal Transduct. Target Ther.* 7 (1), 311. doi:10.1038/s41392-022-01127-3

Lin, M. P., Ovbiagele, B., Markovic, D., and Towfighi, A. (2016). Association of secondhand smoke with stroke outcomes. *Stroke* 47 (11), 2828–2835. doi:10.1161/STROKEAHA.116.014099

Liu, Y., Wang, W., Cui, X., Lyu, J., and Xie, Y. (2024). Exploring genetic associations of 3 types of risk factors with ischemic stroke: an integrated bioinformatics study. *Stroke* 55 (6), 1619–1628. doi:10.1161/STROKEAHA.123.044424

Lu, R., Qin, Y., Xie, C., Tan, X., Zhu, T., Tan, J., et al. (2024). Secondhand smoke exposure can increase the risk of first ischemic stroke: a 10.7-year prospective cohort study in China. *Ann. Epidemiol.* 92, 25–34. doi:10.1016/j.annepidem.2024.02.005

Maida, C. D., Norrito, R. L., Rizzica, S., Mazzola, M., Scarantino, E. R., and Tuttolomondo, A. (2024). Molecular pathogenesis of ischemic and hemorrhagic strokes: background and therapeutic approaches. *Int. J. Mol. Sci.* 25 (12), 6297. doi:10.3390/ijms25126297

Martin, S. S., Aday, A. W., Almarzooq, Z. I., Anderson, C. A. M., Arora, P., Avery, C. L., et al. (2024). 2024 heart disease and stroke statistics: a report of US and global data from the American heart association. *Circulation* 149 (8), e347–e913. doi:10.1161/CIR.0000000000001209

Notley, C., Gentry, S., Livingstone-Banks, J., Bauld, L., Perera, R., Conde, M., et al. (2025). Incentives for smoking cessation. *Cochrane database Syst. Rev.* 1 (1), Cd004307. doi:10.1002/14651858.CD004307.pub6

Okafor, C. N., Okafor, N., Kaliszewski, C., and Wang, L. (2022). Association between electronic cigarette and combustible cigarette use with cardiometabolic risk biomarkers among U.S. adults. *Ann. Epidemiol.* 71, 44–50. doi:10.1016/j.annepidem.2022.02.002

Papadopoulou, A., Harding, D., Slabaugh, G., Marouli, E., and Deloukas, P. (2024). Prediction of atrial fibrillation and stroke using machine learning models in UK Biobank. *Heliyon* 10 (7), e28034. doi:10.1016/j.heliyon.2024.e28034

Qasim, H., Karim, Z. A., Rivera, J. O., Khasawneh, F. T., and Alshbool, F. Z. (2017). Impact of electronic cigarettes on the cardiovascular system. *J. Am. Heart Assoc.* 6 (9), e006353. doi:10.1161/JAHA.117.006353

Qin, W., Zhang, L., Li, Z., Xiao, D., Zhang, Y., Zhang, H., et al. (2020). Endothelial to mesenchymal transition contributes to nicotine-induced atherosclerosis. *Theranostics* 10 (12), 5276–5289. doi:10.7150/thno.42470

Rezk-Hanna, M., Mosenifar, Z., Benowitz, N. L., Rader, F., Rashid, M., Davoren, K., et al. (2019). High carbon monoxide levels from charcoal combustion mask acute endothelial dysfunction induced by hookah (waterpipe) smoking in young adults. *Circulation* 139 (19), 2215–2224. doi:10.1161/CIRCULATIONAHA.118.037375

Shi, J., Xiong, L., Guo, J., and Yang, Y. (2023). The association between combustible/electronic cigarette use and stroke based on national health and nutrition examination survey. *BMC Public Health* 23 (1), 697. doi:10.1186/s12889-023-15371-x

Siegel, J., Patel, S. H., Mankaliye, B., and Raval, A. P. (2022). Impact of electronic cigarette vaping on cerebral ischemia: what we know so far. *Transl. Stroke Res.* 13 (6), 923–938. doi:10.1007/s12975-022-01011-w

Singh, M., Kumar, A., Khanna, N. N., Laird, J. R., Nicolaides, A., Faa, G., et al. (2024). Artificial intelligence for cardiovascular disease risk assessment in personalised framework: a scoping review. *EClinicalMedicine* 73, 102660. doi:10.1016/j.eclinm.2024.102660

Sood, A. K., Kesic, M. J., and Hernandez, M. L. (2018). Electronic cigarettes: one size does not fit all. *J. Allergy Clin. Immunol.* 141 (6), 1973–1982. doi:10.1016/j.jaci.2018.02.029

Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Anderson, C. A. M., Arora, P., Avery, C. L., et al. (2023). Heart disease and stroke statistics-2023 update: a report from the American heart association. *Circulation* 147 (8), e93–e621. doi:10.1161/CIR.0000000000001123

Tu, W. J., Wang, L. D., and Special Writing Group of China Stroke Surveillance Report (2023b). China stroke surveillance report 2021. *Mil. Med. Res.* 10 (1), 33. doi:10.1186/s40779-023-00463-x

Tu, W. J., Zhao, Z., Yin, P., Cao, L., Zeng, J., Chen, H., et al. (2023a). Estimated burden of stroke in China in 2020. *JAMA Netw. Open* 6 (3), e231455. doi:10.1001/jamanetworkopen.2023.1455

Wang, C., Liu, C., Shi, J., Li, H., Jiang, S., Zhao, P., et al. (2023). Nicotine exacerbates endothelial dysfunction and drives atherosclerosis via extracellular vesicle-miRNA. *Cardiovasc Res.* 119 (3), 729–742. doi:10.1093/cvr/cvac140

Wang, X., Liu, X., O'Donnell, M. J., McQueen, M., Sniderman, A., Pare, G., et al. (2024b). Tobacco use and risk of acute stroke in 32 countries in the INTERSTROKE study: a case-control study. *EClinicalMedicine* 70, 102515. doi:10.1016/j.eclinm.2024.102515

Wang, Y., Ge, Y., Yan, W., Wang, L., Zhuang, Z., and He, D. (2024a). From smoke to stroke: quantifying the impact of smoking on stroke prevalence. *BMC Public Health* 24 (1), 2301. doi:10.1186/s12889-024-19754-6

Whitehead, A. K., Erwin, A. P., and Yue, X. (2021). Nicotine and vascular dysfunction. *Acta Physiol. (Oxf)* 231 (4), e13631. doi:10.1111/apha.13631