International Journal of General Medicine

Dovepress

open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Stroke Risk Factors in United States: An Analysis of the 2013–2018 National Health and Nutrition Examination Survey

Zhouming Ren
Xinzheng Fu

Department of Neurology, Haining People's Hospital, Haining, Zhejiang, People's Republic of China

Purpose: This research intended to identify significant risk factors of stroke among the elderly population in the United States using the k-means clustering method.

Patients and Methods: In this cross-sectional study, we analyzed data of 4346 subjects aged ≥60 years using the National Health and Nutrition Examination Survey (NHANES) 2013–2018 datasets. Questionnaire data, dietary data, and laboratory data were accessed to acquire measurements of the potential risk factors. A pre-defined classification method was used based on the Medical Condition Questionnaire to define the stroke group. K-means clustering analysis used all potential risk factors for differentiating both groups. A stepwise logistic regression analysis examined the association between significant risk factors and the odds of stroke.

Results: Age (OR:1.053, 95% CI:1.029–1.077), diabetes (OR: 28.019, 95% CI: 19.139–41.020), glycohemoglobin (OR: 2.309, 95% CI: 1.818–2.934), plasma fasting glucose (OR: 1.017, 95% CI: 1.010–1.024), hypertension (OR: 2.343, 95% CI: 1.602–3.426), dietary fiber consumption (OR:0.980, 95% CI:0.964–0.995), and education level (OR:0.541, 95% CI: 0.411–0.713) were identified as significant risk factor for stroke among the elderly population in the k-means clustering method. In the pre-defined grouping method, age (OR:1.093, 95% CI:1.054–1.132), diabetes (OR:2.228, 95% CI: 1.432–3.466), hypertension (OR:2.295, 95% CI:1.338–3.938), and dietary fiber consumption (OR: 0.966, 95CI%:0.947–0.985) were found to influence to the risk of stroke.

Conclusion: Age, hypertension, dietary fiber consumption, and education level are the significant risk factors of stroke among elders aged >60 years. Among all the risk factors, diabetes is the strongest predictor of stroke. Glycohemoglobin and plasma fasting glucose are also associated with stroke risks, implying that glycemic control is particularly crucial in stroke prevention and management among older adults.

Keywords: diabetes, hyperglycemia, k-means clustering method

Introduction

Stroke is the second leading cause of death worldwide, resulting in 15.2 million deaths in 2015.[1] Besides high mortality rate, disability is a significant component of the disease burden of stroke. As the third leading cause of disability adjusted life year (DALYs), stroke costs $35.8 billion annually.[2,3] The American Heart Association (AHA) estimates 7.6 million (2.7%) Americans aged ≥20 years having had a stroke.[4] The evidence of high occurrence rate and substantial burden of stroke has led to a vast research endeavor. As population ages, stroke risk is expected to go

Correspondence: Xinzheng Fu
Department of Neurology, Haining People's Hospital, No. 2 Qianjiang West Road, Haining, Zhejiang, People's Republic of China
Tel +86-15325739220
Email fxzncfxzncfxznc@126.com

Received: 29 June 2021
Accepted: 9 September 2021
Published: 28 September 2021

up.[5] Therefore, precisely identifying the risk factors of the disease is pivotal to reduce the impact of stroke.

Current studies on stroke risk factor are mainly review studies, analyzing previous research that uses pre-defined stroke outcomes. The pre-defined stroke group may categorize borderline or undiagnosed individuals as not having a stroke.[6] However, these borderline cases may possess shares similar characteristics with the stroke patients. Therefore, this research aims to examine prominent risk factors of stroke using a clustering method.

The k-means clustering is an unsupervised learning that groups the non-explicitly labeled data while maximizing the heterogeneity among groups.[7] The method can be used to reveal similarities of unknown groups in a complex dataset. Unlike classification by the pre-defined outcomes, k-means clustering uses vector quantization for grouping elements. Thus, the k-means clustering identifies the potential stroke risk factors based on the characteristics of the study participants, ignoring any pre-defined criteria.

In this research, we intend to examine potential significant risk factors proposed in previous studies[3,8–12] by a k-means clustering method and compare it with the analysis using a pre-defined stroke group, aiming to provide more accurate identification.

Materials and Methods

Study Design

The current study is a cross-sectional research, retrieving data from 2013–2014, 2015–2016, and 2017–2018 National Health and Nutrition Examination Survey (NHANES) year. The NHANES is a continuous nation-wide health program conducted by the National Center for Health Statistics (NCHS).[13] Approximately 5000 people were sampled each year. These people distributed in counties across the country, and 15 counties were visited every year. The data collection process consisted of two parts, an in-person interview, and a physical examination performed in the Mobile Examination Center (MEC). All collected data was de-identified and released for public use, available on the NHANES official website (https://www.cdc.gov/nchs/nhanes/index.htm). NHANES was conducted in agreement with the Helsinki Declaration, the protocols of which were approved by the National Center for Health Statistics Ethics Review Board.[14]

Study Participants

In the NHANES 2013–2018 dataset, elders aged 60 years or older with complete medical condition information were eligible for the study (n=5261). Participants with missing data in dietary and baseline characteristics were excluded (n=915). In total, 4346 participants were included in the final analyses. The detailed selection of eligible participants was presented in Figure 1.

Stroke Assessment

The pre-defined stroke groups were determined based on the Medical Condition Questionnaire (MCQ). During the in-person interview, question MCQ 160f "Has a doctor or other health professional ever told {you/SP} that {you/s/he} had a stroke?" was asked by trained interviewers. Participants who answered "Yes" were identified as having had a stroke and classified in the stroke group in the analysis using pre-defined outcomes.

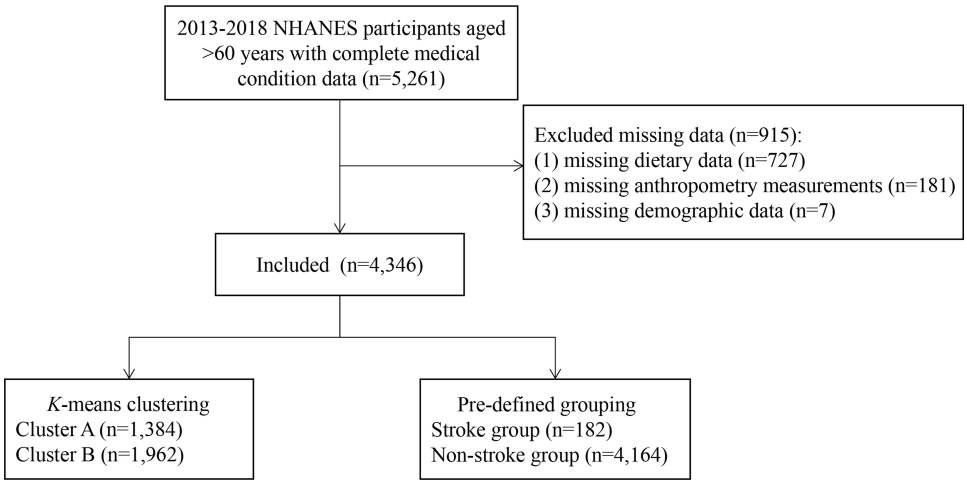

Figure 1 Flow chart of selecting eligible participants.

Risk Factors Measurements

Risk factors were assessed using the questionnaire data, examination data, and laboratory data. During the MEC interview, NHANES questionnaires were administered by trained interviewers using the Audio computer assisted personal self interview (ACASI) Computer-Assisted Personal Interview (CAPI) system.[15] The NHANES examination was performed in the MEC where participants underwent the anthropometry examination under a controlled environment.[16] The data was collected through a computerized data collection process with a built-in data entry quality control checks. Biospecimens, including blood, urine, oral rinse, and vaginal/penile swabs, were collected during the MEC examination to provide a detailed evaluation of the participants' health conditions and nutritional status.[17] Collected data were entered directly into a computerized database and underwent internal and external quality assurance and quality control.

Demographic Characteristics

Demographic variables were retrieved from the Demographic Variables and Sample Weights file (DEMO). Information regarding age, gender, race, education level, marital status, physical activity, and poverty income ratio (PIR) was extracted from the DEMO data files.

Weight

Bodyweight was measured by a calibrated digital weight scale, and height was measured using a stadiometer. Body mass index (BMI) was calculated and rounded to one decimal place. The BMI data of the study participants was available in the exanimation dataset Body Measures datafile.

Hypertension

Blood Pressure & Cholesterol Questionnaire (BPQ) question BPQ020 asked, "{Have you/Has SP} ever been told by a doctor or other health professional that {you/s/he} had hypertension, also called high blood pressure?" Participants who answered "yes" were considered as having hypertension.

Diabetes

Diabetes Questionnaire (DIQ) question DIQ010 asked, "The next questions are about specific medical conditions. {Other than during pregnancy, {have you/has SP}/{Have you/Has SP}} ever been told by a doctor or health professional that {you have/{he/she/SP} has} diabetes or sugar diabetes?" Participants who answered "yes" were considered diabetic.

Cardiovascular Disease

If the participant answered "Yes" to any of the following questions in the MCQ, the individual was considered as having cardiovascular disease.

(i) MCQ160b: "Has a doctor or other health professional ever told {you/SP} that {you/s/he} had congestive heart failure (CHF)?"
(ii) MCQ160c Has a doctor or other health professional ever told {you/SP} that {you/s/he} had coronary heart disease (CHD)?
(iii) MCQ160d: "Has a doctor or other health professional ever told {you/SP} that {you/s/he} had angina, also called angina pectoris?"
(iv) MCQ160e: "Has a doctor or other health professional ever told {you/SP} that {you/s/he} had a heart attack (HA), also called myocardial infarction)?"

Smoking

Smokers were defined using the Smoking-Cigarette Use Questionnaire (SMQ). Participants who answered "yes" to question SMQ020 "{Have you/Has SP} smoked at least 100 cigarettes in {your/his/her} entire life?" were classified as smokers.

Dietary Intake

Dietary intake was estimated by 24-hour dietary recall, a validated USDA Automated Multiple-Pass Method.[18] The specific intake of each nutrient was available in the Dietary Interview-Total Nutrients Intakes. Consumptions of dietary fiber, vitamin A, vitamin E, vitamin C, vitamin D, polyunsaturated fatty acids (PUFA), and alcohol were retrieved from the dietary data. Alcohol consumers were identified if the alcohol consumption was >0 mg/day. PUFA was categorized into six groups on a 5 g incremental basis.

Laboratory Assessment

Laboratory data was accessed to acquire plasma biomarkers and indicators of lipid profile and glycemic control. Cholesterol-High-Density Lipoprotein, Cholesterol-Low-Density Lipoproteins & Triglycerides, Cholesterol-Total, Glycohemoglobin, and Plasma Fasting Glucose data files were used to extract high-density lipoprotein (HDL), low-

Table 1 Baseline Characteristics of the Overall Study Participants, NHANES 2013–2018 (n=4346)

Baseline Characteristics	Total (n=4346)
Age, years, M [Q_1,Q_3]	68.00 [63.00,75.00]
Gender, n (%)	
Male	2178 (45.85)
Female	2168 (54.15)
Race, n (%)	
Mexican American	521 (3.98)
Non-Hispanic Black	925 (8.24)
Non-Hispanic White	1985 (78.09)
Others	915 (9.69)
BMI, kg/m^2, n (%)	
<18.5	1194 (27.57)
18.5-	42 (0.97)
25.0-	1709 (38.32)
30.0-	1401 (33.14)
Education level, n (%)	
Less than 12th grade	1030 (12.69)
High school or above	3316 (87.31)
Marital status, n (%)	
Divorced/separated	772 (14.92)
Married	2545 (64.72)
Widowed	788 (16.24)
Single	241 (4.11)
PIR, M [Q_1,Q_3]	3.18 [1.77,5.00]
Smoking, n (%)	
No	3268 (79.25)
Yes	1078 (20.75)
Alcohol consumption, n (%)	
No	3532 (77.82)
Yes	814 (22.18)
CHF, n (%)	
Yes	288 (5.35)
No	4058 (94.65)
CHD, n (%)	
Yes	421 (10.15)
No	3925 (89.85)
Angina, n (%)	
Yes	210 (5.09)
No	4136 (94.91)
HA, n (%)	
Yes	375 (7.84)
No	3971 (92.16)
Diabetes, n (%)	
Yes	1181 (21.75)
No	3165 (78.25)
Hypertension, n (%)	
Yes	2635 (57.39)
No	1711 (42.61)
HDL, mg/dL, M [Q_1,Q_3]	54.00 [44.00,67.00]
TG, mg/dL, M [Q_1,Q_3]	107.00 [54.00,210.00]
LDL, mg/dL, M [Q_1,Q_3]	101.00 [80.00,124.00]
TC, mg/dL, M [Q_1,Q_3]	189.00 [160.00,217.00]
GHb, %, M [Q_1,Q_3]	5.70 [5.40,6.10]
GLU, mg/dL, M [Q_1,Q_3]	107.00 [99.00,121.00]
Dietary fiber, g, M [Q_1,Q_3]	14.90 [10.10,21.30]
Vitamin A, mcg, M [Q_1,Q_3]	524.00 [304.00,833.00]
Vitamin E, mg, M [Q_1,Q_3]	7.37 [4.77,10.96]
Vitamin C,mg, M [Q_1,Q_3]	55.60 [23.50,109.60]
Vitamin D, mcg, M [Q_1,Q_3]	3.20 [1.40,5.80]
PUFA, n (%)	
<5	315 (5.29)
5-	854 (17.96)
10-	968 (21.74)
15-	785 (19.17)
20-	544 (13.33)
25-	880 (22.52)

Abbreviations: BMI, body mass index; PIR, poverty income ratio; CHF, congestive heart failure; CHD, coronary heart disease; HA, heart attack; HDL, high-density lipoprotein; LDL, low-density lipoproteins; TG, triglycerides; TC, total cholesterol; GHb, glycohemoglobin; GLU, plasma fasting glucose; PUFA, polyunsaturated fatty acids.

density lipoproteins (LDL), triglycerides (TG), total cholesterol (TC), glycohemoglobin (GHb), and plasma fasting glucose (GLU) levels.

Statistical Analysis

Data extraction was performed by R 4.0.2. The SPSS Statistics 23.0 (IBM Corporation. Armonk, NY, USA) was used for clustering. The SAS 9.4 (SAS Institute, inc. Cary, NC, USA) was used to identify risk factors. A p value of less than 0.05 was defined as significant. Sample weights (WTINT2YR) were applied to all analyses to ensure the representativeness of the study sample.

Continuous variables were examined for normality by the Shapiro normality test. Normally distributed continuous variables were presented in mean and standard

deviation (mean±SD) and compared using the independent *t*-test. Non-normally distributed variables, displayed in median and interquartile range [M(Q1–Q3)], were compared by the Mann–Whitney *U*-test. Categorical variables were expressed in frequencies and proportions (n%) and compared using the Pearson's chi-square test (χ^2) and Fisher's exact test when appropriate.

K-means clustering method was implemented to define subgroups of stroke. All risk factors were applied as clustering variables in this research. Each clustering variable served as an axis to cluster the observations. The observations were assigned to the nearest centroid. The grouping process was completed when all centroids had become static, and all observations had been positioned. Once the stroke subgroups were developed, intergroup comparisons were made to identify variables that were significantly different. Multivariate stepwise regression was implemented to investigate the potential stroke risk factors and obtain the odds ratio (OR), 95% confidence interval (95% CI), and p values. Receiver Operator Characteristic (ROC) curves were applied to evaluate and compare the performance of classification.

Results

Study Population

Characteristics of the study population were summarized in Table 1. Of the included 4346 people, the median age was 68 years, with more female participants than male participants (54.15% vs 45.85%). Most participants were non-Hispanic whites (78.09%), followed by non-Hispanic blacks (8.24%), others (9.69%), and Mexican Americans (3.98%). A total of two-thirds of the populations were observed to be overweight (38.32%) and obese (33.14%). There were more married (64.72%) participants than widowed (16.24%), divorced or separated (14.92%), and single participants (4.11%). Most participants were non-smokers (79.25%), and alcohol consumption was noted in 22.18% of population. Most people were not diagnosed with CHF, CHD, angina, and HA, corresponding to 94.65%, 89.85%, 94.91%, and 92.16% of the overall population. The study population consisted of 21.75% diabetic patients. More than half of the study participants (57.39%) were diagnosed with hypertension. The median of HDL, TG, LDL, and TC level was 54.00 mg/dL,

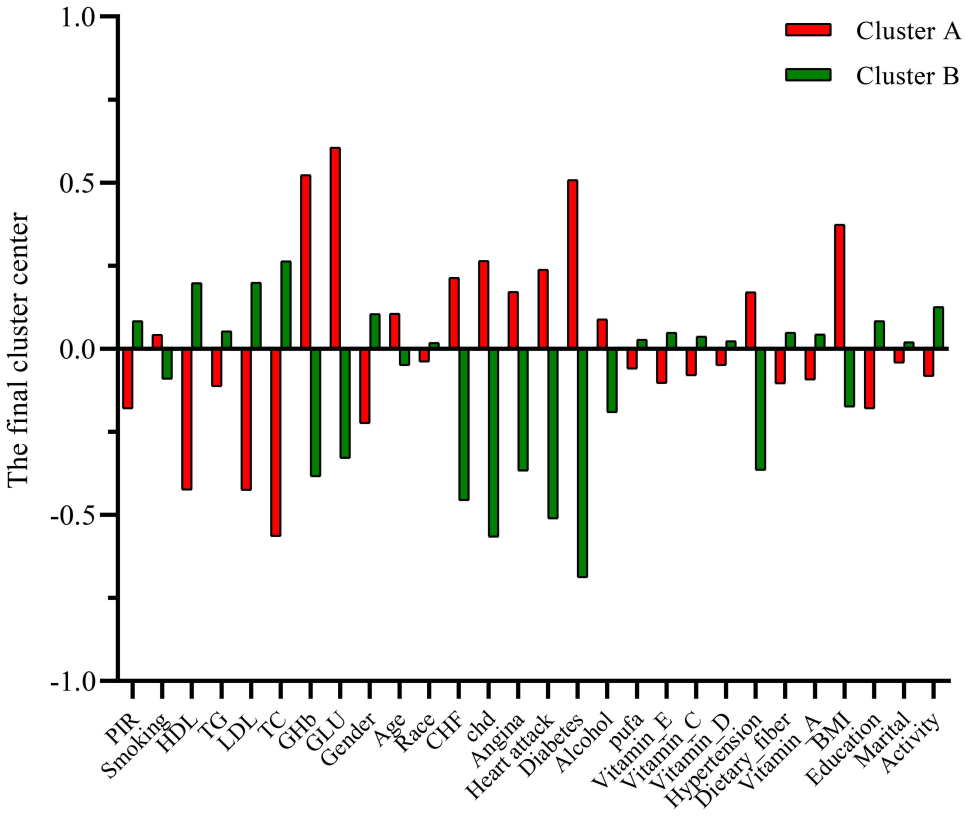

Figure 2 K-means clustering: centroids of each cluster.

Abbreviations: PIR, poverty income ratio; HDL, high-density lipoprotein; TG, triglycerides; LDL, low-density lipoproteins; TC, total cholesterol; GHb, glycohemoglobin; GLU, plasma fasting glucose; CHF, congestive heart failure; CHD, coronary heart disease; HA, heart attack; PUFA, polyunsaturated fatty acids; BMI, body mass index.

107.00 mg/dL, 101.00 mg/dL, and 189.00 mg/dL, respectively. The median GHb and GLU level was 5.70% and 107.00 mg/dL, respectively. The median dietary fiber, vitamin A, vitamin E, vitamin C, and vitamin D intake was 14.90 mg, 524.00 mcg, 7.37mg, 55.60 mg, 3.20 mcg, respectively.

K-Means Clustering

When applying the k-means clustering analysis, the study population was grouped into two clusters, Cluster A and Cluster B. There were 1384 participants in Cluster A and 1962 participants in Cluster B. The final clustering centers, as known as the centroids, were presented in Figure 2. The overall risk of stroke was 4.19%. The risk of stroke in Cluster A was 7.56% (Figure 3), while the risk of stroke in Cluster B was 2.60%. A significant difference in the stroke incidence was detected (χ^2=57.965, P<0.001) between Cluster A, 7.56%, and Cluster B, 2.60%.

When comparing the demographic characteristics (Table 2), age (Z=667.598, P<0.001), gender (χ^2=46.793, P<0.001), race (χ^2=43.418, P<0.001), education level (χ^2=38.397, P<0.001), and PIR (Z=−999.692, P<0.001) were significantly different between Cluster A and Cluster B. The proportion of physical activity (χ^2=434.774, P<0.001), alcohol consumers (χ^2=60.299, P<0.001), CHF patients (χ^2=91.344, P<0.001), CHD patients (χ^2=126.416, P<0.001), angina patients (χ^2=60.128, P<0.001), HA patients (χ^2=124.904, P<0.001), diabetic patients (χ^2=461.741, P<0.001), and hypertension patients (χ^2=91.259, P<0.001) were also significantly between Cluster A and Cluster B. Additionally, disparities were observed in the following: BMI (χ^2=3.123, P<0.001), HDL (Z=−2490.96, P<0.001), TG (Z=−230.891, P<0.001), LDL (Z=−2183.04, P<0.001), TC (Z=−2921.49, P<0.001), GHb (Z=3940.41, P<0.001), GLU (Z=3081.21, P<0.001), Dietary fiber (Z=−472.144, P<0.001), vitamin A (Z=−450.076, P<0.001), vitamin E (Z=−556.366, P<0.001), vitamin C (Z=−570.544, P<0.001), vitamin D (Z=−202.439, P<0.001), PUFA (χ^2=9.001, P=0.109).

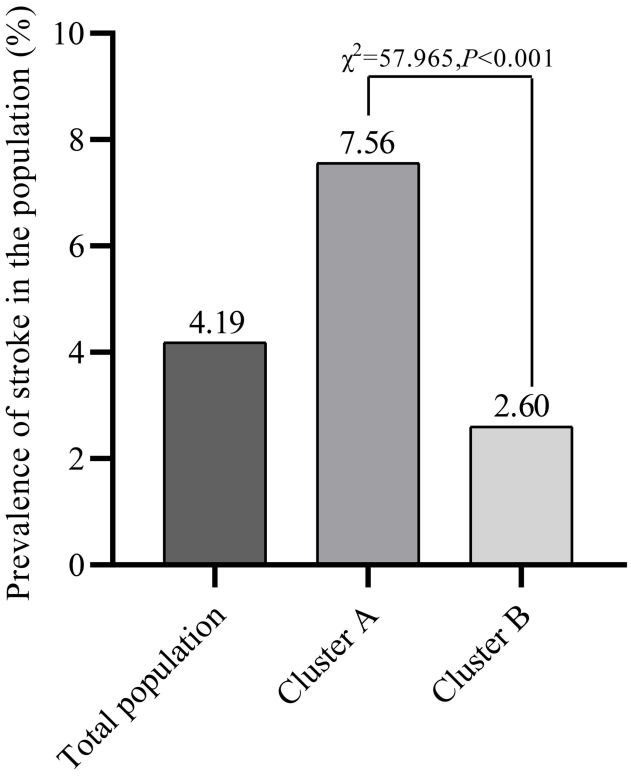

Figure 3 K-means clustering: the risk of stroke of each cluster.

Pre-Defined Stroke Group

When defining the stroke subgroups based on the MCQ, the stroke group included 182 participants, and the non-stroke group contained 4164 people. As summarized in Table 3, the demographic comparison discovered significant differences in age (Z=958.729, P<0.001), race (χ^2=8.974, P=0.048), education level (χ^2=7.614, P=0.008), physical activity (χ^2=11.529, P=0.009), and PIR (Z=−473.070, P<0.001) between the stroke and non-stroke group. In terms of cardiovascular diseases, only the prevalence of CHF was significantly different between the stroke and non-stroke groups (χ^2=4.236, P=0.045). The stroke group consisted of a higher proportion of diabetic participants than the non-stroke patients (χ^2=13.591, P=0.001). A larger percentage of the hypertension patients was in the stroke group than the non-stroke group (χ^2=19.385, P <0.001). The HDL (Z=−358.113, P<0.001), TG (Z=−143.048, P<0.001), LDL (Z=−457.960, P<0.001), TC (Z=−559.595, P<0.001), GHb (Z=299.764, P<0.001), and GLU (Z=370.373, P<0.001) levels were also significantly different between the stroke and non-stroke group. Dietary intakes of fiber (Z=−455.767, P<0.001), vitamin A (Z=−112.521, P<0.001), vitamin E (Z=−313.950, P<0.001), vitamin C (Z=−138.531, P<0.001), and vitamin D (Z=−82.555, P<0.001) were significantly different between the stroke and non-stroke group.

Stepwise Logistic Regression Analysis

After stepwise logistic regression analysis, age, diabetes, hypertension, dietary fiber consumption, education level,

Table 2 Baseline Characteristics According to the Risk of Stroke, k-Means Clustering Method

Variables	K-Means Clustering		Statistics	*P*
	Cluster A* n=1384	Cluster B† n=1962		
Age, years, M [Q_1,Q_3]	70.00 [65.00,76.00]	68.00 [63.00,75.00]	Z=667.598	<0.001
Gender, n (%)‡				
Male	849 (61.03)	1329 (40.71)	χ^2=46.793	<0.001
Female	535 (38.97)	1633 (59.29)		
Race, n (%)				
Mexican American	224 (6.61)	297 (3.09)	χ^2=43.418	<0.001
Non-Hispanic Black	294 (10.08)	631 (7.61)		
Non-Hispanic White	571 (71.59)	1414 (80.30)		
Others	295 (11.72)	620 (9.00)		
BMI, kg/m^2, n (%)				
<18.5	226 (14.04)	968 (32.15)	χ^2=3.123	<0.001
18.5-	5 (0.25)	37 (1.22)		
25.0-	506 (34.85)	1203 (39.50)		
30.0-	647 (50.86)	754 (27.13)		
Education level, n (%)				
Less than 12th grade	434 (19.70)	596 (10.32)	χ^2=38.397	<0.001
High school or above	950 (80.30)	2366 (89.68)		
Marital status, n (%)				
Divorced/separated	220 (13.50)	552 (15.40)	χ^2=5.586	0.150
Married	829 (63.32)	1716 (65.20)		
Widowed	264 (18.63)	524 (15.44)		
Single	71 (4.55)	170 (3.96)		
Physical activity, n (%)				
Sedentary	2182 (45.98)	804 (54.97)	χ^2=34.774	<0.001
Insufficient	681 (16.43)	208 (16.34)		
Moderate	545 (14.90)	154 (13.25)		
High	938 (22.69)	222 (15.44)		
PIR, M [Q_1,Q_3]	2.53 [1.43,4.48]	3.43 [1.89,5.00]	Z=−999.692	<0.001
Smoking, n (%)				
No	1095 (81.11)	2173 (78.61)	χ^2=1.598	0.213
Yes	289 (18.89)	789 (21.39)		
Alcohol consumption, n (%)				
No	1228 (87.41)	2304 (74.57)	χ^2=60.299	<0.001
Yes	156 (12.59)	658 (25.43)		
CHF, n (%)				
Yes	249 (18.24)	39 (0.98)	χ^2=91.344	<0.001
No	1135 (81.76)	2923 (99.02)		
CHD, n (%)				
Yes	366 (32.67)	55 (2.52)	χ^2=126.416	<0.001
No	1018 (67.33)	2907 (97.48)		
Angina, n (%)				
Yes	176 (15.64)	34 (1.52)	χ^2=60.128	<0.001
No	1208 (84.36)	2928 (98.48)		

(*Continued*)

Table 2 (Continued).

Variables	K-Means Clustering		Statistics	*P*
	Cluster A* n=1384	Cluster B† n=1962		
HA, n (%)				
Yes	318 (25.81)	57 (1.75)	χ^2=124.904	<0.001
No	1066 (74.19)	2905 (98.25)		
Diabetes, n (%)				
Yes	1046 (74.73)	135 (3.80)	χ^2=461.741	<0.001
No	338 (25.27)	2827 (96.20)		
Hypertension, n (%)				
Yes	1086 (77.46)	1549 (50.59)	χ^2=91.259	<0.001
No	298 (22.54)	1413 (49.41)		
HDL, mg/dL, M [Q_1,Q_3]	45.00 [38.00,54.00]	57.00 [47.00,70.00]	Z=−2490.96	<0.001
TG, mg/dL, M [Q_1,Q_3]	111.00 [43.00,197.00]	106.00 [56.00,219.00]	Z=−230.891	<0.001
LDL, mg/dL, M [Q_1,Q_3]	84.00 [68.00,104.00]	108.00 [87.00,129.00]	Z=−2183.04	<0.001
TC, mg/dL, M [Q_1,Q_3]	156.00 [139.00,183.00]	199.00 [173.00,223.00]	Z=−2921.49	<0.001
GHb, %, M [Q_1,Q_3]	6.60 [6.00,7.40]	5.60 [5.40,5.80]	Z=3940.41	<0.001
GLU, mg/dL, M [Q_1,Q_3]	128.00 [108.00,163.00]	104.00 [97.00,113.00]	Z=3081.21	<0.001
Dietary fiber, g, M [Q_1,Q_3]	14.20 [9.50,19.70]	15.10 [10.30,22.00]	Z=−472.144	<0.001
Vitamin A, mcg, M [Q_1,Q_3]	486.00 [274.00,770.00]	538.00 [311.00,851.00]	Z=−450.076	<0.001
Vitamin E, mg, M [Q_1,Q_3]	6.78 [4.30,10.01]	7.49 [4.94,11.30]	Z=−556.366	<0.001
Vitamin C,mg, M [Q_1,Q_3]	45.90 [20.60,95.00]	58.60 [24.90,115.70]	Z=−570.544	<0.001
Vitamin D, mcg, M [Q_1,Q_3]	3.10 [1.30,5.50]	3.20 [1.50,5.90]	Z=−202.439	<0.001
PUFA, n (%)				
<5	110 (6.57)	205 (4.86)	χ^2=9.001	0.109
5-	295 (20.29)	559 (17.17)		
10-	311 (21.38)	657 (21.86)		
15-	238 (18.46)	547 (19.40)		
20-	159 (10.97)	385 (14.13)		
25-	271 (22.32)	609 (22.59)		

Notes: *Cluster A, high incidence of stroke, 7.15%; †Cluster B, low incidence of stroke, 2.80%; ‡n%, sample weights were applied to the all the percentages.
Abbreviations: BMI, body mass index; PIR, poverty income ratio; CHF, congestive heart failure; CHD, coronary heart disease; HA, heart attack; HDL, high-density lipoprotein; LDL, low-density lipoproteins; TG, triglycerides; TC, total cholesterol; GHb, glycohemoglobin; GLU, plasma fasting glucose; PUFA, polyunsaturated fatty acids.

GHb, and GLU were identified as risk factors in the k-means clustering analysis (Table 4). The most prominent risk factor was diabetes, associated with a 28.02 times increased risk of stroke (OR: 28.019, 95% CI: 19.139–41.020, *P*<0.001). The analysis of biomarkers yielded similar results, with a 1% increase in GHb showing a 1.31 increase in the risk of stroke (OR: 2.309, 95% CI: 1.818–2.934, *P*<0.001). As the level of GLU increased by 1 mg/dL, the risk of stroke elevated 0.017 (OR: 1.017, 95% CI: 1.010–1.024, *P*<0.001). Hypertension was associated with 2.34 times higher risk of stroke (OR: 2.343, 95% CI: 1.602–3.426, *P*<0.001). The risk of stroke increased 0.05 in each 1-year increase in age (OR:1.053, 95% CI:1.029–1.077, *P*<0.001). Every 1 g increase in dietary fiber intake was linked with a 0.02 decrease in the stroke risk (OR:0.980, 95% CI:0.964–0.995, *P*=0.016). Higher education level also had a protective effect (OR:0.541, 95% CI: 0.411–0.713, *P*<0.001).

Fewer risk factors were identified when using the pre-defined stroke groups, including age, diabetes, hypertension, and dietary fiber consumption. The effect of hypertension (OR:2.295, 95% CI:1.338–3.938, *P*=0.002) was more significant than diabetes (OR:2.228, 95% CI: 1.432–3.466, *P*<0.001) on the risk of stroke. Each 1-year increase in age was associated with 0.093 higher risk of stroke (OR:1.093, 95% CI:1.054–1.132, *P*<0.001). Dietary fiber illustrated a protective effect on stroke, each 1 g increase

Table 3 Baseline Characteristics According to the Risk of Stroke, Pre-Defined Grouping Method

Variables	Pre-Defined Grouping		Statistics	*P*
	Stroke* n=182	No Stroke n=4164		
Age, years, M [Q_1,Q_3]	74.00 [69.00,80.00]	68.00 [63.00,75.00]	Z=958.729	<0.001
Gender, n (%)†				
Male	93 (45.35)	2085 (45.87)	χ^2=0.007	0.933
Female	89 (54.65)	2079 (54.13)		
Race, n(%)				
Mexican American	12 (2.29)	509 (4.05)	χ^2=8.974	0.048
Non-Hispanic Black	34 (6.89)	891 (8.29)		
Non-Hispanic White	107 (81.55)	1878 (77.95)		
Others	29 (9.27)	886 (9.70)		
BMI, kg/m^2, n(%)				
<18.5	55 (25.85)	1139 (27.64)	χ^2=7.290	0.063
18.5-	1 (0.19)	41 (1.00)		
25.0-	74 (41.69)	1635 (38.19)		
30.0-	52 (32.26)	1349 (33.17)		
Education level, n(%)				
Less than 12th grade	57 (23.24)	973 (12.26)	χ^2=7.614	0.008
High school or above	125 (76.76)	3191 (87.74)		
Marital status, n(%)				
Divorced/separated	27 (11.74)	745 (15.05)	χ^2=5.340	0.165
Married	93 (59.63)	2452 (64.93)		
Widowed	53 (25.63)	735 (15.86)		
Single	9 (3.00)	232 (4.16)		
Physical activity				
Sedentary	120 (62.52)	2062 (45.31)	χ^2=11.529	0.009
Insufficient	21 (13.17)	660 (16.56)		
Moderate	17 (8.68)	528 (15.15)		
High	24 (15.64)	914 (22.97)		
PIR, M [Q_1,Q_3]	2.23 [1.42,4.19]	3.22 [1.78,5.00]	Z=−473.070	<0.001
Smoking, n(%)				
No	143 (85.19)	3125 (79.00)	χ^2=3.193	0.081
Yes	39 (14.81)	1039 (21.00)		
Alcohol consumption, n(%)				
No	152 (76.90)	3380 (77.85)	χ^2=0.033	0.856
Yes	30 (23.10)	784 (22.15)		
CHF, n(%)				
Yes	23 (10.36)	265 (5.15)	χ^2=4.236	0.045
No	159 (89.64)	3899 (94.85)		
CHD, n(%)				
Yes	34 (17.33)	387 (9.86)	χ^2=4.236	0.077
No	148 (82.67)	3777 (90.14)		
Angina, n(%)				
Yes	20 (10.77)	190 (4.86)	χ^2=2.5030	0.121
No	162 (89.23)	3974 (95.14)		
HA, n(%)				
Yes	23 (11.45)	352 (7.69)	χ^2=1.789	0.188
No	159 (88.55)	3812 (92.31)		

(*Continued*)

Table 3 (Continued).

Variables	Pre-Defined Grouping		Statistics	*P*
	Stroke* n=182	No Stroke n=4164		
Diabetes, n(%)				
Yes	77 (41.24)	1104 (20.96)	χ^2=13.591	0.001
No	105 (58.76)	3060 (79.04)		
Hypertension, n(%)				
Yes	143 (79.72)	2492 (56.49)	χ^2=19.385	<0.001
No	39 (20.28)	1672 (43.51)		
HDL, mg/dL, M [Q_1,Q_3]	50.00 [41.00,59.00]	54.00 [44.00,67.00]	Z=−358.113	<0.001
TG, mg/dL, M [Q_1,Q_3]	102.00 [44.00,187.00]	108.00 [55.00,210.00]	Z=−143.048	<0.001
LDL, mg/dL, M [Q_1,Q_3]	88.00 [74.00,113.00]	102.00 [81.00,125.00]	Z=−457.960	<0.001
TC, mg/dL, M [Q_1,Q_3]	173.00 [145.00,199.00]	189.00 [161.00,218.00]	Z=−559.595	<0.001
GHb, %, M [Q_1,Q_3]	5.90 [5.50,6.60]	5.70 [5.40,6.10]	Z=299.764	<0.001
GLU, mg/dL, M [Q_1,Q_3]	113.00 [102.00,133.00]	107.00 [99.00,120.00]	Z=370.373	<0.001
Dietary fiber, g, M [Q_1,Q_3]	12.50 [8.80,18.00]	15.00 [10.20,21.50]	Z=−455.767	<0.001
Vitamin A, mcg, M [Q_1,Q_3]	466.00 [259.00,834.00]	526.00 [307.00,833.00]	Z=−112.521	<0.001
Vitamin E, mg, M [Q_1,Q_3]	6.45 [4.30,9.74]	7.39 [4.80,11.00]	Z=−313.950	<0.001
Vitamin C,mg, M [Q_1,Q_3]	40.70 [20.50,121.20]	56.50 [23.60,108.60]	Z=−138.531	<0.001
Vitamin D, mcg, M [Q_1,Q_3]	3.10 [1.20,5.50]	3.20 [1.40,5.80]	Z=−82.555	<0.001
PUFA, n(%)				
<5	18 (10.05)	297 (5.09)	χ^2=16.570	0.005
5-	47 (24.58)	807 (17.69)		
10-	44 (23.53)	924 (21.67)		
15-	31 (17.60)	754 (19.23)		
20-	18 (13.07)	526 (13.34)		
25-	24 (11.18)	856 (22.98)		

Notes: *Stroke, the stroke group was defined using the medical condition questionnaire; †n%, sample weights were applied to the all the percentages.
Abbreviations: BMI, body mass index; PIR, poverty income ratio; CHF, congestive heart failure; CHD, coronary heart disease; HA, heart attack; HDL, high-density lipoprotein; LDL, low-density lipoproteins; TG, triglycerides; TC, total cholesterol; GHb, glycohemoglobin; GLU, plasma fasting glucose; PUFA, polyunsaturated fatty acids.

consumption of which was associated with 0.034 times lower risk of stroke (OR: 0.966, 95CI%:0.947–0.985, *P*<0.001).

Since a significant effect of diabetes was detect, ROC curves were plotted to interpret the performance of each group method, as presented in Figure 4. The sensitivity and specificity of the k-means clustering analysis were significantly better than that of the pre-defined grouping method. The area under curve (AUC) of the k-means clustering was 0.854 (95% CI:0.842–0.866), while the AUC of the pre-defined grouping method was 0.579 (95% CI:0.542–0.616). The AUCs of the two classification methods were significantly different (Z=13.934, *P*<0.001).

Discussion

As a deadly and debilitating disease, stroke poses profound physiological, psychological, and economic effects on patients' life, particularly among the aging population. Accurately identifying the risk factor is crucial in minimizing the burdens of stroke. Using the k-means clustering, we identified seven significant risk factors associated with the risk of stroke in the elderly population. Age, diabetes, hypertension, GHb, and GLU were positively associated with stroke incidence, while dietary fiber and educational attainment were inversely correlated with the risk of stroke.

The pre-defined grouping method yielded a smaller sample size of the stroke group when compared with the k-means clustering (182 vs 1384). Furthermore, certain risk factors were not detected using the pre-defined stroke group, including gender, BMI, marital status, smoking, alcohol consumption, CHD, angina, and HA, which were established risk factors.[11,19–21] Additionally, the ROC reflects a significantly superior specificity and sensitivity of the k-means clustering methods. Therefore, the k-means clustering analysis can detect potential significant risk factors that are ignored using the pre-defined criteria.

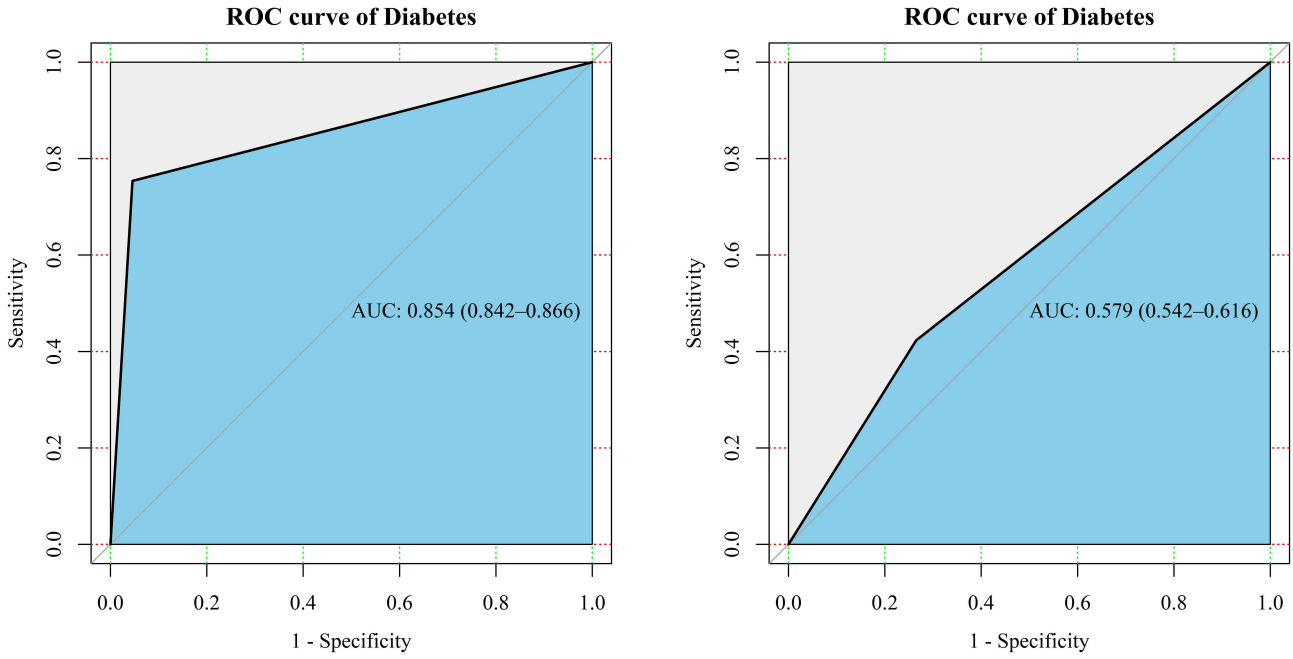

Figure 4 ROC curves evaluating the classification of diabetes.
Abbreviations: ROC, Receiver Operator Characteristic; AUC, area under curve.

Hypertension was proposed as the most potent risk factor.[22] In the present research, hypertension is also linked with stroke occurrence. However, we observed diabetes as the strongest predictor of stroke, increasing the risk of stroke by 27 times using the k-means clustering methods. The elevation of GHb and GLU level predicted the increased risk of stroke, ascertaining the effect of diabetes on stroke incidence. In contrast, diabetes was linked with a less significant impact on stroke occurrence using the pre-defined classification method, and the biomarkers were not associated with the odds of stroke. Several studies suggested that physical activity was associated with a reduced risk of stroke.[23,24] In our study, there was a statistical difference in physical activity between the stoke and non-stroke groups. However, physical activity was not related to the risk of stroke in the Logistic regression analysis. The possible explanation was that the physical activity level of the included population was unevenly distributed, and more people were distributed in sedentary and insufficient physical activity levels. Evidence suggested that the level of physical activity was associated with the risk of stroke, and light physical activity may not be related to the risk of stroke.[25–27]

The putative mechanism of diabetes's influence involves several aspects. The nitric oxide (NO)-mediated vasodilation is compromised among diabetic patients, resulting in endothelial dysfunction and a cascade reaction of atherosclerosis.[28] The reduced arterial elasticity and elevated inflammatory biomarkers among diabetic patients, such as C-reactive protein, interleukin-1, interleukin-6, and tumor necrosis factor-α, may also contribute to the higher risk of stroke. Furthermore, hyperglycemia may increase the vulnerability of vertebrobasilar arteries in diabetic patients by sympathetic denervation, elevating the risk of thrombotic infarction in the posterior cerebral circulation.[29]

Although diabetes has been established as a risk factor of stroke in previous studies,[28,30,31] the influence is not as potent as that in the current study. The significantly higher risk detected in this research may suggest the vital role of glycemic control among the elderly population and imply the accurate classification of k-means clustering methods, which discerns borderline stroke patients and reveals the critical role of diabetes in affecting the risk of stroke. The superiority of the clustering analysis has also been confirmed in previous risk factor studies analyzing the NHANES dataset.[32,33] Other strengths of the current study are the use of nationally representative sample and adequate sample size.

The shortcomings of our study are mainly the study design. The cross-sectional design limits the

Table 4 Logistic Regression Analysis of Stroke Risk Factors, Comparing the k-Means Clustering Method and the Pre-Defined Grouping Method

Variables	K-Means Clustering				Pre-Defined Grouping			
	Unadjusted		Adjusted		Unadjusted		Adjusted	
	OR (95% CI)	*P*	OR (95% CI)	*P*	OR (95% CI)	*P*	OR (95% CI)	*P*
Age*	1.031 (1.018–1.045)	<0.001	1.053 (1.029–1.077)	<0.001	1.097 (1.059–1.137)	<0.001	1.093 (1.054–1.132)	<0.001
Diabetes	74.026 (52.765–103.854)	<0.001	28.019 (19.139–41.020)	<0.001	2.647 (1.681–4.167)	<0.001	2.228 (1.432–3.466)	0.001
Hypertension	3.386 (2.649–4.328)	<0.001	2.343 (1.602–3.426)	<0.001	3.028 (1.773–5.172)	<0.001	2.295 (1.338–3.938)	0.002
Dietary fiber†	0.981 (0.971–0.991)	<0.001	0.980 (0.964–0.995)	0.011	0.961 (0.943–0.980)	<0.001	0.966 (0.947–0.985)	0.001
Education level‡	0.473 (0.388–0.577)	<0.001	0.541 (0.411–0.713)	<0.001				
GHb§	11.945 (8.947–15.947)	<0.001	2.309 (1.818–2.934)	<0.001				
GLU¶	1.049 (1.041–1.058)	<0.001	1.017 (1.010–1.024)	<0.001				

Notes: *Age, every 1-year increase in age; †Dietary fiber, every 1-gram increase in dietary fiber consumption; ‡Education level, received education of high school or above; §GHb, every 1% increase in the GHb level; ¶GLU, every 1 mg/dL increase in the GLU level.

Abbreviations: OR, odds ratio; 95% CI, 95% confidence interval; GHb, glycohemoglobin; GLU, plasma fasting glucose.

interpretation of the bidirectional relationship between stroke and the risk factors. Moreover, we were unable to separate ischemic stroke patients from hemorrhagic stroke patient since the NHANES questionnaire did not specify the stroke types. Thus, the impact of each risk factor on different types of stroke was uncertain. Yet, findings of previous studies suggest similar risk factors of ischemic stroke and hemorrhagic stroke,[28,34] possibly due to the overlapping pathophysiology of the two stroke types. Additionally, of the 101.5 million global incidences of stroke, ischemic stroke accounts for 76.1% (77.2 million) cases. Therefore, the results of this research may provide general information regarding the primary prevention and secondary management of stroke.

Besides maintaining normal blood pressure and adopting a healthy diet and lifestyle, the findings of this research underscore the importance of glycemic control in stroke prevention in the aging population. Future research examining the risk factor of stroke may specify the stroke types to obtain a more comprehensive understanding. When examining risk factors of other diseases, the k-means clustering method used in this method may achieve a more objective appraisal.

Conclusion

In summary, age, diabetes, GHb, GLU, hypertension, dietary fiber consumption, and education level are the risk factors of stroke among populations aged >60 years. Interestingly, diabetes, a modifiable risk factor, is associated with an approximately 27 times higher risk of stroke when using the k-means clustering. This research elucidates the significance of diabetes to the risk of stroke. Future studies are required to investigate the impact of each risk factor on stroke subtypes.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Katan M, Luft A. Global burden of stroke. *Semin Neurol*. 2018;38(2):208–211. doi:10.1055/s-0038-1649503
2. Kim J, Thayabaranathan T, Donnan GA, et al. Global stroke statistics 2019. *Int J Stroke*. 2020;15(8):819–838. doi:10.1177/1747493020909545
3. Yan LL, Li C, Chen Jet al. *Stroke*. The International Bank for Reconstruction and Development/The World Bank; 2017.
4. Virani Salim S, Alvaro A, Benjamin Emelia J, et al. Heart disease and stroke statistics—2020 update: a report from the American Heart Association. *Circulation*. 2020;141(9):e139–e596. doi:10.1161/CIR.0000000000000757

5. World Health Orgnization. Ageing and health. Available from: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health. Accessed November 12, 2020.
6. Mai X, Liang X. Risk factors for stroke based on the national health and nutrition examination survey. *J Nutr Health Aging*. 2020;24(7):791–795. doi:10.1007/s12603-020-1430-4
7. Hennig C, Meila M, Murtagh F, Rocci R. *Handbook of Cluster Analysis*. 1st ed. Chapman and Hall/CRC; 2015. doi:10.1201/b19706
8. Guzik A, Bushnell C. Stroke epidemiology and risk factor management. *Continuum (Minneap Minn)*. 2017;23(1):15–39. doi:10.1212/CON.0000000000000416
9. Feigin VL, Norrving B, George MG, Foltz JL, Roth GA, Mensah GA. Prevention of stroke: a strategic global imperative. *Nat Rev Neurol*. 2016;12(9):501–512. doi:10.1038/nrneurol.2016.107
10. Caprio FZ, Sorond FA. Cerebrovascular disease: primary and secondary stroke prevention. *Med Clin North Am*. 2019;103(2):295–308. doi:10.1016/j.mcna.2018.10.001
11. Virani Salim S, Alvaro A, Aparicio Hugo J, et al. Heart disease and stroke statistics—2021 update. *Circulation*. 2021;143(8):e254–e743. doi:10.1161/CIR.0000000000000950
12. Gillum R. Education, poverty, and stroke incidence in whites and blacks The NHANES I Epidemiologic Follow-up Study. *J Clin Epidemiol*. 2003;56(2):188–195. doi:10.1016/S0895-4356(02)00535-8
13. Centers for Disease Control and Prevention. NHANES - National Health and Nutrition Examination Survey Homepage; January 3, 2019. Available from: https://www.cdc.gov/nchs/nhanes/index.htm. Accessed January 16, 2019.
14. National Center for Health Statistics. NHANES - NCHS research ethics review board approval. May 8, 2019. Available from: https://www.cdc.gov/nchs/nhanes/irba98.htm. Accessed March 5, 2021.
15. National Center for Health Statistics. NHANES 2017–2018 questionnaire instruments. Available from: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/questionnaires.aspx?BeginYear=2017. Accessed March 25, 2021.
16. National Center for Health Statistics. NHANES 2017–2018 examination data overview. Available from: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overviewexam.aspx?BeginYear=2017. Accessed March 25, 2021.
17. National Center for Health Statistics. NHANES 2017–2018 laboratory data overview. Available from: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overviewlab.aspx?BeginYear=2017. Accessed March 25, 2021.
18. Blanton CA, Moshfegh AJ, Baer DJ, Kretsch MJ. The USDA automated multiple-pass method accurately estimates group total energy and nutrient intake; 2006. Available from: https://pubag.nal.usda.gov/catalog/10039. Accessed November 24, 2020.
19. Roy-O'Reilly M, McCullough LD. Age and sex are critical factors in ischemic stroke pathology. *Endocrinology*. 2018;159(8):3120–3131. doi:10.1210/en.2018-00465
20. Andersen KK, Olsen TS. Stroke case-fatality and marital status. *Acta Neurol Scand*. 2018;138(4):377–383. doi:10.1111/ane.12975
21. Howard VJ, Madsen TE, Kleindorfer DO, et al. Sex and race differences in the association of incident ischemic stroke with risk factors. *JAMA Neurol*. 2019;76(2):179–186. doi:10.1001/jamaneurol.2018.3862
22. Fryar CD, Ostchega Y, Hales CM, Zhang G, Kruszon-Moran D. Hypertension prevalence and control among adults: United States, 2015–2016. *NCHS Data Brief*. 2017;289:1–8.
23. Soares-Miranda L, Siscovick DS, Psaty BM, et al. Physical activity and risk of coronary heart disease and stroke in older adults: the cardiovascular health study. *Circulation*. 2016;133(2):147–155. doi:10.1161/circulationaha.115.018323
24. Saunders DH, Sanderson M, Hayes S, et al. Physical fitness training for stroke patients. *Cochrane Database Syst Rev*. 2020;3(3):Cd003316. doi:10.1002/14651858.CD003316.pub7
25. Yang D, Bian Y, Zeng Z, et al. Associations between intensity, frequency, duration, and volume of physical activity and the risk of stroke in middle- and older-aged Chinese people: a cross-sectional study. *Int J Environ Res Public Health*. 2020;17(22):8628. doi:10.3390/ijerph17228628
26. Kyu HH, Bachman VF, Alexander LT, et al. Physical activity and risk of breast cancer, colon cancer, diabetes, ischemic heart disease, and ischemic stroke events: systematic review and dose-response meta-analysis for the Global Burden of Disease Study 2013. *BMJ*. 2016;354:i3857. doi:10.1136/bmj.i3857
27. Kramer SF, Hung SH, Brodtmann A. The impact of physical activity before and after stroke on stroke risk and recovery: a narrative review. *Curr Neurol Neurosci Rep*. 2019;19(6):28. doi:10.1007/s11910-019-0949-4
28. Chen R, Ovbiagele B, Feng W. Diabetes and stroke: epidemiology, pathophysiology, pharmaceuticals and outcomes. *Am J Med Sci*. 2016;351(4):380–386. doi:10.1016/j.amjms.2016.01.011
29. Kuroda J, Matsuo R, Yamaguchi Y, et al. Poor glycemic control and posterior circulation ischemic stroke. *Neurol Clin Pract*. 2019;9(2):129–139. doi:10.1212/CPJ.0000000000000608
30. O'Donnell MJ, Chin SL, Rangarajan S, et al. Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. *Lancet Lond Engl*. 2016;388(10046):761–775. doi:10.1016/S0140-6736(16)30506-2
31. Lau L, Lew J, Borschmann K, Thijs V, Ekinci EI. Prevalence of diabetes and its effects on stroke outcomes: a meta-analysis and literature review. *J Diabetes Investig*. 2019;10(3):780–792. doi:10.1111/jdi.12932
32. Ghassib IH, Batarseh FA, Wang H, Borgnakke WS. Clustering by periodontitis-associated factors—A novel application to NHANES data. *J Periodontol*. 2021;92:1136–1150. doi:10.1002/JPER.20-0489
33. Bancks MP, Casanova R, Gregg EW, Bertoni AG. Epidemiology of diabetes phenotypes and prevalent cardiovascular risk factors and diabetes complications in the National Health and Nutrition Examination Survey 2003–2014. *Diabetes Res Clin Pract*. 2019;158:107915. doi:10.1016/j.diabres.2019.107915
34. Boehme AK, Esenwa C, Elkind MSV. Stroke risk factors, genetics, and prevention. *Circ Res*. 2017;120(3):472–495. doi:10.1161/CIRCRESAHA.116.308398