ISSN: 2508-7894
KJAI website: http://accesson.kr/kjai
doi: http://dx.doi.org/10.24225/kjai.2024.12.2.1

Selecting Optimal Algorithms for Stroke Prediction: Machine Learning-Based Approach

Kyung Tae CHOI[1], Kyung-A KIM[2], Myung-Ae CHUNG[3], Min Soo KANG[4]

Received: February 14, 2024. Revised: April 08, 2024. Accepted: June 05, 2024.

Abstract

In this paper, we compare three models (logistic regression, Random Forest, and XGBoost) for predicting stroke occurrence using data from the Korea National Health and Nutrition Examination Survey (KNHANES). We evaluated these models using various metrics, focusing mainly on recall and F1 score to assess their performance. Initially, the logistic regression model showed a satisfactory recall score among the three models; however, it was excluded from further consideration because it did not meet the F1 score threshold, which was set at a minimum of 0.5. The F1 score is crucial as it considers both precision and recall, providing a balanced measure of a model's accuracy. Among the models that met the criteria, XGBoost showed the highest recall rate and showed excellent performance in stroke prediction. In particular, XGBoost shows strong performance not only in recall, but also in F1 score and AUC, so it should be considered the optimal algorithm for predicting stroke occurrence. This study determines that the performance of XGBoost is optimal in the field of stroke prediction.

Keywords : Stroke prediction, Machine learning, KNHANES, XGBoost

Major Classifications: Artificial Intelligence, etc

1. Introduction

Cerebrovascular disease is the fifth cause of death in Korea after cancer, heart disease, COVID-19, and pneumonia. According to statistics on causes of death as of 2022 reported by Statistics Korea, 49.6 people per 100,000 died from cerebrovascular disease (Statistics Korea, 2022).

In particular, stroke is an important cause of death worldwide, and when it occurs, it has a high mortality rate and causes serious aftereffects. Despite current advances in medical technology, the development of treatments for stroke began in the 1980s, but no treatment has been successfully developed . Deaths due to stroke are continuously decreasing, but the risk of death is still high, and even after recovery after treatment, the frequency of cognitive impairment after stroke is reported to vary widely, ranging from 10 to 82%. The resulting physical and mental disabilities increase the burden of medical costs (Statistics Korea, 2022).

From 2018 to 2022, medical expenses for stroke patients increased by 29.0%, reaching 2.044 trillion won (Na,2020). Once a stroke occurs, it leaves permanent damage, and

* This work was supported by the research grant of the KODISA Scholarship Foundation in 2024.

1 First Author. Undergraduate
student, Dept. of Medical IT, Eulji University, Korea. Email: dohi@kakao.com
2 Second Author. Doctoral student, Dept. of Medical Artificial Intelligence, Eulji University, Korea. Email: kyungakim@naver.com
3 Third Author. Professor, Dept. of BigData Medical Convergence, Eulji University, Korea. Email: machung@eulji.ac.kr
4 Corresponding Author. Professor Dept. of BigData Medical Convergence, Eulji University, Korea. Email: mskang@eulji.ac.kr

despite active rehabilitation treatment, it is difficult to return to social life.

If stroke occurs in middle age, when stroke occurs most often, it has a greater impact on the family and community, and if stroke occurs in old age, when stroke occurs most often, the issue of maintaining a healthy life becomes more important at a time when health is weakening. There are various causes of this type of stroke, but the main cause is high blood pressure. In general, if you have high blood pressure, your risk of stroke increases from two to four times compared to other healthy people, and it is the most important cause of hemorrhagic stroke . In Korea, among people aged 20 or older with high blood pressure, the awareness rate is 67%, the treatment rate is 63%, and the control rate is 47%. It is estimated that there are about 1.27 million people in their 20s and 30s with high blood pressure, but the recognition rate is 17% and the treatment rate is 14%. is very low . Blood pressure management has a great long-term effect in preventing stroke, and once a stroke occurs, no treatment available to date can restore damaged brain tissue and its function, so prevention and early detection of stroke are important. In this context, research predicting stroke onset is of great importance. Therefore, in this paper, we analyzed stroke risk factors to find the most optimized algorithm for predicting stroke onset through machine learning and compared three algorithm models (Logistic Regression, Random Forest, and XGBoost) for predicting stroke onset.

2. Theoretical background

2.1. Stroke Risk Factors

Stroke is a cardiovascular disease in which symptoms appear suddenly after onset. Most causes of stroke are bad lifestyle habits or chronic diseases. Bad lifestyle habits include smoking, eating habits, obesity, excessive drinking, and lack of exercise, and chronic diseases include high blood pressure, diabetes, hyperlipidemia, heart disease, and existing stroke. According to the Korean Stroke Association's treatment guidelines, stroke risk factors are divided into uncontrollable risk factors, proven controllable risk factors, and potential controllable risk factors. Risk factors that cannot be controlled include age, gender, low birth weight, and genetic factors, and proven controllable risk factors include high blood pressure, smoking, diabetes, atrial fibrillation, other heart diseases, dyslipidemia, asymptomatic carotid artery stenosis, and postmenopausal hormones. Treatment, diet and nutrition, physical activity, obesity, and potential controllable risk factors include metabolic syndrome, alcohol use, drug abuse, and oral contraceptives. These include breathing problems during sleep, migraines, hyperhomocysteinemia, and hypercoagulability (Korean Stoke Society, 2023).

2.2. Logistic Regression

Logistic Regression is a probability model designed by Cox (1958) and is a statistical model used when the dependent variable is binary (one of two types). Logistic regression analysis is used to predict the probability of an event occurring using a linear combination of independent variables. The probability of occurrence is limited to a value between 0 and 1, and the results are divided into a specific classification, so it can be viewed as a classification technique.

2.3. Random Forest

Random Forest, an ensemble learning method that creates a single result using multiple results as a protocol, is a plan to combine multiple decision trees to achieve more predictive power and implement a model. Random forests, rather than decision trees, are often used to provide sufficient scaling, reduce overfitting, and improve prediction performance.

2.4. XGBoost

XGBoost is a tree boosting algorithm commonly used in data science and machine learning and is based on gradient boosting technology. Through this, decision trees are sequentially combined to form a powerful prediction model, and regulation techniques are used to reduce model complexity and prevent overfitting. XGBoost calculates the importance of each feature to facilitate data interpretation and ensures excellent performance in large-scale and high-dimensional data sets through effective parallel processing (Chen & Guestrin, 2016).

2.5. Evaluation indicators

ROC AUC represents the area under the ROC curve, a graph that visually represents the performance of a binary classification model. The ROC curve represents the recall rate and false positive rate at all possible thresholds, and the AUC value, which calculates the area under the ROC curve, is between 0 and 1. Model performance evaluation through AUC: “$0.5 \geq AUC$” cannot be used for model prediction, “$0.5 < AUC < 0.7$” model prediction performance is acceptable, “$0.7 \leq AUC < 0.8$” model prediction performance is acceptable, “$0.8 \leq AUC$” model prediction performance is acceptable. It can be seen as having excellent performance and is widely used because the performance of models can be easily compared (Draelos, 2019).

3. Main Subject

3.1. Experimental Environment

The performance of the experimental environment used for model comparison in this paper is summarized in Table 1.

Table 1: Equipment performance used for learning

Type	Content
CPU	Intel i5-13400F 2.50 GHz
GPU	NVIDIA GeForce RTX 3070
Memory	16 GB
OS	Windows 10 Pro

3.2. Experimental Data Analysis

In this paper, using raw data from the 8th National Health and Nutrition Examination Survey (2019-2021), the analysis targets were adults aged 19 to 80 who participated in the health survey and checkup survey and responded to whether they currently had a stroke. The data was analyzed by categorizing continuous variables into categories. Categorization has the advantages of improving classification model performance, improving simpler interpretation, reducing outliers and noise, considering non-linear relationships between explanatory variables, and reducing model complexity. However, note that categorization can lead to loss of information, and in some cases it may be more appropriate to use continuous variables as is. When deciding on categorization, the characteristics of the data and the requirements of the model should be considered.

Since the National Health and Nutrition Examination Survey data were extracted through a complex sample design, the analysis was performed considering the weights provided by the data. However, when using raw data from the National Health and Nutrition Examination Survey and integrating data by year within each group, the integrated weight is calculated by multiplying the existing weight by the ratio of the number of survey districts by year. In the case of the 8th period, due to COVID-19, a value proportional to the survey period of each year was assigned, and then the integrated ratio was calculated using the value assigned for each year within the integrated period and its subtotal, and then integrated by multiplying the yearly weight and the integrated ratio. Weights were calculated and used (Korea Disease Control and Prevention Agency, 2022).

Factors that affect the occurrence of stroke include risk factors such as socio-demographic characteristics, personal history, family history, health questionnaire, and health checkup results, so that they can all be reflected in the model, and variables that have a significant impact on the occurrence of stroke The Rao&Scott chi-square test was used to select them. If the P-Value value is less than 0.05 using the Rao&Scott chi-square test, it can be considered a variable that has a significant impact on the occurrence of stroke. The data may appear significant if a chi-square test is performed without applying weights, but since it can be seen as lacking in explanatory power on the basis of Korea as a whole, the test was conducted with weights applied, and the relevant information is in table 2.

The data used for analysis was a total of 17,899 people, including 7,939 men and 9,960 women. As a result of weighted frequency (%) analysis, there were 291 (0.43%) current stroke patients, 159 (0.69%) current stroke patients among men, and 132 (0.38%) current stroke patients among women. Through this analysis, age, gender, recipient of basic livelihood security, income, education, medical diagnosis of hypertension, medical diagnosis of dyslipidemia, medical diagnosis of myocardial infarction or angina pectoris, monthly drinking, exercise status, BMI, abdominal obesity, fasting blood sugar, Glycated hemoglobin and systolic blood pressure can be considered significant variables for the onset of stroke.

Table 2: variable description

Independent Variable	Categories	Current Status of stroke		P-Value
		No (n = 17608)	Yes (n = 291)	
Age	19-39	4569(24.5)	3(0)	<0.001
	40-49	3047(25.9)	9(0)	
	50-59	3256(26.0)	25(10.3)	
	60-69	3293(16.8)	80(31.1)	
	70-79	2410(4.95)	116(37.5)	
	≥80	1033(1.74)	58(21.1)	
sex	Male	7780(40.1)	159(64.6)	0.004
	Female	9828(59.9)	132(35.4)	
Recipient of national basic living	No	16441(94.5)	238(79.2)	<0.001
	Yes	1152(5.5)	53(20.8)	
Income	<200	4567(17.4)	178(58.6)	0.013
	200-400	4353(25.9)	61(19.8)	
	400-600	3791(25.9)	27(11.4)	
	≥600	4807(30.8)	25(10.2)	
edu	Elementary school graduate or less	2894(12.5)	135(40.2)	0.011
	Middle school graduate	1569(8.1)	52(21.0)	
	High Scool graduate	5603(34.8)	66(33.0)	
	University graduate and higher	6294(44.6)	35(5.58)	
Stress	Very much	822(6.04)	10(0.12)	0.251
	Much	3811(25.0)	40(4.60)	

	A little	9933(55.9)	147(68.7)	
	Almost not at all	2818(13.1)	91(26.6)	
self-perceived health	Excellent	806(3.1)	4(0.1)	0.348
	Good	4294(23.0)	38(8.3)	
	Fair	8291(55.2)	115(67.9)	
	Poor	2500(16.1)	92(13.4)	
	Very Poor	554(2.5)	41(10.3)	
family history of hypertension	No	8765(51.4)	121(71.0)	0.243
	Yes	7285(48.6)	119(29.0)	
family history of Stroke	No	13514(87.3)	162(79.4)	0.5
	Yes	2019(12.7)	66(20.6)	
Doctor-diagnosed hypertension	No	13027(73.0)	78(22.7)	0.001
	Yes	4580(27.0)	213(77.3)	
Doctor-diagnosed Dyslipidemia	No	13855(77.4)	146(52.2)	<0.001
	Yes	3750(22.6)	145(47.8)	
Doctor-diagnosed Myocardial infarction or Angina pectoris	No	15903(97.4)	153(55.5)	<0.001
	Yes	513(2.63)	138(44.5)	
monthly alcohol consumption	less than one glass per month or non-drinking	8459(34.3)	207(67.9)	0.028
	more than one glass per month	8944(65.7)	82(32.1)	
current smoke	No	14493(68.3)	250(84.2)	0.258
	Yes	2896(31.7)	39(15.8)	
Exercise	No	9427(59.3)	215(74.7)	<0.001
	Yes	6919(40.7)	71(25.3)	
BMI	<23	7135(41.1)	95(28.4)	0.048
	23-25	3982(23.3)	74(12.6)	
	≥25	6258(35.6)	112(59.1)	
Abdominal obesity	No	10878(41.7)	125(20.7)	<0.001
	Yes	6531(58.3)	161(79.3)	
Glucose	<126	15648(84.9)	215(612)	0.037
	≥126	1519(15.1)	59(38.8)	
HbA1C	<6.5	15141(82.0)	201(55.5)	0.038
	≥6.5	2020(18.0)	72(44.5)	
Systolic blood pressure	<130	12993(71.0)	170(32.1)	0.004
	≥130	4382(29.0)	119(67.9)	
Diastolic blood pressure	<90	16089(83.3)	268(74.7)	0.516
	≥90	1286(16.7)	21(25.3)	
Total cholesterol	<200	10435(51.0)	226(40.5)	0.791
	200-240	4959(48.4)	34(59.5)	
	≥240	69(0.6)	1(0.0)	
HDL cholesterol	≥40	14498(53.1)	201(62.2)	0.59
	<40	2669(46.8)	73(37.8)	
Triglyceride	<160	12485(71.6)	247(67.2)	0.917
	160-200	2355(17.7)	57(19.7)	
	≥200	2327(10.7)	27(13.1)	

3.3. Data Preprocessing

Data preprocessing is the work performed on a data set before using it to train a model. There are many missing values in the raw data of the 8th National Health and Nutrition Examination Survey (2019-2021), and there are problems in which noise must be taken into account. In this paper, significant variables resulting from the analysis and provable risk factors such as smoking and high blood pressure were extracted as variables, and rows with missing values were removed because missing values were difficult to use in the model. Through this data preprocessing, it was refined into a data set consisting of 17 columns and 16,096 rows needed for this paper. Tables 3 show information about the properties of the refined data set.

Table 3: Attributes of the Dataset

Variable	Variable description
agr_gr	1 = age is in the range [19-39] 2 = age is in the range [40-49] 3 = age is in the range [50-59] 4 = age is in the range [60-69] 5 = age is in the range [70-79] 6 = age is in the range [80+]
sex	1 = Male 2 = Female
edu	1 = Elementary school graduate or less 2 = Middle school graduate 3 = High Scool graduate 4 = University graduate and higher
allownc	0 = Not a recipient of national basic living 1 = A recipient of national basic living
ainc_gr	1 = Income is in the range [0-200] 2 = Income is in the range [200-400] 3 = Income is in the range [400-600] 4 = Income is in the range [600+]
DI1_dg	0 = The doctor has never diagnosed

	hypertension. 1 = The doctor diagnosed hypertension
DI2_dg	0 = The doctor has never diagnosed Dyslipidemia. 1 = The doctor diagnosed Dyslipidemia
DI4_dg	0 = The doctor has never diagnosed Myocardial infarction or Angina pectoris 1 = The doctor diagnosed Myocardial infarction or Angina pectoris
drink	0 = Does not drink 1 = Drinks more than once a month
smoke	0 = Not a current smoker 1 = A current smoker
excrcise_gr	0 = does not exercise 1 = exercises
BMI_gr	0 = BMI is in the range [0-23] 1 = BMI is in the range [23-25] 2 = BMI is in the range [25+]
sbp_gr	0 = sbp is in the range [0-130] 1 = sbp is in the range [130+]
wc_gr	0 = Not abdominal obesity 1 = Abdominal obesity
glu_gr	0 = Glucose is in the range [0-126] 1 = Glucose is in the range [126+]
HbA1c_gr	0 = HbA1c is in the range [0-6.5] 1 = HbA1c is in the range [6.5+]
stroke	0 = Not a stroke patient. 1 = A stroke patient.

As shown in Figure 1, the data to be used in this paper is biased toward people who have not had a stroke in terms of stroke, so it can be seen that the data set is unbalanced.

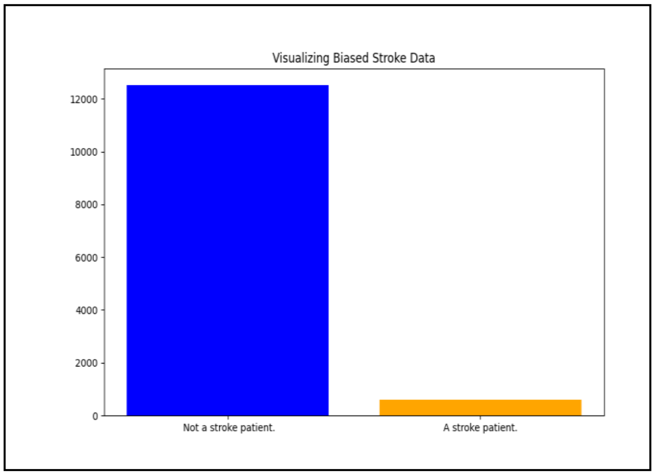

Figure 1: Visualizing Biased Storke Data

If the data is biased, accuracy may be good, but precision and recall may be low, resulting in poor results. There are two ways to solve this problem: oversampling and undersampling. Unlike oversampling, there is a problem with data loss in undersampling. In this paper, oversampling was used, and in order to prevent overfitting, it was divided into training data and test data 80:20 for training rather than the actual data set, and SMOTE among the oversampling techniques was used only on the training data. Figure 2 shows that SMOTE is used to resolve imbalance in the training data.

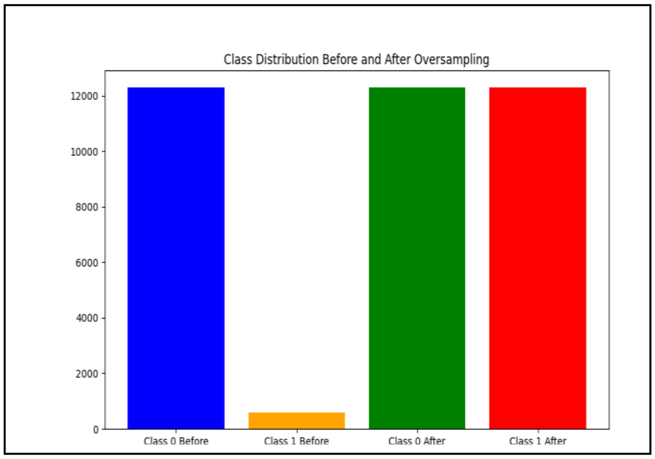

Figure 2: Before and after oversampling

4. Results

The first experiment is Logistic Regression, a probability model designed by Cox (1958) and a statistical model used when the dependent variable is binary (one of two types). Accuracy: 0.9065 Precision: 0.3156 Recall: 0.7346 F1 score: 0.4416 ROC AUC score: 0.9304 was obtained. Figure 3 visualizes the Logistic Regression model.

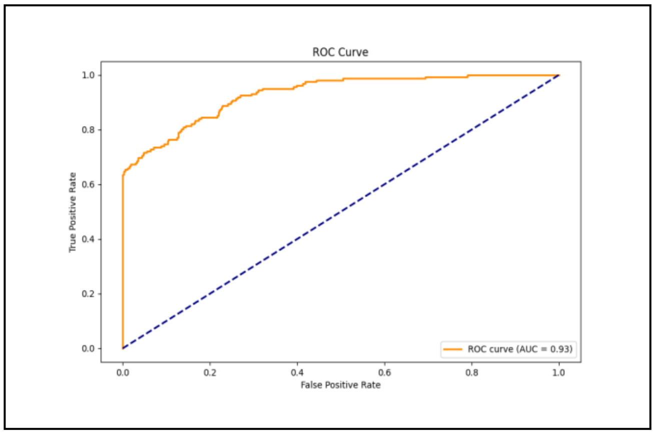

Figure 3: Logistic Regression

The second experiment is Random Forest, an algorithm that uses a decision tree as an ensemble learning method. Accuracy: 0.9795 Precision: 0.9444 Recall: 0.6296 F1 score: 0.7555 ROC AUC score: 0.9401 was obtained. Figure 4 is a visualization of the Decision Tree model.

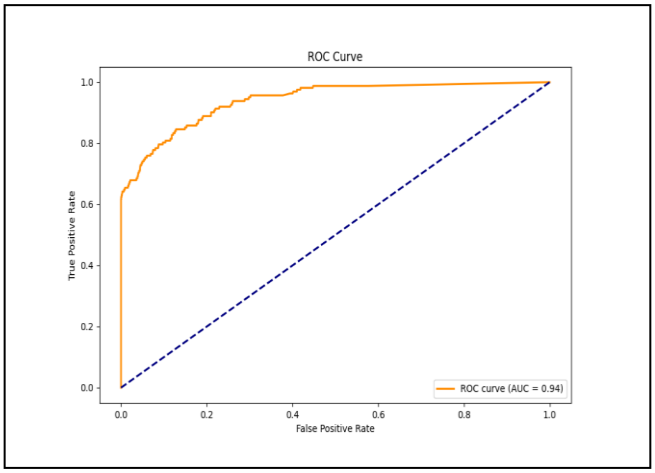

Figure 4: Random Forest

The third experiment is XGBoost, a tree boosting algorithm commonly used in data science and machine learning, and is based on gradient boosting technology. Through this, decision trees are sequentially combined to form a powerful prediction model, and regulation techniques are used to reduce model complexity and prevent overfitting. Accuracy: 0.9745 Precision: 0.8174 Recall: 0.6358 F1 score: 0.7153 ROC AUC score: 0.9066 was obtained. Figure 5 visualizes the XGBoost model.

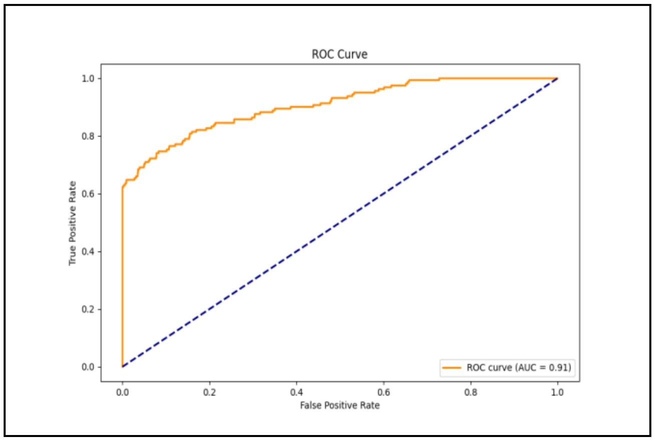

Figure 5: XGBoost

Table 4 is a table comparing the performance resulting from experiments between each model.

Table 4: Performance comparison for each model

Type	Logistic Regression	Random Forest	XGBoost
Accuracy	0.9065	0.9795	0.9745
Precision	0.3156	0.9444	0.8174
Recall	0.7346	0.6296	0.6358
F1 Score	0.7555	0.7555	0.7153
ROC AUCscore	0.9401	0.9401	0.9066

5. Conclusions

In this paper, a total of three experiments were performed: Logistic Regression, Random Forest, and XGBoost to select the optimal algorithm for predicting stroke onset. Comparisons can be made through evaluation indicators that indicate model performance, such as accuracy, precision, recall, F1 score, and ROC AUC score (AUC). First, in the case of AUC, the closer the model is to 1, the better the model's performance, and all models showed values over 0.9 and close to 1. Therefore, it can be seen that the performance of the models is suitable according to the above criteria. Because it is more important to accurately predict stroke patients first, recall was used as an evaluation index, and F1 score was used as an auxiliary index. Since it is not appropriate to evaluate models based on accuracy when the data is imbalanced, the F1 score was used as a criterion for evaluating only models with a score of 0.5 or higher to evaluate biased classification. First of all, in order of highest F1 score, Random Forest: 0.7555, The two models, in order of highest recall, were XGBoost: 0.6358 and Random Forest: 0.6296. To find the optimal algorithm for stroke prediction, we experimented with applying various machine learning models. As a result of comparing the F1 score and recall, logistic regression had the highest recall, but the F1 score did not meet the above criteria, so it was excluded. It can be seen that XGBoost has the highest recall among models that satisfy the above criteria. This shows that XGBoost is the optimal algorithm.

References

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Draelos, R. (2019, February 23). Measuring Performance: AUC (AUROC). *glassbox*. https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/

Korean Stoke Society. (2023). *Stroke treatment guidelines*

Korea Disease Control and Prevention Agency. (2022). *Raw data usage guidelines*.

Na, J.(2020). Stroke: "Early detection and prevention are important". *Biotime*. Retrieved from:

https://www.biotimes.co.kr/news/articleView.html?idxno=3915

Statistics Korea. (2022). *Cause of death statistics*