Analysis on the Risk Factors of Stroke

Janneese Palmer, Lucy Baptist, Joseph Coffey, Kate Conway

Summer Institute in Biostatistics and Data Science, 2023
National Institutes of Health (NIH)

Advisor: Dr. Lun-Ching Chang

Introduction

A stroke is a cardiovascular disease that occurs when the blood supply to part of the brain is blocked or when a blood vessel in the brain bursts. There are two types of strokes: ischemic and hemorrhagic. According to the Centers for Disease Control and Prevention, every year in the United States, over 795,000 people have a stroke. Stroke is a leading cause of serious long-term disability and death around the world (Centers for Disease Control and Prevention, 2023). Anyone at any age is susceptible to having a stroke; however, there are several factors that can increase one's chances of experiencing a stroke. Risk factors for strokes include high blood pressure, high cholesterol, heart disease, diabetes, obesity, sickle cell disease, tobacco use, age, etc. (Centers for Disease Control and Prevention, 2023).

In another study conducted using Korean data, researchers developed a logistic regression model aiming to improve stroke prevention strategies for high-risk patients. The model included variables such as age, BMI, cholesterol levels, hypertension, diabetes, smoking habits (status and intensity), physical activity, alcohol consumption, past medical history (hypertension and coronary heart disease), and family history (stroke and coronary heart disease) (Min et al., 2018).

Dataset

The dataset used for this project was taken from Kaggle: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset. We used this dataset to explore how various risk factors impact one's risk for stroke and how these factors interact with each other. The dataset includes 5110 subject's risk factors in terms of health and demographic information and whether they had a stroke.

The variables in this dataset are:

- **id**: unique identifier for each subject
- **gender**: male, female, or other
- **age**: 0-82 years old
- **hypertension**: 0 if the subject does not have hypertension, 1 if the subject has hypertension
- **heart_disease**: 0 if the subject does not have any heart disease, 1 if the subject has a heart disease
- **ever_married**: Yes or No
- **work_type**: children, govt_job, never_worked, private, or self-employed

- **Residence_type**: urban or rural
- **avg_glucose_level**: average glucose level in blood; range: 55-272
- **bmi**: body mass index; range: 10-98
- **smoking_status**: formerly smoked, never smoked, smokes, unknown
- **stroke**: 0 if subject never had a stroke, 1 if the subject suffered a stroke before

Statistical Analysis

The null hypothesis for the t tests and chi square tests was that there is no difference between the stroke and no stroke groups.

	Stroke **n=249 (.05)**	**No Stroke** **n=4861 (.95)**	**P Value**
Age	67.73 (SD 12.73)	41.97 (SD 22.29)	< 0.001
Male	108 (.43)	2007 (.41)	0.560
Ever Married (Yes)	220 (.88)	3133 (.64)	< 0.001
Work Type			< 0.001
Children	2 (.01)	685 (.14)	
Government	33 (.13)	624 (.13)	
Private	149 (.60)	2776 (.57)	
Self Employed	65 (.26)	754 (.16)	
Residence Type			0.298
Urban	135 (.54)	2461 (.51)	
Rural	114 (.46)	2400 (.49)	
Smoker Status			< 0.001

Never	90 (.36)	1802 (.37)	
Former	70 (.28)	815 (.17)	
Current	42 (.17)	747 (.15)	
BMI	30.47 (SD 6.33)	28.82 (SD 7.91)	< 0.001
Average Glucose Level	132.54 (SD 61.92)	104.8 (SD 43.85)	< 0.001
Hypertension Present	66 (.27)	432 (.09)	< 0.001
Heart Disease Present	47 (.19)	229 (.09)	< 0.001

Figure 1: Statistical Analysis of Variables

The above table gives an overview of the data and gives the p value for a t-test or a chi-square test that was used to compare the stroke and non-stroke groups. Of the sample population, 5% experienced stroke. It indicates that larger proportions of the stroke population than the non-stroke population were associated with greater age, being male, having been married, working in a private or self-employed setting, living in an urban area, being a current or former smoker, having heart disease or hypertension, having a higher average glucose level, and having a greater BMI. T-tests were used for the continuous variables (age, BMI, average glucose level) and chi square tests were used for categorical variables (all others). Hypothesis tests were conducted at the .05 significance level. According to the table, all variables except gender and residence type were significantly different between the groups at this level. Some variables are related, such as age with hypertension, heart disease, and marriage status, which can contribute to inaccurate inflation of the significance of these variables if one is associated with stroke.

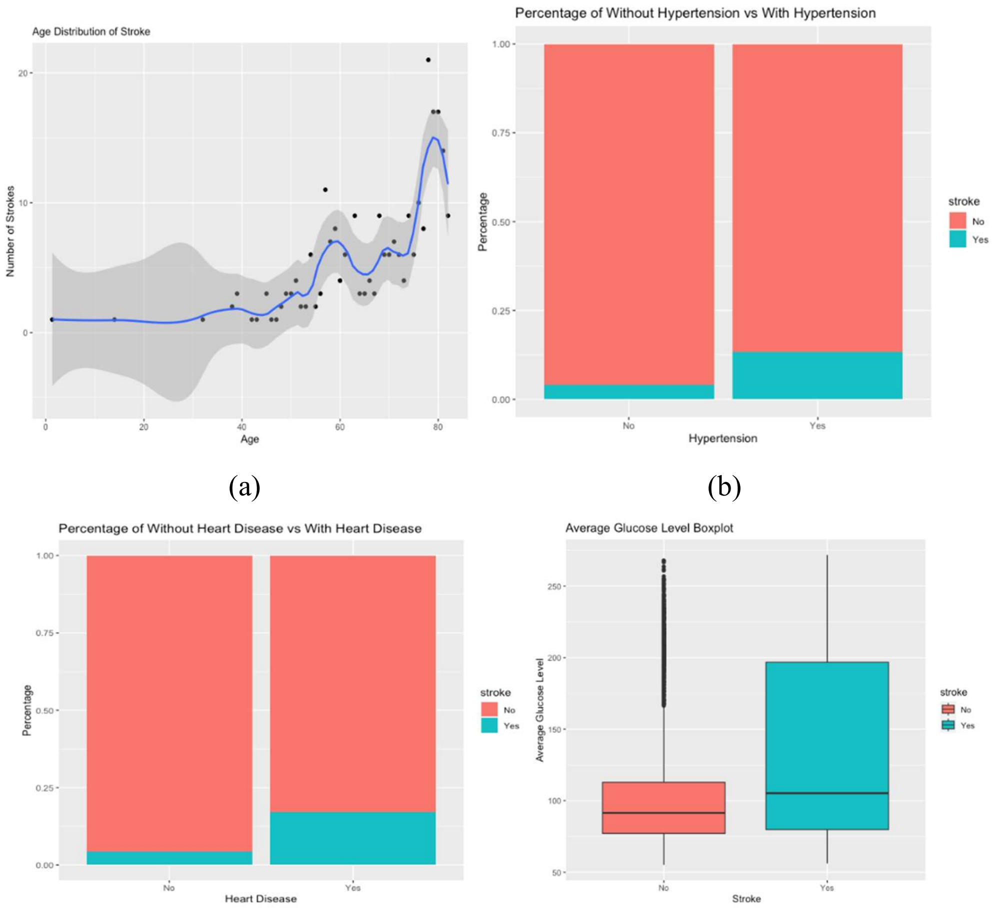
Logistic Regression Model

A logistic regression model was formed using all the variables in the table as predictors, as well as the interaction variables between age and gender, age and BMI, and age with heart disease. Using mixed stepwise regression (mixed, using AIC), a reduced model was created with the predictors age, BMI, hypertension, average glucose level, heart disease, age and BMI interaction, and age and heart disease interaction. The final model was created by selecting the predictors that were then significant at the .1 level (age, hypertension, heart disease, glucose). The coefficients and their p values are given below.

Variable	Coefficient	P Value
Age	0.069	< .001
Hypertension (Present)	0.381	0.019
Heart Disease (Present)	0.004	< 0.001
Average Glucose Level	0.330	0.079

Figure 2: Significant Predictors

These results indicate that all these variables in the table above were significant at the .1 level, and the first three (age, hypertension, and heart disease) were significant at the .05 level. Below are graphs illustrating the relationship between the main predictors and the occurrence of a stroke.



(a)

(b)

(c)

(d)

Figure 3: (a) Age Distribution of Stroke, (b) Percentage of Subjects With or Without Hypertension, (c) Percentage of Subjects With or Without Heart Disease, (d) Average Glucose Level to Stroke Boxplot

Stroke Prediction R Shiny Application

We developed a user-friendly application utilizing R Shiny, which allows users to select variables and obtain the probability of a stroke based on the chosen characteristics. The application comprises three pages: Visualization, Analysis, and Prediction. The Visualization page enables users to observe the impact of specific variable combinations on the occurrence of strokes. Through interactive visualizations, users can gain insights into the relationship between different factors and stroke risk. On the Analysis page, we present and analyze our final model alongside its predictors. This section provides a comprehensive overview of the model's performance and the significance of various predictors in determining the likelihood of a stroke. The Prediction page allows users to estimate the probability of a stroke using our finalized model. By inputting relevant variables, users can obtain personalized predictions tailored to their specific characteristics.

For the source code of the application, please refer to our GitHub repository, where you can access and explore the code in detail. https://github.com/l-baptist/FL_SIBDS23_Stroke_Project/blob/main/shinyappp

Limitations and Future Work

As indicated in the introduction, numerous factors contribute to an individual's susceptibility to stroke. However, in our dataset, we identified only four predictors that exhibited significance at the .1 level, and three predictors that demonstrated significance at the .05 level. Furthermore, we indicated the presence of two distinct stroke types, namely ischemic and hemorrhagic. However, the dataset lacks information regarding the specific stroke type experienced by the subjects who were identified as having had a stroke. In future endeavors, we aspire to work with a dataset that encompasses a broader range of high stroke risk factors. By doing so, we aim to construct an improved logistic model that enhances the accuracy of stroke prediction.

Conclusion

In this study, our objective was to construct a model centered around the high-risk factors associated with strokes. Our aim was to assist individuals in making strides towards predicting the likelihood of experiencing a stroke. By doing so, we hope to empower both individuals and medical practitioners to proactively undertake the essential measures for stroke prevention.

References

Centers for Disease Control and Prevention. (2023, May 4). *About stroke*. Centers for Disease Control and Prevention. https://www.cdc.gov/stroke/about.htm

Min, S. N., Park, S. J., Kim, D. J., Subramaniyam, M., & Lee, K.-S. (2018). Development of an algorithm for stroke prediction: A National Health Insurance Database Study in Korea. *European Neurology*, *79*(3–4), 214–220. https://doi.org/10.1159/000488366