JOURNAL OF
EMERGING INVESTIGATORS

Article

A statistical analysis and generalized linear models of cerebral stroke

Peter L. Jin[1], **Veera Holdai**[2]

[1] James M. Bennett High School, Salisbury, Maryland

[2] Department of Mathematical Sciences, Salisbury University, Salisbury, Maryland

SUMMARY

Cerebral stroke, a life-threatening condition that has a high mortality and morbidity rate, is the second leading cause of death worldwide. A stroke occurs when the blood supply to the brain is interrupted or reduced, resulting in potential neurological damage. Unlike many previous studies that focused on a single personal attribute related to stroke or estimated the probability of stroke, we conducted statistical analyses to investigate whether and how stroke and other variables are influenced together and amongst each other. Next, we modeled stroke, hypertension, and heart disease based on the data from 43,400 patients. We hypothesized that stroke, hypertension, and heart disease are statistically correlated to age, body mass index (BMI), glucose level, work type, marriage status, and gender, but are not related to residence type. Descriptive statistics of the data were computed and examined with statistical tests. To model the data, we performed logistic and linear regression. Our results showed that all of the variables in the dataset were related to each other except for residence type. Our data also suggested that heart disease had the strongest association with whether an individual had stroke, while smoking status had the strongest association with hypertension or heart disease. The analysis and models provide a context of risk factors with respect to cerebral stroke.

INTRODUCTION

Stroke is a significant public health issue affecting over 795,000 people each year in the United States with about 1 in 4 people having a stroke in their lifetime (1, 2). The five-year survival rate among ischemic stroke patients is 31%, while for stroke caused by intracerebral hemorrhage, the survival rate is even lower at 24% (3). Not only a detriment to societal health, stroke also has an immense negative impact on the economy with more than $56 billion in stroke-related costs in the United States in a single year, which includes health care services and absent work time (1).

Although the full complement of pathogenic factors associated with stroke are not completely known, stroke is known to be related to lifestyle choices, genetics, environmental exposures, and personal characteristics and attributes (1). Identifying and assessing the risk factors of stroke is necessary to aid future stroke prevention (1). Furthermore, establishing what mechanisms are unrelated to stroke is also important to reduce misdiagnosis and patient morbidity and mortality.

Previous studies have focused on estimating the probability of cerebral stroke or determining whether a single personal attribute is correlated with stroke (4, 5). However, they have not directly addressed how much stroke is connected to other variables with the effects of particular groupings and combinations of those variables. A specific study tracked individuals who participated in the Northern Manhattan Study (NOMAS), analyzing how diabetes duration status was related to stroke based on prospective population-based cohort data; however, the researchers did not address whether having diabetes along with groupings of other factors influenced the risk of stroke (6). Another study by Fekadu *et. al.* addressed stroke risk factors for patients in Ethiopia and Sub-Saharan Africa (7). That study was based on single characteristics including personal conditions and attributes, but not directly with social variables or multiple variables together (7).

In this study, we analyzed the relationships between stroke and other variables, and quantified how combinations of stroke, hypertension, and heart disease along with dataset features affect each other. We examined the probabilities of having stroke, hypertension, or heart disease based on multiple factors. We hypothesized that the prior conditions are dependent between themselves and significantly related to age, body mass index (BMI), glucose level, work type, marriage status, and gender, while being independent from residence type. We hypothesized that the prior conditions may not be related to residence type as their prevalence has been found to be similar between rural and urban patients in other studies (8). Overall, our findings suggest that stroke, hypertension, and heart disease are associated among themselves and have a relationship to age, BMI, glucose level, smoking status, gender (except with stroke), marriage status, and work type, while being unrelated to residence type. Future work could examine the potential interactions between additional variables to gain a more comprehensive understanding of their combined effects or explore machine learning techniques for enhanced stroke prediction.

RESULTS

Our rationale for this study was to clarify the impact of risk factors onto the chance of having a stroke. The data in this study was sourced from Mendeley Data and originally collected from HealthData.gov (8, 9). This dataset comprises of 43,400 patients with 12 features and with the target column being stroke. It has adequately large sample sizes. Some of the smallest sample sizes were that 200 patients had stroke and hypertension, and 583 patients had stroke but didn't have hypertension; 177 and 606 patients had stroke with and without heart disease, respectively, and 399 urban and 384 rural patients had stroke; all are greater than the widely-used minimum sample size of 30. We calculated descriptive

https://doi.org/10.59720/23-236

statistics of age, body mass index (BMI), and glucose level stratified by stroke, hypertension, and heart disease, residence type, and smoking status. Numerical statistics are being used as they have all been connected to stroke risk (10).

Descriptive Statistics

Mean age, BMI, and glucose level is lower for non-stroke patients as opposed to those who had a stroke, hinting that stroke could be related to those statistics. There is an especially large difference of approximately 26 years between the mean ages of stroke versus non-stroke patients (mean ± standard deviation of age in years: 68.14 ± 12.32 and 41.74 ± 22.39, respectively) (**Figure 1 and 2**). In quartile summaries of age, BMI, and glucose level by stroke, hypertension, heart disease, residence type, and smoking status, the median value was always greater for stroke, hypertension, or heart disease patients compared to patients without those conditions, suggesting that they may be related to each other (**Figure 1 and 3**). However, there was no clear difference in the median age, BMI, and glucose level between urban and rural patients, which supports our hypothesis that residence type is not related to the other factors of interest. Additionally, the median age, BMI, and glucose level for smoking status categories, from greatest to least, was always in the order of formerly smoked, currently smokes, and never smoked, suggesting that patients who never smoked may be the healthiest (in terms of BMI and glucose levels).

To measure variability in the data, we calculated interquartile range (IQR). The largest difference in IQR between stroke and non-stroke individuals occurred for glucose level, with the IQR's being 104.5 mg/dL (81.1 mg/dL to 185.6 mg/dL) and 34.2 mg/dL (77.5 mg/dL to 111.7 mg/dL), respectively, meaning that glucose level for stroke patients varied significantly unlike for non-stroke patients. However, the standard deviation of age and BMI is less for stroke patients, meaning that stroke patients may have less variability in age and BMI.

Overlaid histograms of age, BMI, and glucose level by stroke, hypertension, heart disease, residence type, and smoking status do not display a normal distribution in several cases (**Figure 2 and 3**). Most patients with stroke,

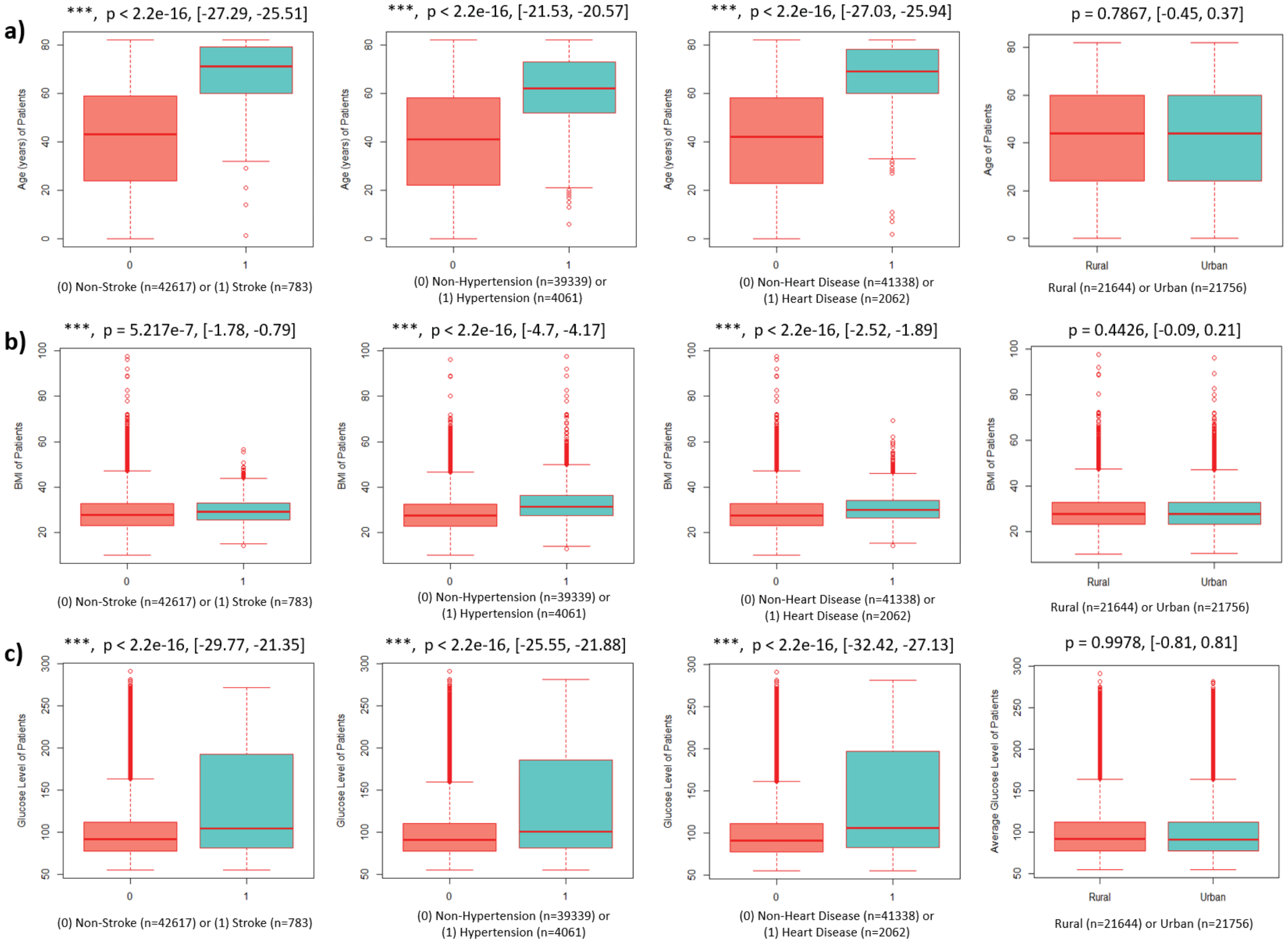

Figure 1: A greater mean age, BMI, and glucose level for stroke patients compared to non-stroke patients. Similar distributions of age, BMI, and glucose level between rural and urban patients. (a) Age, (b) BMI, and (c) glucose level by stroke, hypertension, heart disease, and residence type. The inner boxes display the median and interquartile range, and the whisker lines extending from the inner boxes signify the range of the values. The circular dots show the data outliers. Statistically significant differences between boxes are shown by ***, and the number of patients (n), *p*-values and 95% confidence intervals are listed (two-sample *t*-test, $p < 0.05$).

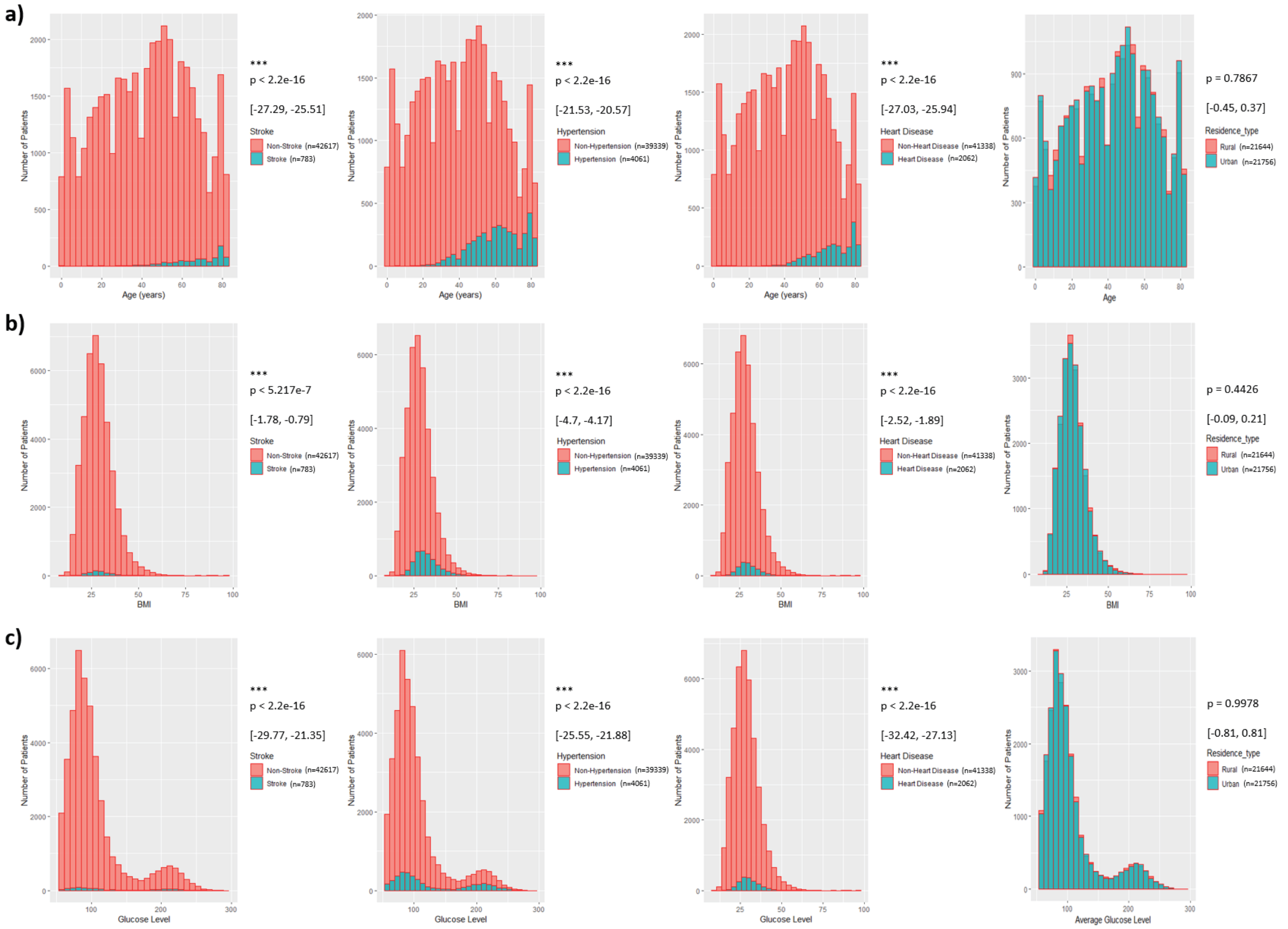

Figure 2: Left-skewed, right-skewed, and bimodal distributions of age, BMI, and glucose level, respectively. Similar distributions of age, BMI, and glucose level between rural and urban patients. (a) Age, (b) BMI, and (c) glucose level by stroke, hypertension, heart disease, and residence type. The bars in each histogram display the number of patients (n) and different colors signify the patient categories. Statistically significant differences in means between histograms are shown by ***, and *p*-values and 95% confidence intervals are listed (two-sample *t*-test, $p < 0.05$).

hypertension, or heart disease are older. For patients with stroke, 90.4% are older than 50 years. For hypertension and heart disease patients, the majority of individuals are also more than 50 years old (77.6% and 90.9%, respectively). All histograms of BMI are skewed right, suggesting that some individuals in the dataset have high BMI's. For example, the majority of patients with stroke (78.8%) are overweight, meaning that they have a BMI greater than 25 (11). All glucose level histograms are bimodal, meaning that there are two distinct groups. For example, the two groups of data for the glucose level of stroke patients are 70 mg/dL to 115 mg/dL and 200 mg/dL to 225 mg/dL. In general, this could suggest that there are two distinct types of individuals with respect to glucose level.

Statistical Comparisons

Our analyses revealed that there was a statistically significant difference in mean age, BMI, and glucose level between stroke and non-stroke, hypertension and non-hypertension, and heart disease and non-heart disease patients by two-sided *t*-tests and confidence intervals (all of $p < 2.2e\text{-}16$, except for BMI by stroke of $p = 5.217e\text{-}7$) (**Figure 1 and 2**). However, with age, BMI, and glucose level by residence type, the differences in means between rural and urban residents were not significant ($p = 0.787$, 0.443, 0.998, respectively), supporting our hypothesis that these factors are not related to residence type. To determine the size of any differences between means, we applied confidence intervals; they suggested that the mean age, for example, was approximately 22.5-27.3 years greater for stroke patients compared to non-stroke patients, while the mean age, BMI, and glucose level were around the same between residence types (**Figure 1 and 2**).

Combinations of variables such as hypertension, heart disease, and smoking status and with age, BMI, and glucose level may reflect on if and how these variables are related. Of these combinations, patients with hypertension and heart disease who have never smoked had the greatest stroke occurrence rate (10.3%), while those without hypertension or heart disease and never smoked had the lowest (1.2%) (**Table 1**). The greatest median age, BMI, and glucose levels were patients with hypertension and heart disease, with

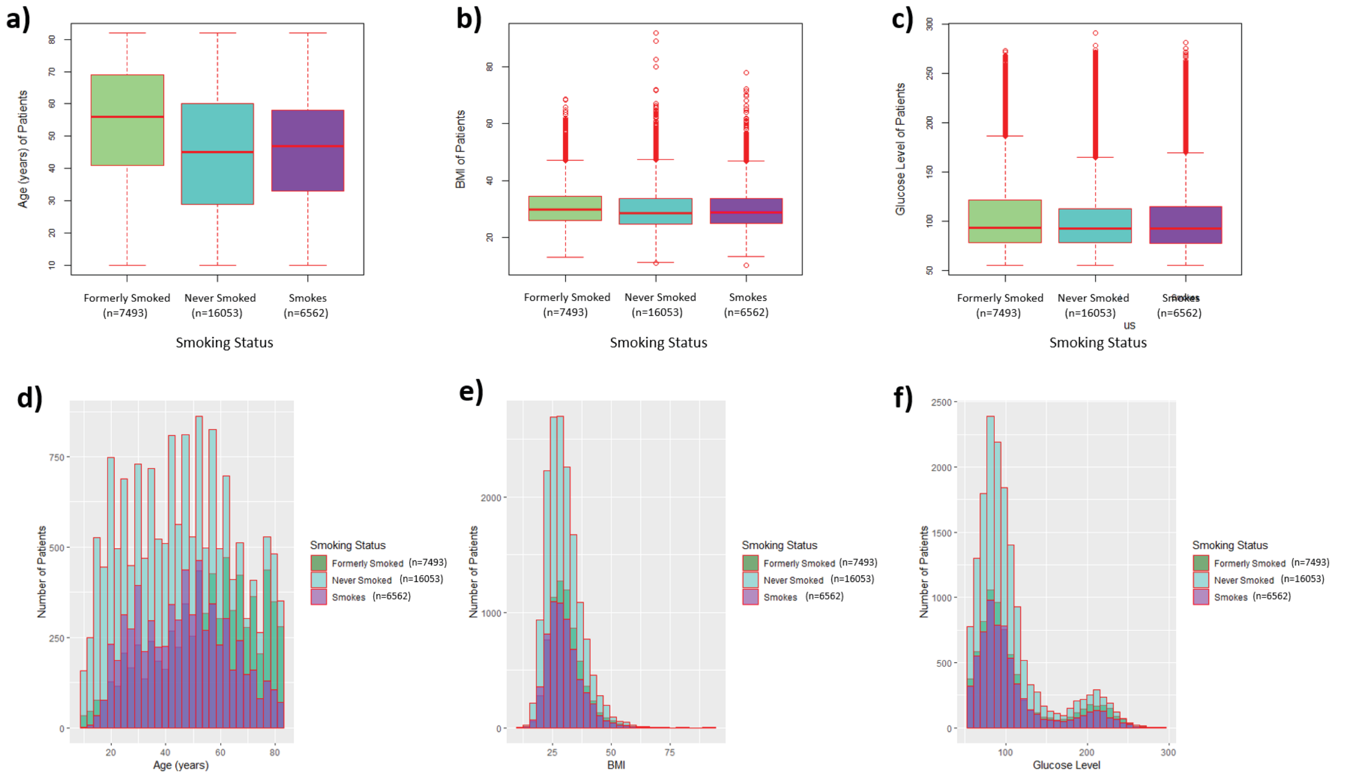

Figure 3: Mean age, BMI, and glucose level by smoking status, from greatest to least, are all in the order formerly smoked, smokes, and never smoked. The moderately symmetric, right skewed, and bimodal distributions of age, BMI, and glucose level, respectively, by smoking status. (a, d) Age, (b, e) BMI, and (c, f) glucose level by smoking status. In the boxplots, inner boxes display the median and interquartile range and whisker lines extending from the inner boxes show the range of the values. The circular dots are data outliers. The bars in each histogram display the number of patients (n) and different colors signify patient categories.

hypertension but without heart disease, and with hypertension and with heart disease, respectively. In addition, we calculated the percentages of patients with stroke, hypertension, and heart disease, based on work type, smoking status, and marriage status (**Table 2**)**.** Patients who are or were married had five times the stroke rate (2.5%) of those who were never married (0.5%). Hypertension was most prevalent in the self-employed (16.1%), and heart disease was most prevalent in patients who formerly smoked (8.6%) (**Table 2**). Between combinations of hypertension and heart disease, patients with both conditions had the highest occurrence of stroke (10.1%) (**Table 1**). This suggests that individuals with hypertension and heart disease have a higher risk for stroke, and around 1 in 10 of them may have a stroke. We found that patients with heart disease but without hypertension had nearly double the percentage of stroke (8.1%) than those with hypertension but without heart disease (4.2%), indicating that the former group may be at a higher risk for stroke than the latter. For patients with both hypertension and heart disease, those who never smoked had the highest occurrence of stroke (10.3%), compared to those who formerly smoked (8.9%) or currently smoke (9.2%), which is puzzling as smoking is known to be highly associated with stroke (12). This result may have been due to the severely imbalanced stroke variable in the dataset and data outliers. These differences in percentages by smoking status were also statistically significant but may not be practically significant.

Then, we investigated whether these variables were related to each other. We found that stroke was related to hypertension, heart disease, smoking status, marriage status, and work type ($p < 0.05$, Chi-Square Test of Independence) (**Table S1**). However, stroke was independent to residence type and gender ($p = 0.67$, 0.06, respectively; Chi-Square Test of Independence). It should be noted that the result with stroke and gender may be due to the outliers in the data and a severely imbalanced dataset. Overall, the tests suggested that residence type was unrelated to stroke, hypertension, and heart disease, while other variables tested were related, supporting our hypothesis. We then calculated their Cramer's *V* values (which go from 0 to 1, with larger values meaning a stronger relationship) to find out how strongly they were related (**Table S1**). Hypertension and heart disease with work type and marriage status were related the strongest (for example: hypertension & marriage status: 0.18; heart disease & work type: 0.12). However, these Cramer's *V* values are moderate overall, which is important because it suggests that these variables are not very strongly related and don't have a very large impact on each other.

To analyze how the numerical variables were related, we calculated Pearson's correlation coefficient. With age and BMI, age and glucose level, and BMI and glucose level, the Pearson's correlation coefficient was respectively 0.36, 0.24, and 0.19, suggesting that they were weakly associated. To find the association between numerical and binary features, we used Point-Biserial correlation. This was implemented for stroke, hypertension, and heart disease with age, BMI,

https://doi.org/10.59720/23-236

	Non-Heart Disease			Heart Disease		
Non-Hypertension	1.21% (n=458)			8.08% (n=125)		
	Smokes: 1.34% (n=74)	Formerly Smoked: 2.03% (n=121)	Never Smoked: 1.20% (n=167)	Smokes: 9.77% (n=30)	Formerly Smoked: 7.98% (n=39)	Never Smoked: 6.98% (n=30)
Hypertension	4.17% (n=148)			10.10% (n=52)		
	Smokes: 2.92% (n=18)	Formerly Smoked: 5.30% (n=47)	Never Smoked: 4.38% (n=69)	Smokes: 9.17% (n=11)	Formerly Smoked: 8.86% (n=14)	Never Smoked: 10.34% (n=18)

Table 1: The percentage and number (n) of patients with stroke for each category between hypertension, heart disease, and smoking status. For each pair of categories, the largest box contains the percentage and number of patients with stroke, and the three smaller boxes show the percentage and number with stroke for each smoking status.

		With Stroke	With Hypertension	With Heart Disease
Work Type	Children	0.03% (n=2)	0.03% (n=2)	0.06% (n=4)
	Government Job	1.64% (n=89)	10.90% (n=593)	4.61% (n=251)
	Never Worked	0.00% (n=0)	5.65% (n=1)	0.00% (n=0)
	Private	1.78% (n=441)	9.54% (n=2369)	4.63% (n=1151)
	Self-Employed	3.69% (n=251)	16.13% (n=1096)	9.66% (n=656)
Smoking Status	Formerly Smoked	2.95% (n=221)	13.93% (n=1044)	8.63% (n=647)
	Never Smoked	1.77% (n=284)	10.89% (n=1748)	3.76% (n=604)
	Smokes	2.03% (n=133)	11.22% (n=736)	6.51% (n=427)
Marriage Status	Married or Was Married	2.51% (n=703)	13.18% (n=3683)	6.79% (n=1897)
	Never Married	0.52% (n=80)	2.44% (n=378)	1.07% (n=165)

Table 2: The percentage and number (n) of patients with stroke, hypertension, and heart disease by work type, smoking status, and marriage status.

https://doi.org/10.59720/23-236

and glucose level. The greatest correlation coefficient was for hypertension and age (0.27), meaning that they were the most strongly correlated. However, 0.27 is a small correlation coefficient overall, suggesting that the variables are weakly related to each other. All of these Point-Biserial correlations had statistically significant p-values (α = 0.05).

Linear Regression Analyses

To measure how age, BMI, and glucose level influenced each other, we developed linear regression models. The models were between age and BMI, age and glucose level, and BMI and glucose level; the model between age and BMI had the highest adjusted R-squared value (0.13), which indicates how well the regression approximates the data. However, this is still low, which indicates that age and BMI do not have a strong relationship and cannot be used to predict one another. For this model in particular, the estimated coefficient was 0.12, meaning that for each one-year increase in age, there was a 0.12 increase in BMI. There was also a 0.45 mg/dL increase in glucose level for each year increase in age, and a 1.03 mg/dL increase in glucose level for a unit increase in BMI. The results of all the linear regression models performed were statistically significant with *p*-values less than 0.05.

Logistic Regression Analyses

We performed logistic regression to compare the effects of other variables on stroke, hypertension, and heart disease (**Table S2**). For the model of stroke based on age, BMI, glucose level, hypertension, heart disease, and smoking status, heart disease had the strongest coefficient (0.75), suggesting that heart disease has the largest impact on whether a patient had a stroke (but doesn't imply causation). This could be because heart disease can cause decreased blood flow in the arteries, leading to reduced blood in the brain (13). Glucose level and BMI had the weakest impact on stroke, and BMI was not a significant predictor of stroke. To measure the overall fit of this model, we calculated the Area under the Receiver Operating Characteristic curve (AUC), which was 0.86, meaning that the model was fairly robust. In logistic regression to model hypertension, smoking status had the strongest impact on hypertension (patient smokes: 0.88), which may be because smoking is known to decrease blood flow and increase blood pressure (14). Glucose level again had the least prediction impact (0.004). The model had an AUC value of 0.81, meaning that it had a fairly robust fit on hypertension. With the regression model of heart disease, smoking status again had the largest impact (0.61) on heart disease, while glucose level had the least (0.005). This model's AUC value was 0.87, suggesting that it was robust in modeling heart disease, and importantly shows that the model can be accurately used.

DISCUSSION

Our statistical analyses have shown that there are significant and non-significant relationships between features in the data. Our results suggest that patients who experience a stroke have a greater age, BMI, and glucose level than those who do not experience a stroke, confirming widely known clinical results (12). Patients who have hypertension or heart disease have greater mean age, BMI, and glucose level than those without. However, the BMI difference between hypertension and non-hypertension patients is much larger than for stroke or heart disease. Histogram distributions of the age of patients with stroke, hypertension, or heart disease are skewed left, which also suggests that the majority of individuals with those conditions are older. We found that there was no significant difference in the mean age, BMI, and glucose level between patients residing in rural versus urban areas, supporting our hypothesis that residence type is not related to age, BMI, and glucose level.

In the data, the self-employed have the highest occurrence of stroke, hypertension, and heart disease of all of the work types, suggesting that they may be at a higher risk for stroke. We speculate that this may be because the self-employed can have worse economic status and less access to healthcare. Patients who formerly smoked and those who are married may have a greater risk of stroke than non-smokers and unmarried patients. In addition, male patients are observed to have a greater rate of hypertension and heart disease than female patients, supporting scientific knowledge reported in literature (15, 16). We found that of the variables stroke, hypertension, heart disease, residence type, smoking status, gender, marriage status, and work type, most were related to each other by the Chi-Square test except for residence type and for stroke and gender. This again suggests that residence type is not associated with the other variables (residence type was not imbalanced). The result of non-association of stroke and gender may have been influenced since the dataset has outliers and is imbalanced in regard to stroke. The Cramer's *V* values were the greatest for hypertension and heart disease with marriage status and work type, suggesting that they were the most strongly associated. Importantly, however, these associations were moderate overall. With the correlation between pairs of numerical and binary variables, all had negligible correlations except for age and BMI, which had a low positive correlation of 0.36. This indicates that the variables being tested are weakly related to each other, in which when the value of one variable increases, the value of the other correlated variable may also increase. From linear regression, we also found that age, BMI, and glucose level have an impact on each other. Our results suggested that these variables are not enormously affected by each other.

We identified the variables with the greatest impact on stroke, hypertension, and heart disease by logistic regression. With respect to stroke, the heart disease coefficient was the greatest, suggesting that it had the most impact on whether an individual had a stroke. Hypertension was also an impactful stroke factor, while BMI and glucose level were not, which is puzzling as both BMI and glucose level have a known association with stroke (17, 18). We speculate that this result could be due to an imbalanced dataset for stroke and far outliers in terms of BMI. For hypertension, whether a patient smoked had the greatest influence on it. However, glucose level had the lowest impact on hypertension. This is odd given the well-known strong association between the two variables (19). With respect to heart disease, smoking and hypertension were the most influential factors, while glucose level was the least influential. This could be due to the prior reasons of dataset imbalance and outliers.

Some factors may have influenced our statistical analysis. There was a severely imbalanced stroke population in that there are many more patients with stroke than those without, which may have influenced the Chi-Square results regarding stroke and gender. This imbalanced dataset could be addressed by under-sampling the non-stroke patients, over-sampling the stroke patients, or using a data weighting mechanism. The dataset included far outliers, although they accounted for a small (< 5% for BMI) part of the data. For example, several patients had a BMI of greater than 90, which is humanly possible but very unlikely. This could have affected the logistic regression models. Our results in terms of stroke and gender may have been different if we had access to an outlier-free and non-imbalanced dataset. In the future, we could remove these outliers from the data entirely before analysis; we would still have enough data, since outliers only make up a small section of the dataset.

As the data has missing values (31% of smoking status and 3% of BMI), a future extension of this research could be to attempt to have the full data from the source, remove the missing values entirely, or impute the missing values into the dataset. Removing these missing values could be problematic, however, since they make up a significant portion of the data. We could also develop machine learning models to predict the occurrence or occurrence time of stroke, hypertension, and heart disease based on the dataset. Research by Dritsas et. al. used nine machine learning models to predict stroke based on influencing factors, for example (20). Complex machine learning models such as neural networks, random forests, and the Naïve Bayes would be best to predict and model these intricate prior conditions (21). The models would be significant to patients in general, as we could make use of these predictions to screen individuals and improve prevention strategies to decrease patient morbidity and mortality.

The findings of this research suggest that stroke, hypertension, and heart disease are associated among themselves and have a relationship to age, BMI, glucose level, smoking status, gender (except with stroke), marriage status, and work type, while being unrelated to residence type. We also quantified the impact of these variables amongst themselves. These results could be used to assist with stroke, hypertension, and heart disease screening techniques, or be incorporated into machine learning prediction models. For example, our results on the relevance of risk factors to stroke could be used to see whether patients have a high chance of developing the disease.

Features	Values
Patient ID	1-72,943
Gender	Male/Female/Other
Age	0.08-82
Hypertension	1 (Yes) / 0 (No)
Heart Disease	1 (Yes) / 0 (No)
Marriage Status	Yes (Married) / No (Not Married)
Work Type	Children/Private/Never Worked/Self-Employed/Government Job
Residence Type	Urban/Rural
Glucose Level	55-291.05
BMI	10.1-97.6
Smoking Status	Smokes/Formerly Smoked/Never Smoked
Stroke	1 (Yes) / 0 (No)

Table 3: The dataset features and their values. For numerical features, the minimum and maximum data values are displayed. For categorical features, all categories are listed.

MATERIALS AND METHODS

Dataset

This research was based on a dataset of 43,400 patients sourced from Mendeley Data and originally collected from HealthData.gov (8, 9). Independent Review Board approval was not required. The dataset has 12 features with only 783 stroke patients out of the total 43,400 and 31% of smoking status and 3% of BMI values missing. Missing values from the data were not excluded for our analyses. It contains patient data but no personal information (**Table 3**).

Data Analysis

The data was analyzed using the statistical programming language R (**Appendix**) (22). Descriptive statistics of variables and correlation coefficients in the data were calculated and then visualized with boxplots and histograms. The boxplots were created on the same frame to allow accurate comparison of the numerical data values for age, BMI, and glucose level. Histograms were used to investigate the overall distribution of age, BMI, and glucose level, and were overlaid for stroke, hypertension, heart disease, residence type, and smoking status. The histograms were created in R using the ggplot2 library (23).

One-sided and two-sided T-tests were used to analyze the differences in sample means of the age, BMI, and glucose level for stroke and non-stroke, hypertension and non-hypertension, heart disease and non-heart disease, and rural and urban patients for statistical significance. The widely accepted significance level of α = 0.05 was used to determine whether there was a statistically significant difference in means.

The Chi-Square Test of Independence was performed between pairs of the categorical variables stroke, hypertension, heart disease, residence type, smoking status, gender, work type, and marriage status. This was to determine whether the variables being tested were independent or dependent. The Chi-Square Test was extended with Cramer's V values, which determines the degree of dependency between the tested variables, using the Rcompanion library (24).

Linear regression was performed to quantify the relationship between age, BMI, and glucose level. To study the impact and influence of dataset variables on and between stroke, hypertension, and heart disease, logistic regression was implemented. The libraries pscl, Metrics, and caret were utilized for logistic regression (25, 26, 27). The robustness of these models was then calculated with the AUC value using the pROC library (28).

Received: September 12, 2023
Accepted: March 8, 2024
Published: December 8, 2024

REFERENCES

1. Tsao, Connie W., *et. al.* "Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association." *Circulation,* vol. 147, 2023. https://doi.org/10.1161/CIR.0000000000001123.
2. Jackson, Geoff, and Chari, Karishma. "National Hospital Care Survey Demonstration Projects: Stroke Inpatient Hospitalizations." *National Health Statistics Reports*, vol. 132, 2023, pp. 1-11.
3. Smajlovic, Dzevdet, *et. al.* "Five-year survival-after first-

https://doi.org/10.59720/23-236

ever stroke." *Bosnian Journal of Basic Medical Sciences*, vol. 6, no. 3, 2006, pp. 17-22. https://doi.org/10.17305/bjbms.2006.3138.

4. Zhao, Dong, *et. al.* "Epidemiological transition of stroke in China: twenty-one-year observational study from the Sino-MONICA-Beijing Project." *Stroke*, vol. 39, no. 6, 2008, pp. 1668-74. https://doi.org/10.1161/strokeaha.107.502807.
5. Yiin, Gabriel S. C., *et. al.* "Age-specific incidence, outcome, cost and projected future burden of atrial fibrillation-related embolic vascular events: a population-based study." *Circulation*, vol. 130, no. 15, 2014, pp. 1236-44. https://doi.org/10.1161/circulationaha.114.010942.
6. Banerjee, Chirantan, *et. al.* "Duration of Diabetes and Risk of Ischemic Stroke: The Northern Manhattan Study." *Stroke*, vol. 43, no. 5, 2012, pp. 1212. https://doi.org/10.1161%2FSTROKEAHA.111.641381.
7. Fekadu, Ginenus, *et. al.* "Risk factors, clinical presentations and predictors of stroke among adult patients admitted to stroke unit of Jimma University Medical Center, South West Ethiopia: prospective observational study." *BMC Neurology*, vol. 19, no. 1, 7 Aug. 2019, pp.187. https://doi.org/10.1186/s12883-019-1409-0.
8. Liu, Tianyu, *et. al.* "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset." *Artificial Intelligence in Medicine*, vol. 101, no. 723, 2019. https://doi.org/10.1016/j.artmed.2019.101723.
9. Liu, Tianyu, *et. al.* "Data for: A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical-datasets." *Mendeley Data*, V1, 2019. data.mendeley.com/datasets/x8ygrw87jw/1.
10. Murphy, Stephen J., and Werring, David J. "Stroke: Causes and clinical features." *Medicine (Abingdon, England: UK ed.)*, vol. 48, no. 9, 2020, pp. 561-566. https://doi.org/10.1016%2Fj.mpmed.2020.06.002.
11. Nuttall, Frank Q. "Body Mass Index: Obesity, BMI, and Health: A Critical Review." *Nutrition Today*, vol. 50, no. 3, 2015, pp. 117-128. https://doi.org/10.1097/nt.0000000000000092.
12. Kelly-Hayes, Margaret. "Influence of Age and Health Behaviors on Stroke Risk: Lessons from Longitudinal Studies." *Journal of the American Geriatrics Society*, vol. 58, pp. 325-328. https://doi.org/10.1111/j.1532-5415.2010.02915.x.
13. Arboix, Adria. "Cardiovascular risk factors for acute stroke: Risk profiles in the different subtypes of ischemic stroke." *World Journal of Clinical Cases*, vol. 3, no. 5, 2015, pp. 418-429. https://doi.org/10.12998/wjcc.v3.i5.418.
14. Shah, Reena S., and Cole, John W. "Smoking and Stroke: the more you smoke the more you stroke." *Expert Review of Cardiovascular Therapy*, vol. 8, no. 7, 2010, pp. 917-932. https://doi.org/10.1586/erc.10.56.
15. Everett, Bethany, and Zajacova, Anna. "Gender Differences in Hypertension and Hypertension Awareness Among Young Adults." *Biodemography Social Biology*, vol. 61, no. 1, 2015, pp. 1-17. https://doi.org/10.1080/19485565.2014.929488.
16. Weidner, Gerdi. "Why do men get more heart disease than women? An international perspective." *Journal of American College Health*, vol. 48, no. 6, 2000, pp. 291-4. https://doi.org/10.1080/07448480009596270.
17. Shiozawa, Masahiro, *et. al.* "Association of Body Mass Index with Ischemic and Hemorrhagic Stroke." *Nutrients*, vol. 13, no. 7, 9 July 2021. https://doi.org/10.3390/nu13072343.
18. Chen, Rong, *et al.* "Diabetes and Stroke: Epidemiology, Pathophysiology, Pharmaceuticals and Outcomes." *The American Journal of Medical Sciences*, vol. 351, no. 4, 2016, pp. 380-386. https://doi.org/10.1016/j.amjms.2016.01.011.
19. Yan, Qun, *et. al.* "Association of blood glucose level and hypertension in Elderly Chinese Subjects: A community-based study." *BMC Endocrine Disorders*, vol. 16, no. 1, 2016. https://doi.org/10.1186/s12902-016-0119-5.
20. Dritsas, Elias, and Trigka, Maria. "Stroke Risk Prediction with Machine Learning Techniques." *Sensors (Basel, Switzerland)*, vol. 22, no. 13, 21 Jun. 2022, pp. 4670. https://doi.org/10.3390/s22134670.
21. Sarker, Iqbal H. "Machine Learning: Algorithms, Real-World Applications and Research Directions." *SN Computer Science*, vol. 2, no. 3, 2021. https://doi.org/10.1007/s42979-021-00592-x.
22. R Core Team. "A Language and Environment for Statistical Computing. R Foundation for Statistical Computing." 2017. www.R-project.org.
23. Wickham, Hadley. "ggplot2: Create Elegant Data Visualizations Using the Grammar of Graphics." R Package Version 3.4.2, 2023. https://cran.r-project.org/package=ggplot2.
24. Mangiafico, Salvatore. "Rcompanion: Functions to Support Extension Education Program Evaluation." R Package Version 2.4.30, 2023. https://cran.r-project.org/package=rcompanion.
25. Jackson, Simon. "pscl: Political Science Computational Laboratory." R Package Version 1.5.5.1, 2023. https://cran.r-project.org/package=pscl.
26. Hamner, Ben, *et. al.* "Metrics: Evaluation Metrics for Machine Learning." R Package Version 0.1.4, 2023. https://cran.r-project.org/package=Metrics.
27. Kuhn, Max, *et. al.* "caret: Classification and Regression Training." R Package Version 6.0-94, 2023. https://cran.r-project.org/package=caret/.
28. Robin, Xavier, *et. al.* "pROC: Display and Analyze ROC Curves." R Package Version 1.18.4, 2023. https://cran.r-project.org/package=pROC.

APPENDIX

	Stroke		Hypertension		Heart Disease	
	p-value	Cramer’s V	p-value	Cramer’s V	p-value	Cramer’s V
Hypertension	2.2e-16	0.0753	n/a		2.2e-16	0.1198
Heart Disease	2.2e-16	0.1138	2.2e-16	0.1198	n/a	
Residence Type	0.6657	0.0022	0.5259	0.0031	0.5831	0.0027
Smoking Status	2.2e-16	0.0469	2.2e-16	0.1274	2.2e-16	0.0997
Gender	0.0559	0.0115	2.104e-6	0.0245	2.2e-16	0.0824
Marriage Status	2.2e-16	0.0719	2.2e-16	0.1766	2.2e-16	0.1288
Work Type	2.2e-16	0.0759	2.2e-16	0.1542	2.2e-16	0.1242

Table S1: Comparison of the dependency between pairs of the variables stroke, hypertension, heart disease, residence type, smoking status, gender, marriage status, and work type through the Chi-Square test. For each pair of variables, the p-value and Cramer’s V score are listed. A significance level of α = 0.05 was used with the p-value.

	Coefficient			Standard Error			p-value		
	Stroke	Hypertension	Heart Disease	Stroke	Hypertension	Heart Disease	Stroke	Hypertension	Heart Disease
Age	0.0714	0.05	0.0768	0.0039	0.0015	0.0024	2e-16	2e-16	2e-16
BMI	-0.0033	0.0532	0.0129	0.0075	0.0029	0.0045	0.6656	2e-16	0.0045
Glucose Level	0.0034	0.0036	0.0049	0.0009	0.0004	0.0005	6.92e-5	2e-16	2e-16
Hypertension	0.4723	n/a	0.3967	0.1102	n/a	0.0735	1.81e-5	n/a	6.69e-8
Heart Disease	0.7466	0.3028	n/a	0.1195	0.0745	n/a	4.15e-10	4.77e-5	n/a
Smoking Status: Formerly Smoked	0.0578	0.6507	0.3285	0.1518	0.078	0.0933	0.7035	2e-16	0.0004
Smoking Status: Never Smoked	0.1474	0.784	-0.1389	0.1411	0.072	0.0928	0.296	2e-16	0.1343
Smoking Status: Smokes	0.3367	0.8769	0.6101	0.1673	0.0824	0.1023	0.0441	2e-16	2.5e-9

Table S2: The coefficient, standard error, and p-value for each variable in logistic regression for stroke, hypertension, and heart disease. The coefficient column signifies the impact of the variable for the model, the standard error column shows the model’s standard error for the variable, and the p-value column suggests the statistical significance of the variable.

#R Code implemented in this research

```
#T-test
only <- strokes[strokes$Residence_type > 0, ]
nononly <- strokes[strokes$Residence_type < 1, ]
(testt <- t.test(only$age, mu=mean(nononly$age), alternative="greater"))
#For BMI
only <- strokes[strokes$stroke > 0, ]
nononly <- strokes[strokes$stroke < 1, ]
(testt <- t.test(only$bmi, mu=mean(nononly$bmi, na.rm=TRUE), alternative="greater"))

#Chi-Square test
chisq.test(strokes$stroke, strokes$work_type)

#Simple Linear Regression
linearreg <- lm(strokes$avg_glucose_level ~ strokes$age)
summary(linearreg)

#Histograms
strokes$heart_disease <- as.character(strokes$heart_disease)
library("ggplot2")
(histogram <- ggplot(strokes, aes(x=avg_glucose_level,fill=heart_disease)) +
   geom_histogram(position="identity", alpha=0.8, colour='red'))
histogram + labs(#title = "Glucose Level by Residence Type",
             x = "Glucose Level",
             y = "Number of Patients",
             fill = "Heart Disease") +
  scale_fill_discrete(labels = c("Non-Heart Disease", "Heart Disease"))

#Boxplots
strokes$age <- as.numeric(strokes$age)
par(cex.main = 0.9)
boxplot(strokes$age ~ strokes$heart_disease,
      #main="Age with and without Stroke",
        xlab="Non-Heart Disease (0) or Heart Disease (1)",
        ylab="Age (years) of Patients",
        border="Red",
        col = c("salmon", "turquoise"))

#Point-Biserial Correlation
cor.test(strokes$heart_disease, strokes$avg_glucose_level)

#Logistic Regression
sample <- sample(c(TRUE,FALSE), nrow(strokes),
                 replace=TRUE, prob=c(0.7,0.3))
train_dataset <- strokes[sample, ]
test_dataset <- strokes[!sample, ]
strokes$smoking_status <- factor(strokes$smoking_status)
logregmodel <- glm(hypertension ~ age + bmi + heart_disease + smoking_status, data =
train_dataset, family="binomial")
summary(logregmodel)
pscl::pR2(logregmodel)["McFadden"]
predicted <- predict(logregmodel, test_dataset, type= "response")
pROC::auc(test_dataset$hypertension, predicted)

#Cramer's V
rcompanion::cramerV(strokes$hypertension, strokes$gender)
```