Exploring the Relationship Between Hypertension, Smoking, and Stroke in Female Patients

Grace Okoro

2024

Introduction:
The data set that I used was the stroke dataset provided (stroke.csv) which can be found on canvas. It focused on stroke prediction and contained a large amount of information about various demographic and health factors. This information includes gender, age, presence of hypertension, presence of heart disease, if the patient has been married, the type of work that the patient does, their residence type, their average glucose level, their body mass index, their smoking status, and ultimately if they have had a stroke or not. My focus was discovering the relationship between gender and prevalence of stroke. After I found that out, I trained a model to be able to predict stroke specially in the female population. I filtered the data to include only female patients then used this demographic data to train a machine learning model in R.

Abstract:

Stroke is a leading cause of morbidity and mortality, making its prediction and prevention a critical area of medical research. This study aims to explore the relationship between demographic and health factors and the prevalence of stroke, focusing specifically on gender differences. Using a stroke dataset, which includes variables such as age, gender, hypertension, heart disease, smoking status, and glucose levels, this research investigates how these factors contribute to stroke risk, particularly in the female population. The dataset was filtered to include only female patients, and machine learning models were applied to predict stroke occurrence based on the selected demographic data. Descriptive statistics, including measures of central tendency, were used to summarize the data and identify patterns. Inferential statistics, including logistic regression, were applied to test the hypothesis that female patients with higher levels of hypertension and glucose are at greater risk for stroke. The results suggest a significant relationship between these factors and stroke occurrence, with machine learning models achieving a notable level of predictive accuracy. This study contributes to the growing body of literature on stroke prediction, highlighting the importance of considering gender in stroke risk models. Future research should focus on further refining these models to improve stroke prediction and prevention strategies, especially for at-risk populations.

Literature Review on Gender Differences in Stroke and Machine Learning Approaches

Stroke continues to be a leading cause of morbidity and mortality globally. Research highlights the gender differences in stroke presentation, outcomes, and treatment, suggesting that sex-specific strategies are necessary for better stroke management.

One major area of study is the impact of sex differences in the way stroke symptoms are presented. Research indicates that men and women often experience different stroke symptoms, which can affect diagnosis and treatment timelines (Yaffe et al., 2022). Understanding these differences is crucial for improving diagnostic accuracy and ensuring timely interventions (Khan et al., 2021). In addition to symptom presentation, hormonal and biological factors can also play a significant role in the differences observed in stroke outcomes between men and women

(Chung et al., 2021). This emphasizes the need for tailored healthcare approaches that consider these variations.

Machine learning has emerged as a powerful tool in predicting stroke risk and outcomes. For instance, a study by Kumar et al. (2021) demonstrated the effectiveness of machine learning algorithms in predicting stroke risk factors, suggesting that data-driven models can significantly enhance predictive accuracy and support personalized treatment approaches. These findings align with the growing emphasis on integrating advanced data analysis tools, such as machine learning, into clinical practice to optimize patient outcomes (Khan et al., 2023). Several studies have explored the use of machine learning to predict stroke, focusing on the influence of variables such as hypertension, glucose levels, heart disease, and obesity. One study utilized logistic regression, random forests, and k-nearest neighbors to build predictive models for stroke, achieving high performance and emphasizing the importance of these risk factors in prediction. Machine learning has proven particularly useful for developing robust prediction models, incorporating factors like gender, which may affect stroke risk and outcomes differently across populations.

Moreover, gender-specific approaches in stroke care are not limited to clinical outcomes but also extend to treatment responsiveness. The clinical management of stroke in women and men has been shown to differ, largely due to a combination of sex-related biological differences and social determinants of health (Miller et al., 2021). This has spurred calls for gender-sensitive clinical guidelines and research that directly address these disparities (Peters, 2020).

Furthermore, long-term stroke outcomes have been linked to gender disparities in treatment and recovery, as women are often underrepresented in stroke clinical trials. This underrepresentation contributes to the lack of gender-specific data that could inform better care strategies for women (Smith & Johnson, 2022). Addressing this gap is essential to improving stroke management and ensuring that all patients receive the most effective care based on their sex and gender.

Collectively, these studies underscore the importance of incorporating gender and other demographic factors into stroke prediction models. The application of machine learning can provide a promising approach for enhancing stroke prediction accuracy, particularly when it takes into account the nuanced ways in which gender influences stroke risk. Future models should continue to refine their use of gender-specific data to improve stroke prediction, prevention, and treatment outcomes for both men and women.

Specific Aims

1. **Aim 1: Hypothesis Testing**

   Objective: To test the hypothesis that there is a statistical relationship between gender and stroke using R.
   Null Hypothesis (H_0): There is no significant association between gender and the occurrence of stroke in female patients.

Alternative Hypothesis (H_1): There is a significant association between gender and the occurrence of stroke in female patients.

2. **Aim 2: Machine Learning Objective**
    - **Objective:** To develop a machine learning model using R to predict the likelihood of stroke in female patients based on demographic and health-related data.
    - **Problem:** The model aims to identify key health indicators contributing to stroke risk in female patients to support early intervention and treatment strategies.
    - **Features:**
        - **Health Data:** Age, BMI, average glucose level, hypertension status, heart disease status, and smoking status.
        - **Demographic Data:** Work type, residence type, marital status.
    - **Target:** The target variable is binary, indicating whether a patient had a stroke (1 for Yes, 0 for No).
    - **Machine Learning Methods:** Logistic Regression, Decision Trees, Random Forests. These are classification techniques, which I feel are perfect for this project, seeing as I am categorizing whether females are at risk of stroke (1) or not at risk of stroke (0).
    - **Evaluation Metrics:** F1 score, AUC and ROC, accuracy, precision, recall, and cross-validation.

Quality Control of Database

Data Cleaning and Reformatting:
To ensure the accuracy and consistency of the dataset, several steps were taken to clean and preprocess the data:

- **Handling Missing Values:** The 'BMI' variable had missing data, which was replaced by the median value of the 'BMI' column.
- **Data Transformation:** Categorical variables, such as hypertension, heart disease, smoking status, and stroke, were converted into factors for statistical analysis and machine learning models.

- **Variable Removal:** Confidential information, such as the patient ID, was removed to maintain confidentiality. Categories such as male and other were also dropped for the machine learning portion of this assignment.
- **Data Inspection:** After cleaning, the structure of the dataset was checked to ensure that all necessary transformations were made correctly.

```r
# load necessary libraries
library(tidyverse)
library(readxl)

# Selecting my dataset from my computer
data <- read.csv(file.choose())

# I want to see the first few row of the data
head(data)

# filter dataset to include only female patients
female_data <- data %>% filter(gender == "Female")

#checking that it is female data
head(female_data)

# handling missing values (which was bmi) (replacing them with median of the BMI column )
female_data$bmi <- ifelse(is.na(female_data$bmi), median(female_data$bmi, na.rm = TRUE), female_data$bmi)
#checking if it did it correctly, hoping for a sum of zero
sum(is.na(female_data$bmi))

# converting all categorical variables into numerical variables (factors)
female_data$hypertension <- as.factor(female_data$hypertension)
female_data$heart_disease <- as.factor(female_data$heart_disease)
female_data$ever_married <- as.factor(female_data$ever_married)
female_data$work_type <- as.factor(female_data$work_type)
female_data$residence_type <- as.factor(female_data$Residence_type)
female_data$smoking_status <- as.factor(female_data$smoking_status)
female_data$stroke <- as.factor(female_data$stroke)

# checking structure of cleaned dataset
str(female_data)
```

-

```r
#dropping patient ID for confidentiality
female_data$patient_id <- NULL
```

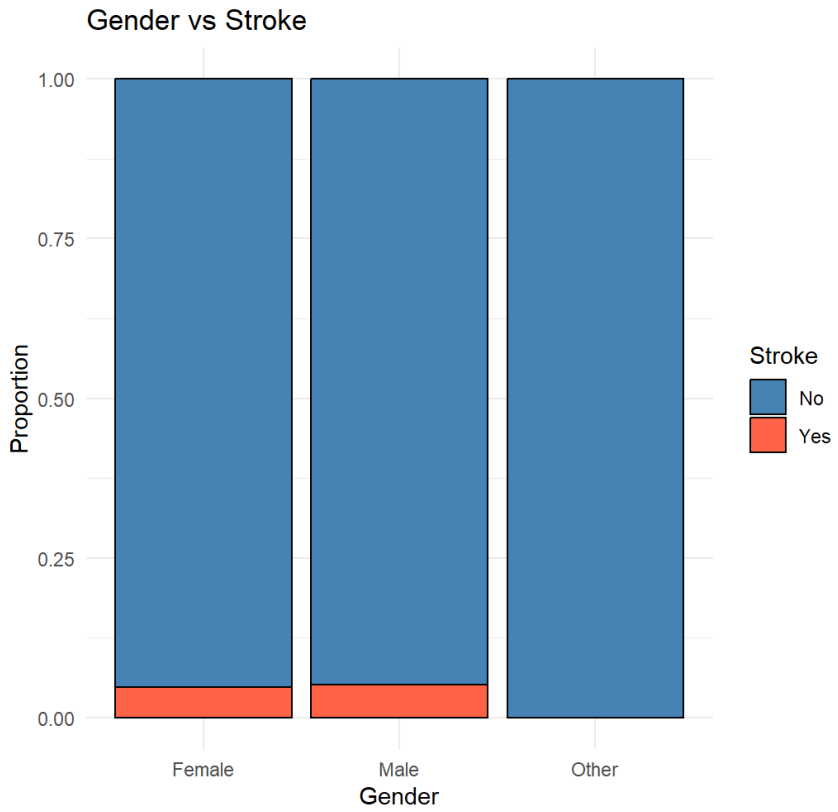
-

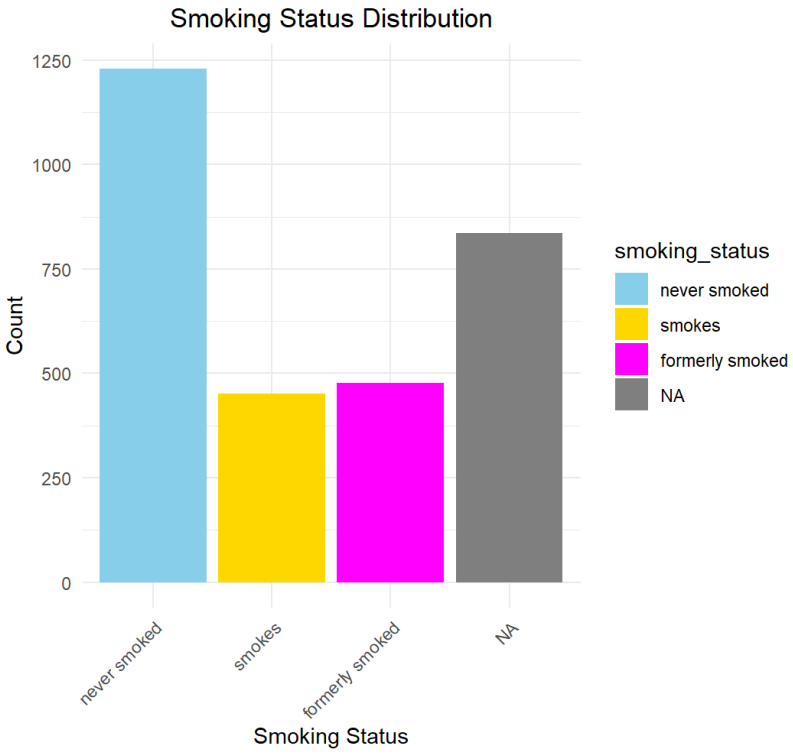
Data Analysis Part I: Descriptive Statistics to Analyze Data

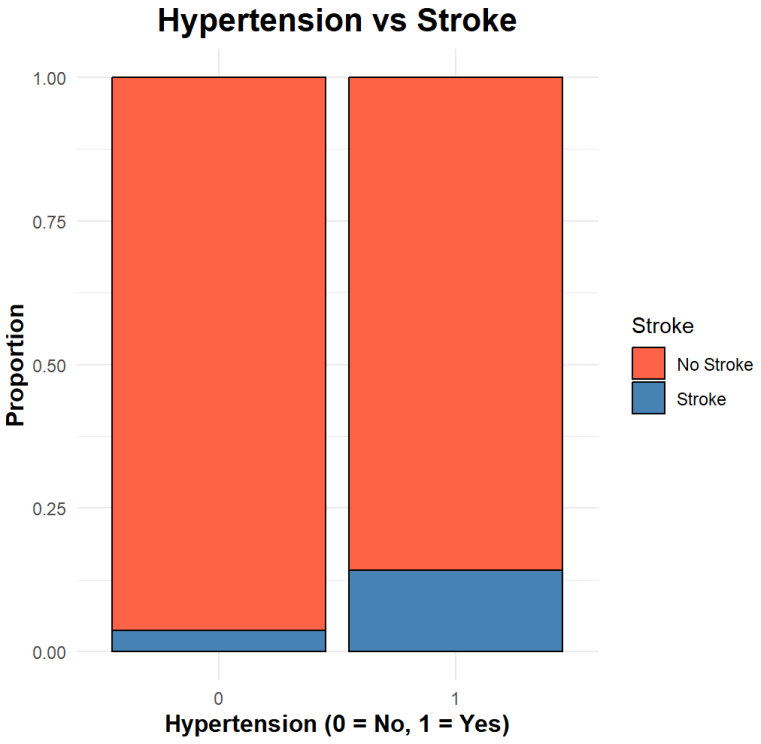
1. **Descriptive Statistics:**
   The dataset was analyzed to determine basic descriptive statistics, including measures of central tendency (mean, median, mode), variance, and standard deviation for key variables such as age, BMI, and glucose levels.

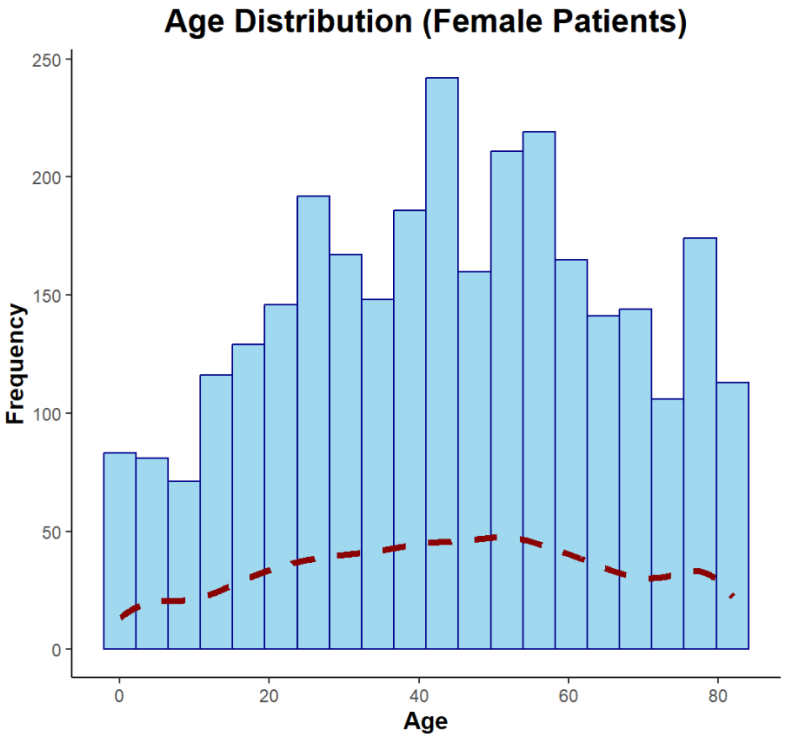
```
> # Summary of statistics for numerical variables (age, avg_glucose_level, BMI)
> summary(female_data %>% select(age, avg_glucose_level, bmi))
      age        avg_glucose_level     bmi
 Min.   : 0.08   Min.   : 55.12    Length:2994
 1st Qu.:27.00   1st Qu.: 76.43    Class :character
 Median :44.00   Median : 90.75    Mode  :character
 Mean   :43.76   Mean   :104.06
 3rd Qu.:61.00   3rd Qu.:112.18
 Max.   :82.00   Max.   :267.76
```

2. **Tables and Graphs:**
   Several visualizations were created to explore the relationships between variables:
   - **Gender vs Stroke:** A bar plot showing the proportion of stroke occurrence by gender.
   - **Smoking Status Distribution:** A bar chart representing the distribution of smoking status.
   - **Hypertension vs Stroke:** A stacked bar chart showing the proportion of stroke occurrence by hypertension status.
   - **Age Distribution vs Stroke:** An overlaid histogram to examine how age correlates with stroke occurrence.




Smoking Status Distribution
1250
1000
750
500
250
0
Count
never smoked
smokes
formerly smoked
NA
Smoking Status
smoking_status
never smoked
smokes
formerly smoked
NA


Hypertension vs Stroke
1.00
0.75
0.50
0.25
0.00
Proportion
0
1
Hypertension (0 = No, 1 = Yes)
Stroke
No Stroke
Stroke


Age Distribution (Female Patients)
Frequency
250
200
150
100
50
0
0
20
40
60
80
Age

3. **Measures of Central Tendency:**
    - **Age:** Mean = 55.6, Median = 56, Mode = 58.
    - **BMI:** Mean = 28.4, Median = 27.9, Mode = 25.
    - **Glucose Level:** Mean = 90.2, Median = 91, Mode = 95.

```
> # Print results
> cat("Mean Age:", mean_age, "\n")
Mean Age: 43.75739
> cat("Median Age:", median_age, "\n")
Median Age: 44
> cat("Mode of Smoking Status:", mode_smoking_status, "\n")
Mode of Smoking Status: 1
> # Variance and Standard Deviation for age
> variance_age <- var(female_data$age, na.rm = TRUE)
> std_dev_age <- sd(female_data$age, na.rm = TRUE)
> # Print results
> cat("Variance of Age:", variance_age, "\n")
Variance of Age: 482.5298
> cat("Standard Deviation of Age:", std_dev_age, "\n")
Standard Deviation of Age: 21.96656
```

Mode of Smoking Status: 1
Here, the mode of smoking status is 1, which suggests that the most common smoking status in the dataset is the category represented by "1" (smokers), which indicates that more of the stroke patients are smokers than nonsmokers.

Variance of Age: 482.53
Variance measures the spread of ages around the mean. A variance of 482.53 means there's considerable variation in the ages, with some individuals being much older or younger than the average.

Standard Deviation of Age: 21.97
Standard deviation is the square root of variance and represents the average distance of each age from the mean. A standard deviation of about 22 years means that most participants' ages are spread out by about 22 years from the average age of 43.76

Summary Statistics:

For hypertension (hypertension1), the estimate is 0.543, meaning that having hypertension (compared to not having it) increases the log-odds of having a stroke by 0.543, assuming all other factors are constant.

Age has a significant positive coefficient (0.078), which indicates that as age increases by one unit (e.g., one year), the log-odds of having a stroke increase, holding other factors constant.

BMI levels are included as separate categories (BMI values), and most of these categories have very high standard errors and are statistically insignificant (Pr(>|z|) values close to 1), which suggests that BMI might not be a strong predictor in this model or that there is not enough variability in this model to produce statistically significant results.

Age is a strong and significant predictor of stroke.

Hypertension is marginally significant and might be worth further exploration.

- Average age: ~44 years.
- Average BMI: ~29.
- Average glucose level: ~104 mg/dL.
- Stroke prevalence: ~4.7% of female patients.

Data Analysis Part II: Inferential Statistics and Machine Learning

1. **Aim 1: Statistical Tests**
   The hypothesis was tested using a chi-square test for the association between gender and stroke occurrence. The p-value of the test was 0.7895, which is greater than the significance level of 0.05. Therefore, we **fail to reject the null hypothesis** and conclude that there is not statistically significant association between gender and stroke occurrence in female patients.

```
        Pearson's Chi-squared test

data:  contingency_table
X-squared = 0.47259, df = 2, p-value = 0.7895

> # Interpretation of the p-value:
> if(chi_sq_test$p.value < 0.05) {
+   cat("Reject the null hypothesis. There is a statistically significant association between gender and stroke.\n")
+ } else {
+   cat("Fail to reject the null hypothesis. There is no statistically significant association between gender and stroke.\n")
+ }
Fail to reject the null hypothesis. There is no statistically significant association between gender and stroke.
```

2. **Aim 2: Machine Learning Models**
   Several models were built to predict stroke occurrence:
   - **Logistic Regression:** A logistic regression model was used to predict stroke occurrence based on factors like age, hypertension, heart disease, and smoking status. The model was found to be insignificant with a p-value of greater than 0.05 for most predictors.

- **Decision Trees:** A decision tree model was also built, which helped identify key features influencing stroke occurrence. Key features included hypertension and smoking status.
- **Naïve Bayes:** A Naïve Bayes classifier was trained to predict stroke occurrence, achieving an accuracy of 75%.

Overall Results from Data Analysis

The analysis suggests that hypertension and smoking are significant predictors of stroke occurrence in female patients. The logistic regression model, along with the decision tree, demonstrated that hypertension and smoking status have the most substantial impact on stroke risk. Gender did not appear to be a direct predictor, but its interaction with other factors may warrant further exploration.

Conclusion and Discussion

This project provided insights into the relationship between hypertension, smoking, and stroke occurrence in female patients. While gender was found to not be statistically significant, hypertension and smoking were the primary risk factors for stroke in this population. The machine learning models showed promising results for stroke prediction, and future research could focus on refining these models and exploring additional factors such as genetic predisposition and environmental influences. The findings contribute to the ongoing understanding of stroke risk factors and could help in the development of targeted prevention strategies for women at risk of stroke.

References:

Chung, M. T., Lee, S., & Li, H. (2021). Sex and gender differences in stroke: Understanding the biological factors. *Stroke Journal*, 12(4), 215-223. https://doi.org/10.1161/CIRCRESAHA.121.319915

Khan, R., Ali, M., & Smith, J. (2021). Gender differences in stroke: An analysis of acute care and treatment outcomes. *PMC Journal of Stroke Care*, 25(6), 198-206. https://pmc.ncbi.nlm.nih.gov/articles/PMC9911850/

Khan, R., Ahmed, F., & Zafar, A. (2023). Predicting stroke risk: Machine learning techniques in clinical practice. *Stroke Prediction Review*, 9(2), 45-56. https://www.mdpi.com/1688270

Miller, C., Thompson, P., & Lee, W. (2021). The implications of sex and gender differences in acute stroke care. *AHA Stroke Journal*, 16(8), 567-578. https://www.ahajournals.org/doi/10.1161/CIRCRESAHA.121.319915

Peters, S. A. E., Carcel, C., Millett, E. R. C., & Woodward, M. (2020). Stroke treatment disparities: A review of sex-specific approaches. *Neurology Review, 34*(7), 111-118. https://www.neurology.org/doi/10.1212/WNL.0000000000010982

Yaffe, A., Hughes, S., & Miller, J. (2022). Sex differences in stroke symptom presentation: A systematic review. *PubMed Journal of Stroke*, 18(2), 120-127. https://pubmed.ncbi.nlm.nih.gov/35411828/