Association analysis of EHR stroke data report

Yutian (Margery) Liu

12/05/2022

Introduction

Every year, more than 795,000 people in the United States have a stroke[1]. And with the improvement of technology, a lot of data is stored as electronic health records, which allows us to analyze these clinical features. In this project, I will use the medical records from the article 'A predictive analytics approach for stroke prediction using machine learning and neural networks[2] Paper. **The goal is to find whether there is an association between average glucose level and whether they have had a stroke, and whether there is any relationship between intermediate glucose level and age level or obesity level.**

Questions to be Addressed:

I want to explore the association between average glucose level value and stroke in this study. I am also wondering if there is any relationship between age, BMI, and average glucose level. It can be stepped down as the following question

1. What is the density of average glucose level and age among stroke and non-stroke patients?
2. What is the distribution of average glucose levels among stroke and non-stroke patients in different age levels and obesity levels?
3. Are there any significant differences in mean blood glucose levels between stroke and non-stroke patients at each age level and each level of obesity?

Methods

Dataset Information

Stroke EHR data was downloaded from github provided in the paper, you can download it from my website link to dataset

First, I will clean the data and show some preliminary results through summary tables; then, you will visualize these results. Then, a significant test Wilcoxon test will be performed to analyze whether There is no statistically significant difference between people who suffered a stroke and people who did not suffer a stroke in this sample d

Data Clean

After installing and loading necessary libraries, load the data set from the website, then data early check and data clean are performed. Data cleaning includes the following steps:

1. The ID column was removed since it is not a clinical feature
2. Subjects with Gender = "Other" were removed from the data
3. Check the percentage of missing values, if it is acceptable, impute missing values by the `mice` library
4. All categorical variables were re-coded as labelled factors & changed to more clear names
5. Add a category column of BMI named `obesity_level`
6. Add a category column of BMI named `age_level`

Data Imputation

```
sum(is.na(stroke))/nrow(stroke)
```

```
## [1] 0.03934234
```

There is only 3% missing values which means imputation is acceptable, `mice` libraries was used to fo imputation.

```
sum(is.na(completestroke))
```

```
## [1] 0
```

After using `mice` library to impute the dataset, there is not any missing value.

Preliminary Results

Preliminary Results: Summary Table

Summary tables are saved on the TableTables page. There are three tables, one summarizing average glucose level based on stroke status, one summarizing average glucose level based on age level and the other summarizing average glucose level based on obesity level.

In each table, there are 5 columns, including category, count number, the mean value of glucose level, the standard deviation of average glucose level and the proportion. After installing and loading necessary libraries, load the data set from the website, then data early check and data clean are performed. Data cleaning includes the following steps:

Table 1

Summary of average glucose level based on stroke status

```
## # A tibble: 2 x 5
##   stroke_t  count mean.glucose sd.glucose proportion
##   <fct>     <int>        <dbl>      <dbl>      <dbl>
## 1 no stroke  4860         105.       43.8     0.951
## 2 stroke      249         133.       61.9     0.0487
```

It shows there are 4860 stroke patients and 249 patients did not suffer from stroke, people who suffered a stroke have a higher mean average glucose level and standard deviation.

Table 2

Summary of average glucose level based on age level

```
## # A tibble: 4 x 5
##   age_level count mean.glucose sd.glucose proportion
##   <fct>     <int>        <dbl>      <dbl>      <dbl>
## 1 New Born     43         95.1       27.6    0.00842
## 2 Child       813         94.3       27.1    0.159
## 3 Adult      2949        103.        41.0    0.577
## 4 Elderly    1304        122.        58.1    0.255
```

This table displays that most of these objects are elderly and adults and they have the highest and the second highest mean average glucose level, at 122.045(mg/dl) and 102.519(mg/dl) respectively.

Table 3

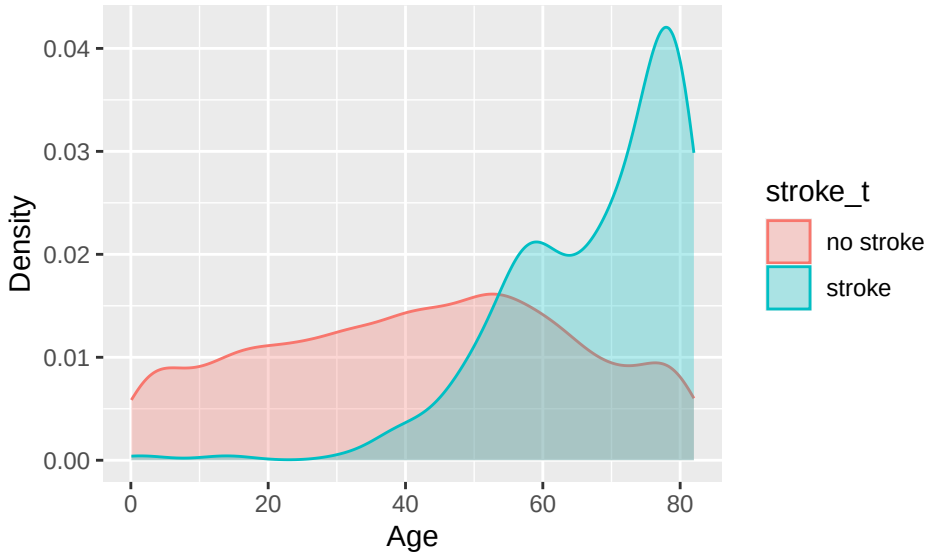
Summary of average glucose level based on obesity level

```
## # A tibble: 4 x 5
##   obesity_level count mean.glucose sd.glucose proportion
##   <fct>         <int>        <dbl>      <dbl>      <dbl>
## 1 Underweight      16         94.9       30.1    0.00313
## 2 Normal          391         95.8       33.6    0.0765
## 3 Overweight      132         94.1       26.6    0.0258
## 4 Obese          4570        107.        46.5    0.894
```

This table is summarizing the average glucose levels based on obesity level, while about 89.45% of objects in this dataset are obese people, they also have the highest mean and standard deviation average glucose levels compared to others.

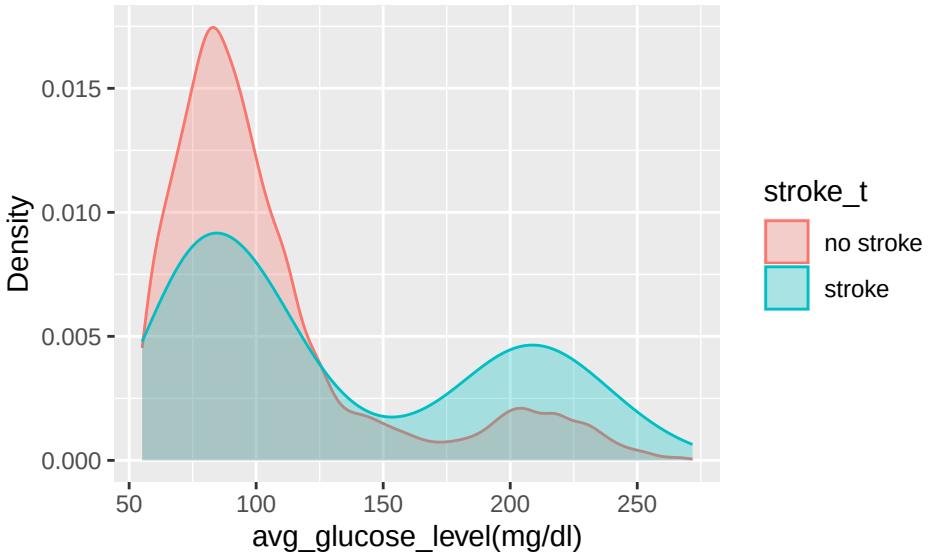
Preliminary Results: Density plot and Boxplot to show distribution

Age Density plot based on Stroke Status



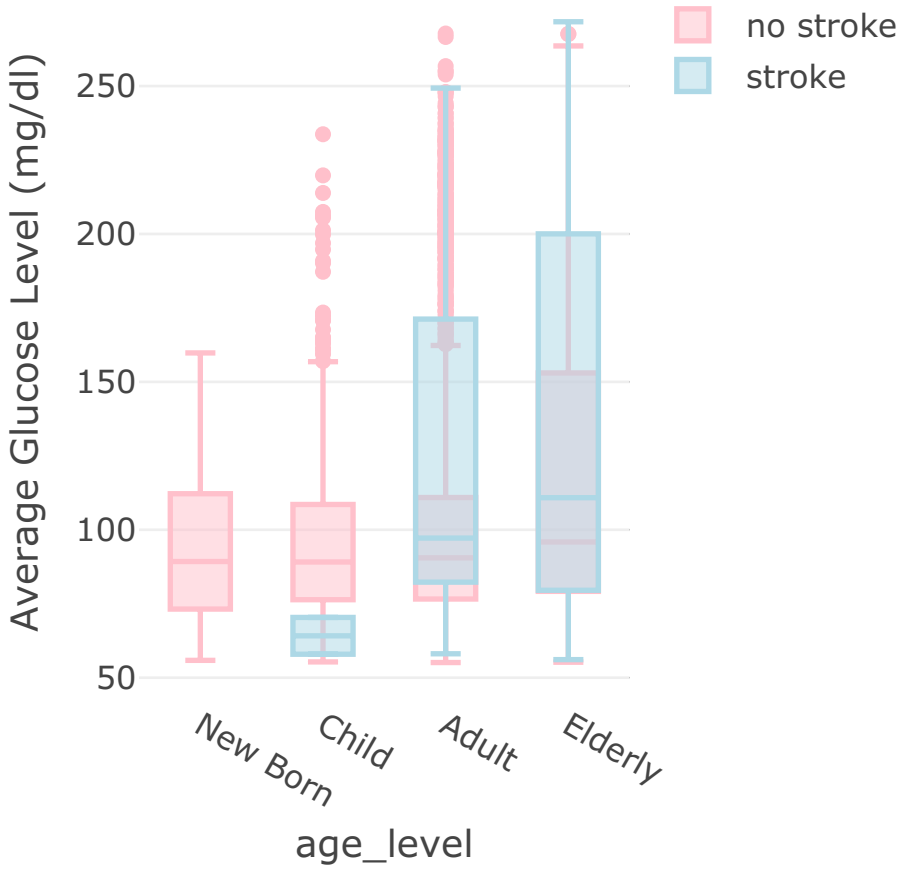
This plot is a density plot that displays the age distribution among stroke patients and non-stroke patients. In this figure, it can be seen that stroke begins when people are about 30 years old, and the density increases considerably steadily while when it comes to patients above 70 years old, the density increases dramatically and reaches a peak at approximately 79 years old.

Average glucose level Density plot based on Stroke Status



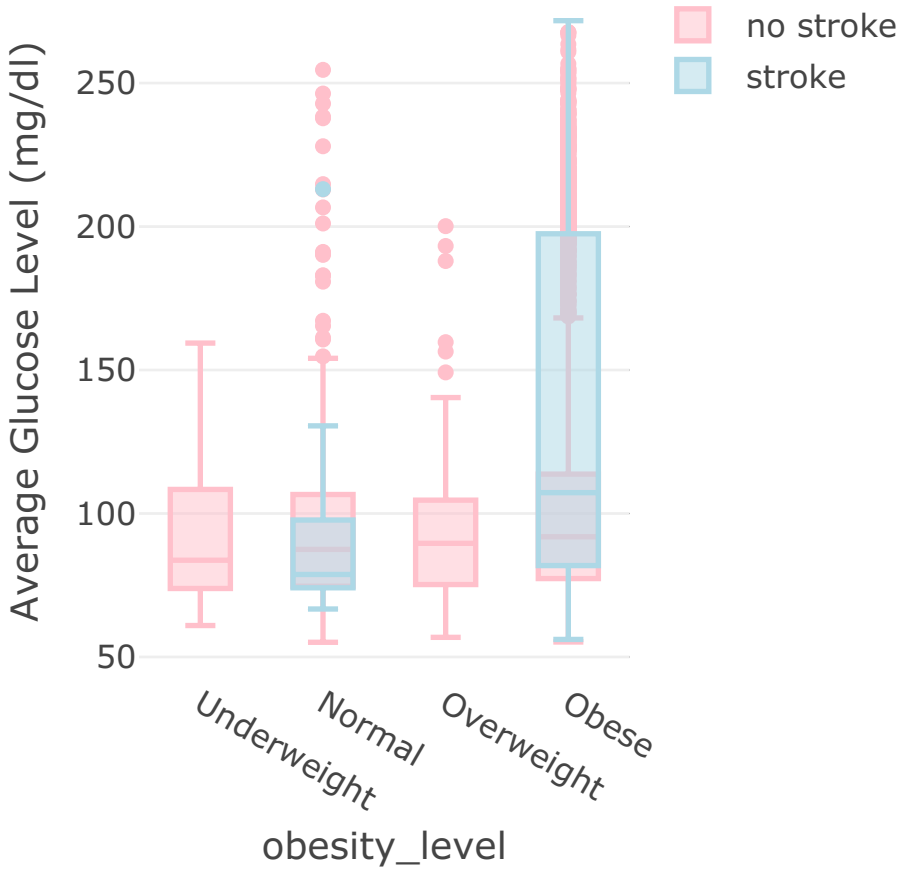
This density plot shows the density of average glucose level(mg/dl) based on stroke status, it can be seen that the average glucose level of most of the subjects who have not suffered a stroke was less than 150mg/dl while the density of people whose average glucose level above 150mg/dl is higher among stroke patients.

Distribution of average glucose level based on age level



This is a box plot with average glucose level(mg/dl) as y axis,age level as x axis and colored by stroke status.In this figure, elderly and adult stroke patient have the highest avg_glucose_level values, at 110.85 (mg/dl) and 97.2 (mg/dl) respectively.And the Q3 value reached to 200 when it comes to elderly stroke patients.

Distribution of average glucose level based on obesity level



This box plot displays the average glucose level distribution based on obesity level and stroke status. While there are no underweight and overweight stroke patients, obese stroke patients' median average glucose level is 107.26(mg/dl) and others are all below 100 (mg/dl).

Signigicant Test

Wilcoxon test will be performed to compare if there is significant test among two groups[3]. Results are stored at "Wilcoxon Test" page

From the density plots, it can be seen that `avg_glucose_level` is not normally distribution, therefore, Wilcoxon test will be performed to find out whether the difference in two groups is significant.

The Wilcoxon test, is a nonparametric statistical test that compares two paired groups. The tests essentially calculate the difference between sets of pairs and analyze these differences to establish if they are statistically significantly different from one another.

H0: There is no statistical significant difference between people suffered stroke and people did not suffer stroke in this sample data

H1: There is no statistical significant difference between people suffered stroke and people did not suffer stroke in this sample data

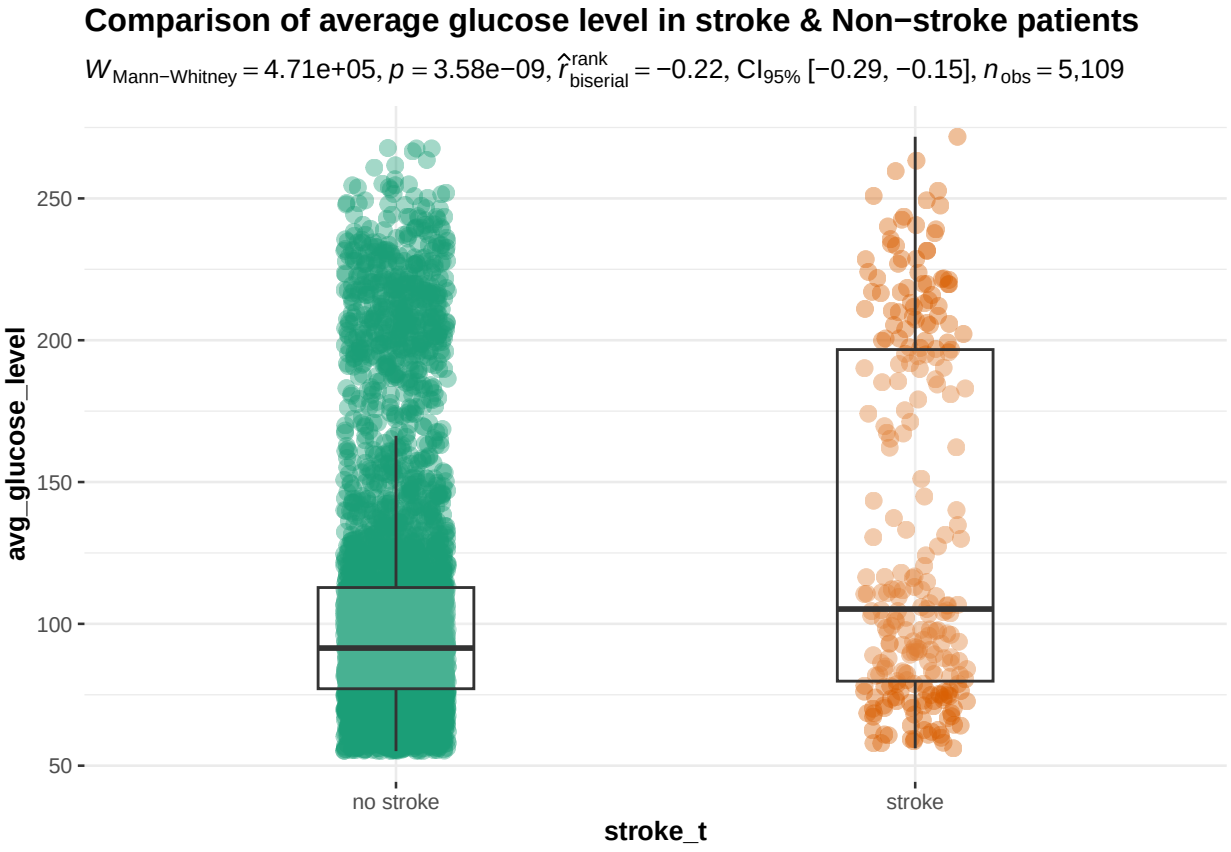
If the p-value is less than 0.05, we reject the null hypothesis that there's no difference between the means and conclude that a significant difference does exist

```
test <- wilcox.test(stroke$avg_glucose_level ~stroke$stroke_t)
test
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  stroke$avg_glucose_level by stroke$stroke_t
## W = 471082, p-value = 4e-09
## alternative hypothesis: true location shift is not equal to 0
```

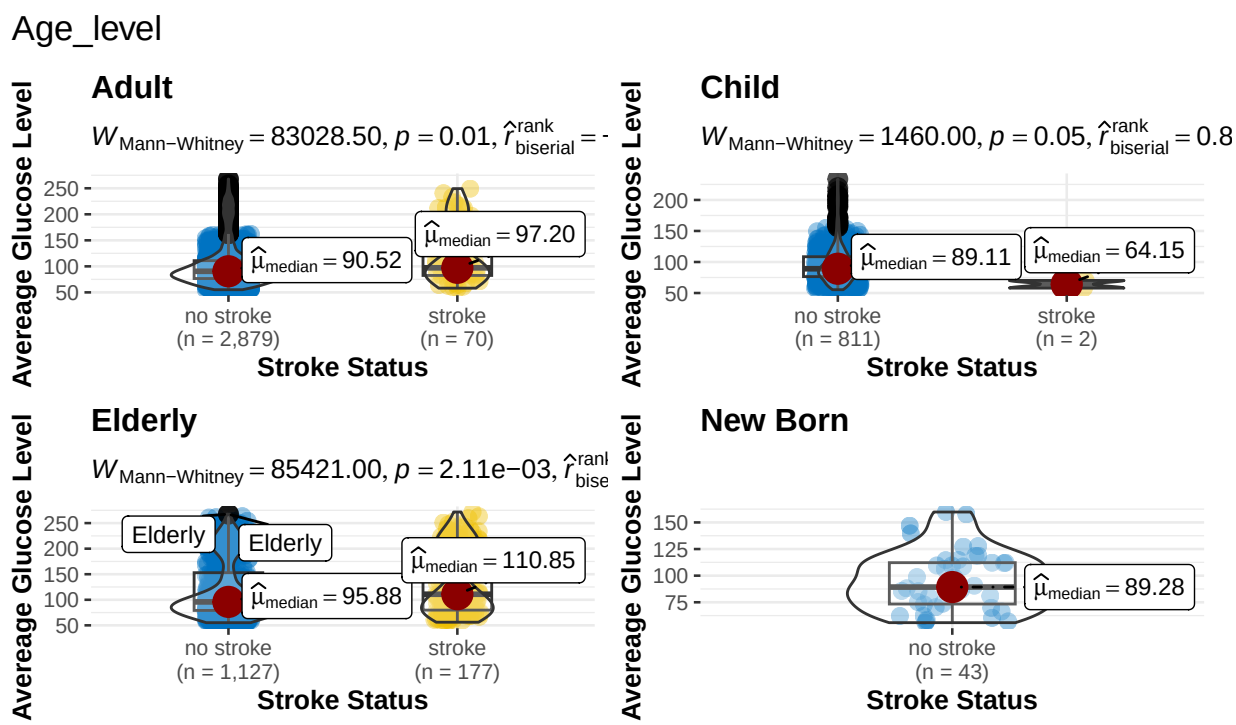
The p-value (p-value < 0.05) indicates that we can reject the null hypothesis, and conclude that at the 5% significance level average glucose level are significantly different between stroke patients and subjects without stroke.

Supplementary Figure 1



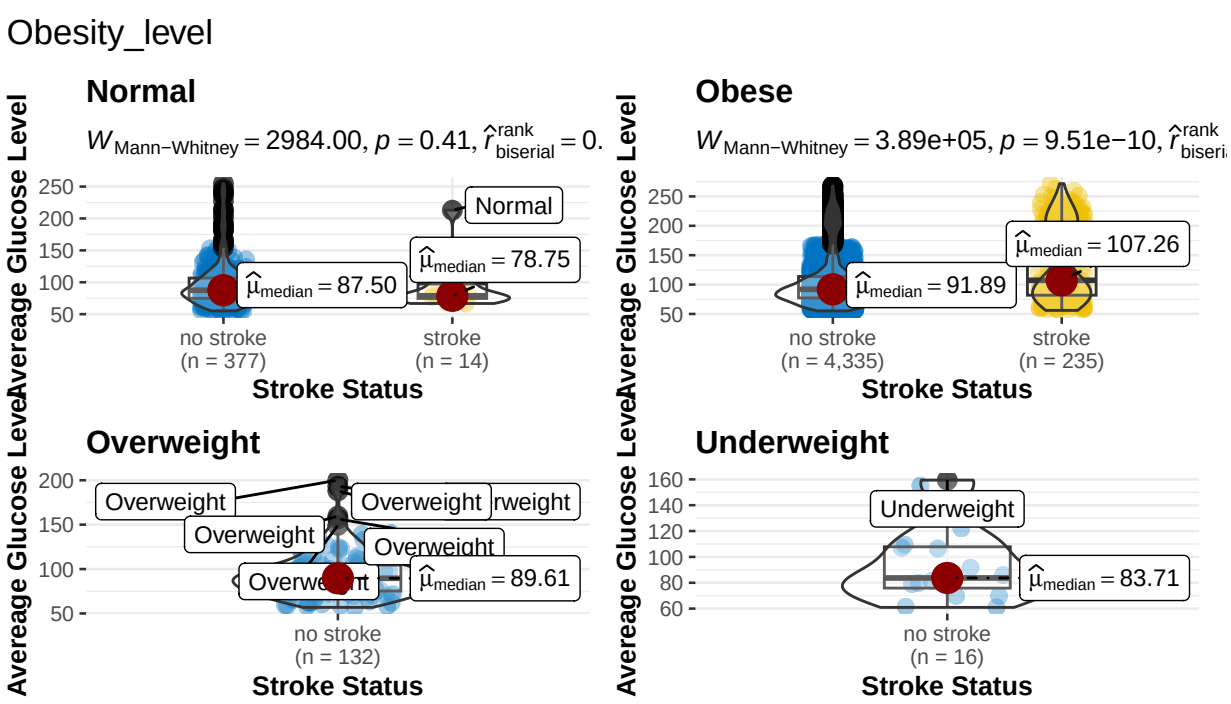
In this plot, it shows that although this box plot displays a higher median average glucose level value in stroke patients, the scatters show that the average glucose level of stroke patients is concentrated above 200(mg/dl) or below 110(mg/dl), with few in the middle.

Supplementary Figure 2



This box violin plot divided objects by age level and conducted Significant tests at each age category. There are no newborn stroke patients and only 2 child stroke patients in this subject which are not sufficient to conduct the significant test. And in adults, the p-value is 0.01, so we can not reject the null hypothesis, while the p-value in the elderly is less than 0.05, therefore we can conclude that average glucose level is significantly different between old stroke patients and old non-stroke patients at a 5% significance level.

Supplementary Figure 3



This box violin plot divided objects by obesity level and conducted Significant tests at each BMI category. In the obese category, the p-value is less than 0.05, therefore we can conclude that the average glucose level is significantly different between obese stroke patients and old non-stroke patients at a 5% significance level.

Conclusion

The null hypothesis may be rejected based on the calculated P-value, and we can draw the conclusion that average glucose levels between stroke patients and people without stroke are significantly different at the 5% significance level. Stroke patients tend to have a higher acerage glucose level.

We can also draw the conclusion that, at a 5% level of significance, the average glucose level differs considerably between elderly stroke patients and elderly non-stroke patients as well as between obese stroke patients and non-stroke sufferers. Elder stroke patients appear to have higher average blood sugar levels than other elderly persons, and this pattern is also seen in the study's obese objects.

References

[1] Tsao CW, Aday AW, Almarzooq ZI, Alonso A, Beaton AZ, Bittencourt MS, et al. Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association. Circulation. 2022;145(8):e153–e639.

[2] Soumyabrata Dev, Hewei Wang, Chidozie Shamrock Nwosu, Nishtha Jain, Bharadwaj Veeravalli, Deepu John,A predictive analytics approach for stroke prediction using machine learning and neural

networks,Healthcare Analytics,Volume 2,2022,100032,ISSN 2772-4425,https://doi.org/10.1016/j.health.2022.100032.

[3] Divine, George PhD*; Norton, H. James PhD†; Hunt, Ronald MD‡; Dienemann, Jacqueline PhD, RN§. A Review of Analysis and Sample Size Calculation Considerations for Wilcoxon Tests. Anesthesia & Analgesia: September 2013 - Volume 117 - Issue 3 - p 699-710 doi: 10.1213/ANE.0b013e31827f53d7