Analysis and Predictions of Stroke Data in which Identifies the Contribution of Average Glucose in Blood and Age

Yen-Lin Lai[1]
1. Department of Biostatistics, Fielding School of Public Health, UCLA, Los Angeles, California 90095, USA.

I. INTRODUCTION

A stroke can appear when the blood supply to part of the brain is blocked or reduced, and it prevents brain tissue from getting oxygen and nutrients [1]. It is one of the top 5 causes of death and a leading cause of disability in the United States [2]. Moreover, there are over 13 million people will have a stroke every year and around 5.5 million people will die as a result in a worldwide view [3]. Thus, reducing stroke cases and deaths remains a critical aspect for better health.

In the past decades, extensive research effort has been directed to depict the causes of a stroke to protect population health. According to Mozaffarian, D., et al., summarizing in Heart disease and stroke statistics - 2016 update, the prevalence of cardio vascular health behaviors factors varies from 1.5% for the healthy diet pattern to up to 78% for the smoking metric, including high blood pressure, body mass index, and heart disease, etc., among US adults [4]. Other risk factor such as the average glucose level in blood has also been shown to act either independently or collectively in determining the incidence of a stroke. The other research conducted by Wang, A., et al. shows that during an 11-year follow-up, after adjusting for confounding variables and comparing with patients in the lowest quartile of baseline Triglyceride-glucose index, those in the third and fourth quartile were associated with an increased risk of stroke (adjusted hazard ratio [HR] 1.22, 95% confidence interval [CI] 1.12–1.33). Furthermore, there exists a linear association between baseline Triglyceride-glucose index with stroke [5].

Since the glucose index and smoking status are some pivotal factors for causing a stroke according to the research mentioned above, an analysis focusing on the association between these factors and stroke would be of great value. Here, we applied a dataset from Kaggle [6], desiring to know the most crucial factor leading to a stroke, as well as establishing the prediction of the possibility of having a stroke. Moreover, we applied several prediction model to decide which one gives the best prediction. This report would be helpful for researchers to get some aspects of causing a stroke.

II. METHODS & MATERIALS

A. Data Set

The dataset *healthcare-dataset-stroke-data.csv Stroke Prediction Dataset* conducted in this study was downloaded from Kaggle [6]. This dataset contains 5110 observations, and 12 variables were included in this dataset covering age, gender, body mass index, average glucose level in blood, marriage status, smoking status, type of work, type of residence, and if the patients have hypertension and heart disease or not, etc.

B. Data Cleaning and Variable Selection

In this study, we used *R Studio Version 1.4.1717* to conduct all data management and analysis. There are 3 numeric variables and 9 categorical variables in this dataset. The variable *id* is eliminated in the first place since there is no information, and we do not want this variable to impact on our later data analysis and to cause unwanted correlation.

When examining the missing value, we found 201 missing values in the variable *bmi*. Although the missing values are small enough to ignore, for the completeness of this dataset, we do the imputation for *bmi* by using MICE (Multivariate Imputation by Chained Equations).

Lastly, we have 11 variables including response variable *stroke*, which will be used in analysis and prediction later on.

C. Statistical Methods

We applied the Logistic Regression Model to all of the explanatory variable to test their respective associations with the probabilities for having a stroke. In our model, the outcome Y represents the probability for having a stroke (0 = the patient did not have a stroke, and 1 = the patient had a stroke). Since there are too many variables, and it will lead to overfitting when we do the classification using Logistic Regression model, we use stepwise method to determine the variables which are significant in this model by choosing the smallest AIC (Akaike information criterion, a criterion to compare the fit of several regression models). We get four of the most significant variables, which are age, hypertension, heart disease, and the average level of glucose. The coefficient β_1 is the expected change in log odds of having a stroke when the participant gets one year older; the coefficient β_2 is the expected change in

log odds of having a stroke when the participant has hypertension; the coefficient β_3 is the expected change in log odds of having a stroke when the participant has any heart diseases; and the coefficient β_4 is the expected change in log odds of having a stroke per unit change in the average level of glucose. An error term ε_i was incorporated into our model as a residual variable that accounts for a lack of perfect goodness of fit.

The equations for the Logistic Regression models in this project are shown as below:

$$Y = \beta_0 + \beta_1 * age + \beta_2 * hypertension + \beta_3 * heart\_disease + \beta_4 * avg\_glucose\_level + \varepsilon_i$$

where:

- Y = the probability for having a stroke
- β_0= intercept of the model
- β_1 = coefficient of *age*
- β_2 = coefficient of *hypertension*
- β_3 = coefficient of *heart_disease*
- β_4 = coefficient of *avg_glucose_level*
- ε_i = error term

The log probabilities for having a stroke are reported as log – average percent changes with 95% CI according to the model results.

III. RESULTS

A. *Logistic regression analysis for the prediction of the probabilities for having a stroke*

The stepwise method of logistic regression has chosen the four most significant variables in this model, which are age, hypertension, heart disease, and the average level of glucose (Table 1). The exponential over the coefficients indicates the estimator odds. If age of the participant increases 1, the estimated odds of having a stroke will increase by 1.071 (the exponential over its estimator). Similarly, for a participant who has hypertension, the estimated odds of having a stroke will increase by 1.464 comparing to whose who has no hypertension (Table 1). Interestingly, we can observe that although the variables *age* and *avg_glucose_level* have lower exponential over the corresponding coefficient than the others, their P-value are relatively lower, which means these two variable are more significant in this model than the other two variables: *hypertension* and *heart_disease*. That is because *age* and *avg_glucose_level* are continuous variables, which means the increase per unit have less effect on this logistic regression model, comparing to *hypertension* and *heart_disease,* which are categorical variables.

Table 1. Logistic regression analysis of estimators of having a stroke

Variables	β[a]	Exp[b](β)	SE[c]β	P-value
Intercept	-7.49	0.0005	0.358	<0.001***
Age	0.069	1.071	0.005	<0.001***
Hypertension (vs. no hypertension)	0.381	1.464	0.163	0.019*
Heart disease (vs. no heart disease)	0.330	1.391	0.188	0.079
Average glucose level	0.004	1.004	0.001	<0.001***

N = 5110 used in this logistic regression analysis; Residual deviance = 1591.5; AIC = 1601.5

[a]: Coefficient of the estimator. [b]: Exponential value. [c]: Standard error.

*P-value < 0.05, ***P-value < 0.001

According to the P-values shown in Table 1, we can determine that age and average glucose level are the two most important variables in predicting whether a patient have a stroke or not (both of the P-vales are less than 0.001). We are interested in which variable is more decisive in the prediction, thus a scatter plot is printed to show the distribution of the patient who had a stroke (Figure 1).

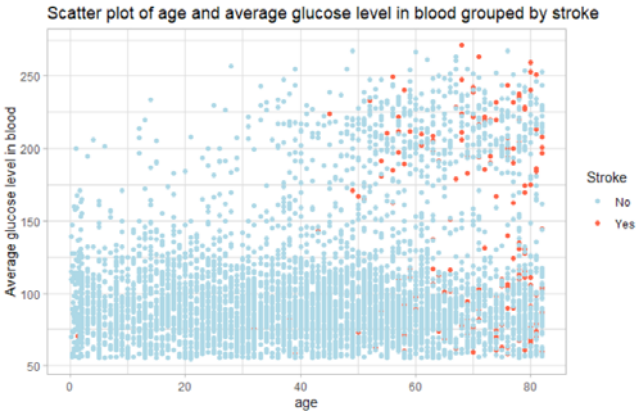

Figure 1. Scatter plot of the most crucial variables – age and average glucose level in blood. The patient who had a stroke is marked by red point, and those who did not have a stroke is marked by blue point. Observed shows that age seems to be more determinant variable while predicting having a stroke or not of a participant.

From Figure 1, we observe that most of the red points (the patients who had a stroke) are distributed in the right side of the scatter plot, i.e., those who had a stroke are mostly older than 45. On the other hand, we do observe the patients whose average glucose level in blood is higher have a higher chance to have a stroke, however, the pattern is not as clear as the variable *age*. We need further examination to determine whether age is more important variable than average glucose level in blood.

The logistic regression using two variables, age and average glucose level in blood, is visualized in Figure 2. Each dot shown is the observation of the

patients had a stroke before or not. Lighter blue dot means the patient has a higher glucose level, while deeper blue dot means the patient has a lower glucose level.

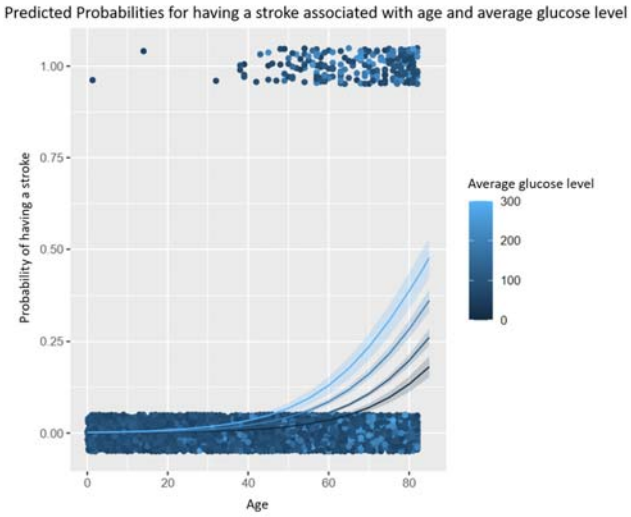

Figure 2. Predicted probabilities for having a stroke associated with age and average glucose level. Observed shows either the patient was identified as having a stroke history "1" or no "0". The four blue line give the predicted probabilities with different average glucose level. Shaded region shows the upper and lower boundary for 95% CI.

The prediction line of the probability for having a stroke indicates that even the participant has a high glucose level (the prediction line of the lightest blue) and is older than 80, we still predict this participant did not have stroke. This result shows that logistic regression is not effective to do a prediction in this dataset, which is highly imbalanced data, meaning the outcome variable stroke is very imbalanced – only 5% of the participants had a stroke comparing to other 95% of them did not. We thus need other means to predict our dataset.

B. *Applying some classification models for the prediction of the probabilities for having a stroke*

To evaluate the prediction, we applied different types of classification, including Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, K-nearest Neighbors Algorithm, Random Forest, and Support Vector Machine, where we get highest accuracy (95.1%) in Logistic Regression, Random Forest, and Support Vector Machine (Table 2).

Other types of classification may seem not as accurate as Logistic Regression, Random Forest, and Support Vector Machine, but we might find some pattern in the analysis. Linear Discriminant Analysis seems more accurate than Quadratic Discriminant Analysis while doing prediction, which is indicating that the model tends to be linear, i.e., there is no quadratic pattern in this logistic regression model (Table 2).

Table 2. Accuracy of the prediction of the testing data

Types of Classification	Accuracy
Logistic Regression[a]	0.951
Linear Discriminant Analysis[a]	0.949
Quadratic Discriminant Analysis[a]	0.885
K-nearest Neighbors Algorithm	0.949
Random Forest	0.951
Support Vector Machine	0.951

[a]: Using the same logistic model in Table 1

However, the result can be biased. As we mentioned before, the dataset we are using is highly imbalanced, which means the high accuracy can be simply achieved by guessing all of the patients did not have a stroke. In fact, the highest accuracy shown in Table 2 is given by predicting all of the participants did not have a stroke, which is not the result we wanted. We will elaborate how to solve this problem and further examine the accuracy of the prediction in the future in the discussion section.

V. DICUSSION

In this project, we conducted several analyses on different variables, including age, hypertension, heart disease, and the average glucose level in blood to investigate their contributions to having a stroke. We also tried to utilize several models to predict the probabilities of having a stroke, including Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, K-nearest Neighbors Algorithm, Random Forest, and Support Vector Machine. We found that the probability of having a stroke increases with increasing ages and the average glucose level in blood. Other than that, having hypertension and having any kind of heart disease also play an important role comparing to other variables we put in our logistic model.

By those analysis, researchers can get some perspectives of the main factors causing a stroke, which in this research, smoking status is excluded, and it is not conforming to other research or the common sense. One can do more statistical analysis to examine if there are some confounding factors between smoking and having a stroke. The classification models we conducted also need more examination to determine whether the accuracy

represents how fit is the model due to the high imbalance of this dataset. By calculating Specificity, Sensitivity, True Positive Rate, False Positive Rate and constructing AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curve for each models we used to do the classification, we can compare the fitting of the model more accurately.

IV. CONCLUSIONS

In this project, we analyzed the probability of having a stroke and conducted some classification models for four variables, including age, hypertension, heart disease, and the average glucose level in blood. We found that the probability of having a stroke increase when a patient is older and has a higher glucose level in blood. By doing classification in different types of model, we also found that Logistic Regression, Random Forest and Support Vector Machine has more fitness to predict whether a participant had a stroke or not in this data.

REFERENCES

1. Mayo Clinic. Symptoms and causes of storke [cited 2021 12/5]; Avilable from: https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113

2. American Stroke Association. About stroke [cited 2021 12/5]; Available from: https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113

3. World Stroke Organization. Why stroke matters – About Stroke [cited 2021 12/6]; Available from: https://www.world-stroke.org/world-stroke-day-campaign/why-stroke-matters/learn-about-stroke

4. Mozaffarian, D., et al., *Heart disease and stroke statistics - 2016 update: a report from the American Heart Association.* American Heart Association, 2016. **133**(4): p. e39-e49.

5. Wang, A., et al., *Triglyceride-glucose index and the risk of stroke and its subtypes in the general population: an 11-year follow-up.* Cardiovascular Diabetology, 2021. **20**(46): p. 1-3.

6. Fedesoriano, Kaggle [cited 2021 12/6]; Available from: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Appendix: R code

Table1

```
full.model <- glm(stroke ~ .,
                  data = stroke.data,
                  family = "binomial")
step.model <- stepAIC(full.model, direction = "both",
                      trace = FALSE)
summary(step.model)
```

Figure1

```
ggplot(stroke.data, aes(x = age, y =
avg_glucose_level)) +
  geom_point(aes(color = factor(stroke)))+
  labs(title = "Scatter plot of age and average
glucose level in blood grouped by stroke",
       y = "Average glucose level in blood",
       color = "Stroke")+
  scale_color_manual(labels = c("No", "Yes"),
values = c("lightblue", "tomato1"))+
  theme_light()
```

Figure 2

```
ggPredict(final.model_numeric,se=TRUE,interactive
=TRUE)
```

Table 2

```
logistic_model <- glm(stroke ~ age + hypertension +
heart_disease + avg_glucose_level,
                      data = train_reg,
                      family = "binomial")
predict_reg <- predict(logistic_model,
                       newdata = test_reg, type
= "response")
predict_reg <- ifelse(predict_reg >0.5, 1, 0)
table(test_reg$stroke, predict_reg)
missing_classerr <- mean(predict_reg !=
test_reg$stroke)
print(paste('Accuracy =', 1 - missing_classerr))
library(caret)
lda.fit=lda(stroke ~ age + hypertension +
heart_disease + avg_glucose_level,
            data = train_reg,
            family = "binomial")
qda.fit=qda(stroke ~ age + hypertension +
heart_disease + avg_glucose_level,
            data = train_reg,
            family = "binomial")
knn.pred=knn(train_numeric,test_numeric,train_nu
meric$stroke,k=10)
rf <- randomForest(as.factor(stroke) ~ .,data =
train_reg, n_tree = 100)
rf.pred.prob <- predict(object = rf, newdata =
test_reg, type = "prob")
svm <- ksvm(as.factor(stroke) ~ .,data = train_reg)
svm.pred.prob <- predict(svm,test_reg, type =
"decision")
svm.pred <- predict(svm,test_reg, type =
"response")
```