The Significance of Risk Factors Associated With Brain Stroke

Christopher Jung

Department of Statistical Sciences, University of Toronto

STA303H1S; Methods of Data Analysis 2

Mohammad K.A. Khan

April 7, 2024

Introduction

Brain stroke is a disease that has been concerning over the years. Due to the unawareness of the significance of other symptoms to the disease, the relationship is often overlooked. Diabetes, obesity (high BMI), and heart disease are significant factors that can cause a stroke. High blood pressure and high cholesterol are also some underlying significant factors. The goal of the analysis is to study the significance of these 5 variables to brain stroke. In one past literature, it was stated that “Diabetes is an important modifiable risk factor for stroke” (Chen, 2016). Obesity, heart disease, heart disease, and high cholesterol which aligns with this analysis (Kurth, 2002), (Arboix 2015).

Method

Study Population

The original data used for this study comes from The Behavioral Risk Factor Surveillance System (BRFSS) survey from 2015 that was conducted annually to study health-related risk factors among Americans. The cleaned version of the data for this study contained 253680 observations, each observation representing an individual (Taboul).

Variables

For this study, Stroke was the outcome variable. Diabetes, BMI, Heart disease, High Blood Pressure, High cholesterol were the measures of interest. There were 22 variables in the full (original) model.

Variable Selection

The data was fit as a full model using a logistic regression under GLM where the canonical link was logit. The model was first studied based on its summary. After the summary study, different variable selection methods such as forward selection based on AIC, forward selection based on BIC, and Least Absolute Shrinkage and Selection Operator (LASSO) selection were applied on the data.

For the forward selection methods, the process started as an empty model and iteratively built on the model based on its AIC/BIC value. It concluded that the resulting model was the model with the lowest AIC/BIC value. The two models were chosen as candidates for the final model of this study. For the LASSO method, the coefficients of variables in the regression that are not significant enough could shrink to 0 which functioned as a natural selection approach to obtain the final model. Containing only the significant variables, the resulting model was chosen as a candidate for the final model.

Diagnostics and Validation

After each variable selection method, diagnostics and validation of the model was studied including Cook's distance, calibration plot, ROC curve. Under diagnostics, Cook's distance plot was fit to identify any influential observations that affected the estimation of all fitted values in each model. If any influential observations were nonsensical, it was removed. The model with the respective nonsensical observations were refit using the respective variable selection. Next, validation tools such as calibration plot and ROC curve were used for validation for each model.

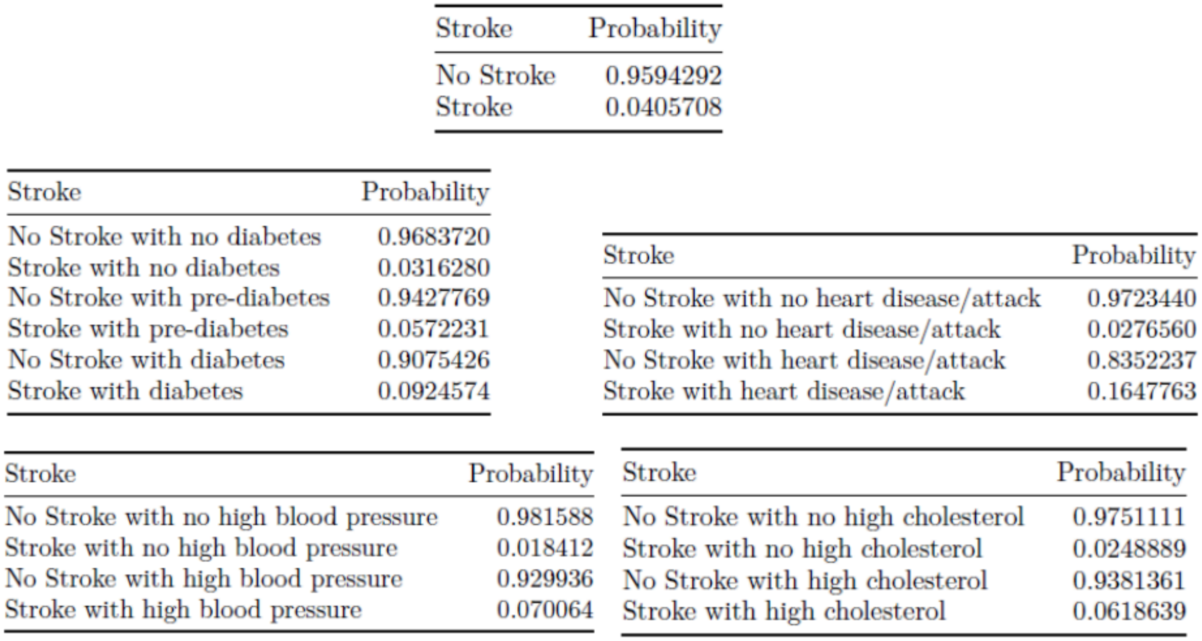
For the calibration plot, cross validation was implemented with 10 repetitions. Using an algorithm of splitting the data into training and testing sets, cross validation assessed prediction accuracy by mean absolute error or mean squared error. The predictions were plotted on the calibration plot, visually showing whether the observed values in the model resembled the estimated values.

For the ROC curve, the main purpose was validation with statistical discrimination under binary classification. For each probability in the model, the classification used a threshold that classified the true outcome into two groups, 1 for having a stroke and 2 for not having a stroke. Plotting the sensitivity against the false positive rate under discrimination, the ROC curve was implemented.

Having three different models, the results from the calibration plot and the ROC curve were compared. The model with the most plausible outcome was selected as the final model. The final model contained all the significant variables from variable selection and was validated by the model validation process. With the final model, significance of each exposure of interest was studied.

Results

Figure 1: Probability Tables of Stroke in the Measures of Interests



Stroke	Probability
No Stroke	0.9594292
Stroke	0.0405708

Stroke	Probability
No Stroke with no diabetes	0.9683720
Stroke with no diabetes	0.0316280
No Stroke with pre-diabetes	0.9427769
Stroke with pre-diabetes	0.0572231
No Stroke with diabetes	0.9075426
Stroke with diabetes	0.0924574

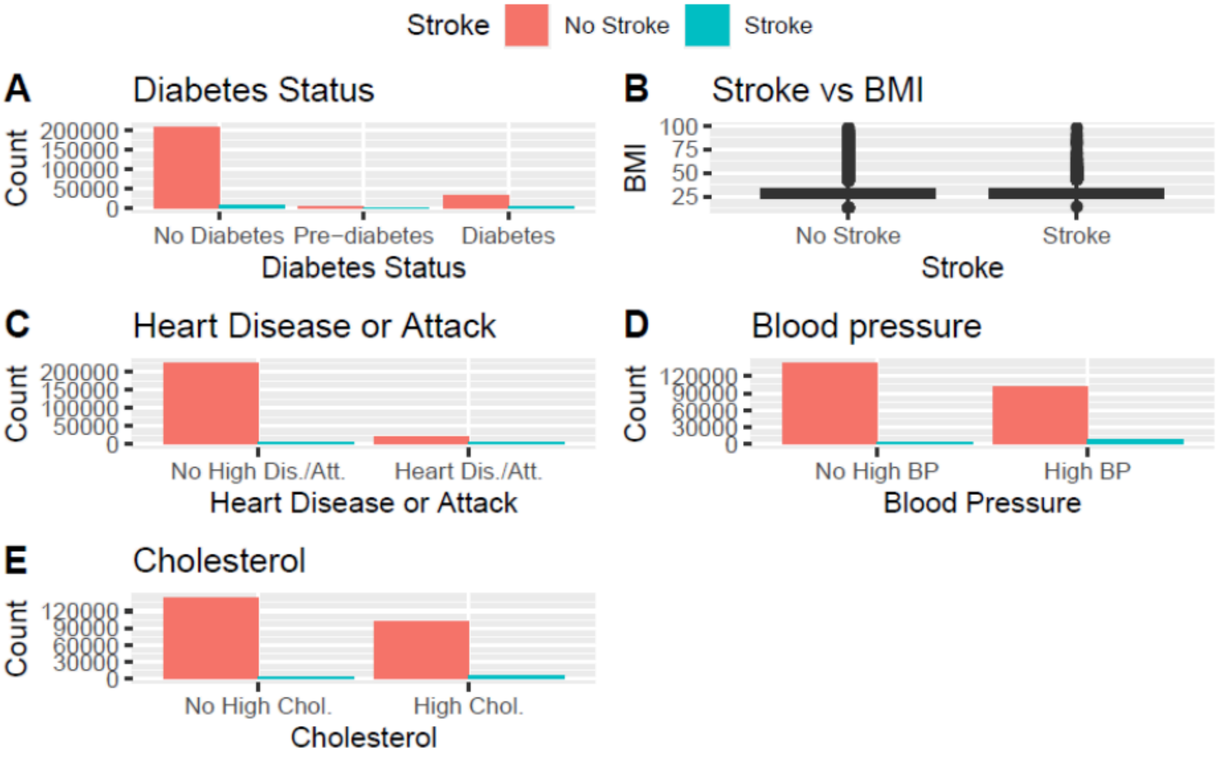
Stroke	Probability
No Stroke with no heart disease/attack	0.9723440
Stroke with no heart disease/attack	0.0276560
No Stroke with heart disease/attack	0.8352237
Stroke with heart disease/attack	0.1647763

Stroke	Probability
No Stroke with no high blood pressure	0.981588
Stroke with no high blood pressure	0.018412
No Stroke with high blood pressure	0.929936
Stroke with high blood pressure	0.070064

Stroke	Probability
No Stroke with no high cholesterol	0.9751111
Stroke with no high cholesterol	0.0248889
No Stroke with high cholesterol	0.9381361
Stroke with high cholesterol	0.0618639

Note: The first table was the study population probability. The bottom four tables were tables of probability of having a stroke in each level of each measure of interest.

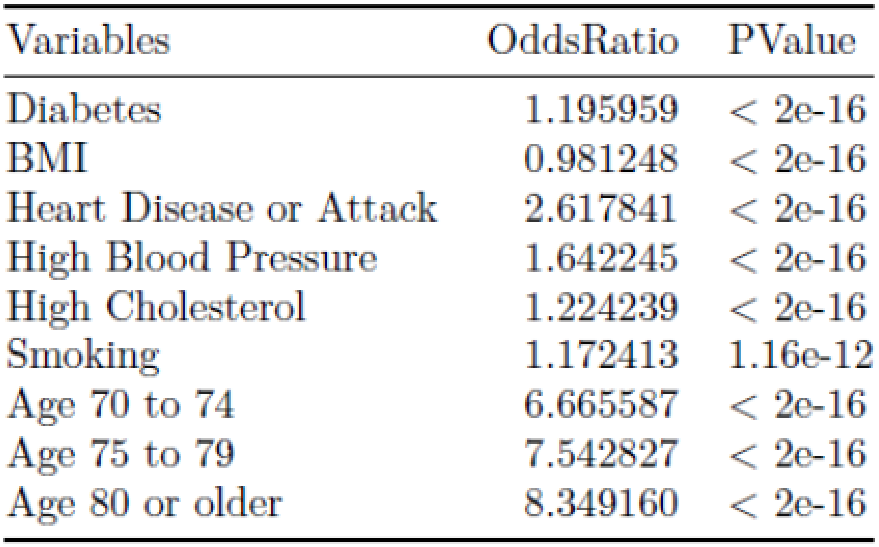
Figure 2: No Stroke Vs. Stroke Plots for Measures of Interests



Note: This figure displays plots that show No Stroke vs Stroke for each category within different measures of interests.

From Figure 1, it was shown that 4 percent of the population had a stroke in the past. From Figure 1 & 2, the group with diabetes, higher BMI, heart disease, high blood pressure, or high cholesterol had a higher proportion of having a stroke than the group without the risk factors respectively. The findings showed that these factors could be important candidates for model variables.

Table 1: Partial Numerical Summary of Full Model



Variables	OddsRatio	PValue
Diabetes	1.195959	$< 2e-16$
BMI	0.981248	$< 2e-16$
Heart Disease or Attack	2.617841	$< 2e-16$
High Blood Pressure	1.642245	$< 2e-16$
High Cholesterol	1.224239	$< 2e-16$
Smoking	1.172413	1.16e-12
Age 70 to 74	6.665587	$< 2e-16$
Age 75 to 79	7.542827	$< 2e-16$
Age 80 or older	8.349160	$< 2e-16$

Note: The table included the measures of interest and other significant variables, such as, smoking and age.

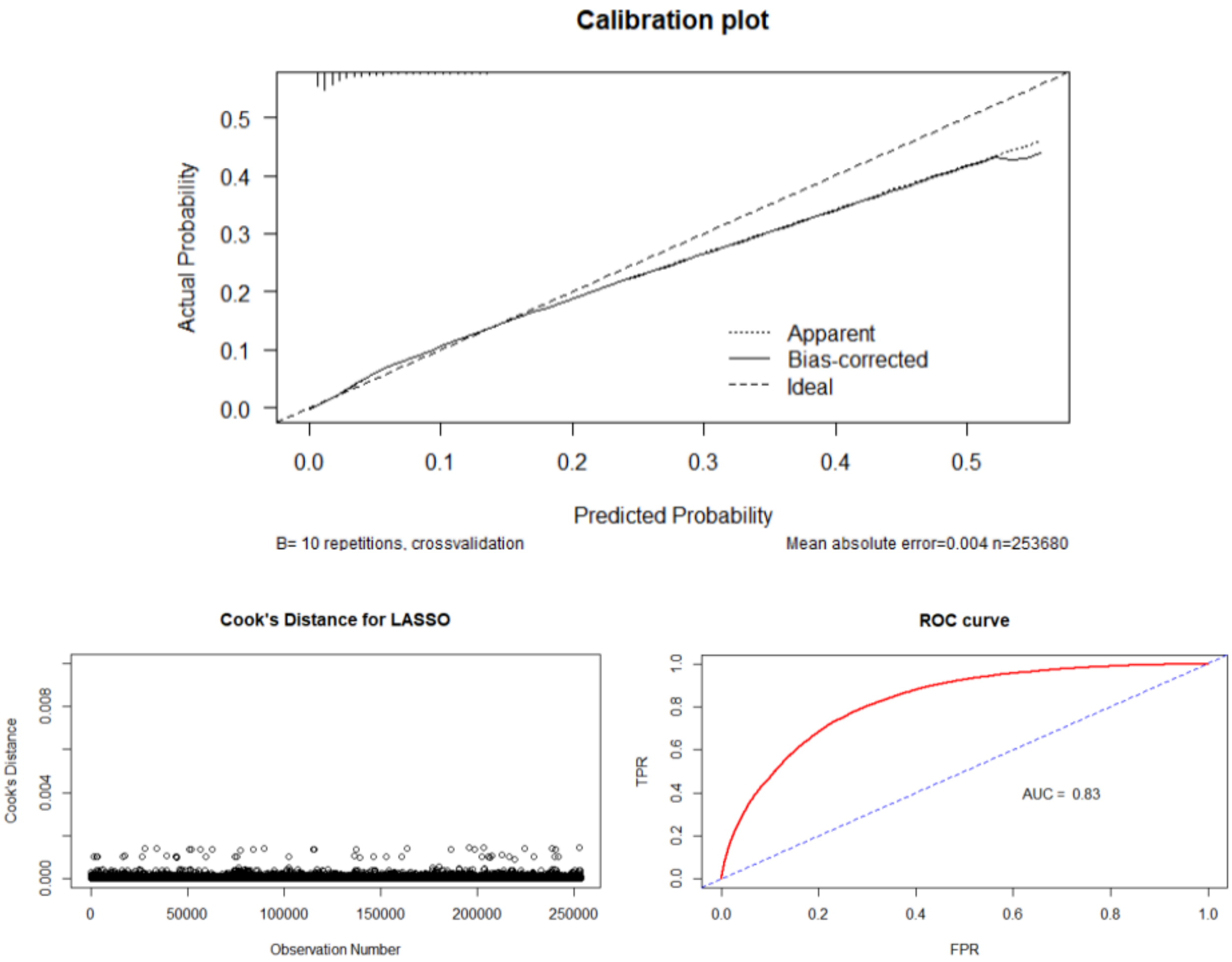
Initially, analysis of the full model was studied to confirm that the EDA findings were aligned with the full model (Table 1). The odds of having a brain stroke in the heart disease or attack group was 2.62 times the odds of having a brain stroke in the non-heart disease or attack group. Diabetes, high blood pressure and high cholesterol also displayed a similar relationship to brain stroke. However, BMI seemed to have an odds ratio close to 1. To determine the final model, different variable selection was implemented as candidates.

From forward selection based on AIC, the measures of interests were all selected where total variables selected were 18. Under diagnostics, Cook's distance showed 0 observations were influential observations. For model validation, the calibration plot showed Mean Absolute Error (MAE) of 0.04 and Mean Squared Error (MSE) of 0.00006. From the ROC curve, AUC value was 0.83. The following values suggested good model performance and a valid model for estimation, hence, the model was a good candidate for the final model.

From forward selection based on BIC, the measures of interests were all selected where total variables selected were 15. Under diagnostics, Cook's distance showed 0 observations were influential observations. For model validation, the calibration plot showed MAE is 0.04 and MSE is 0.00006. From the ROC curve, AUC value was 0.83. The model was a good candidate for the final model from good model performance and a valid model for estimation.

From LASSO selection, the measures of interests were all selected where total variables selected were 14. Under diagnostics, Cook's distance showed 0 observations were influential observations. For model validation, from Figure 3, the calibration plot showed MAE is 0.05 and MSE is 0.00006. From the ROC curve, AUC value was 0.83. The model was a good candidate for the final model.

Figure 3: Model Diagnostics and Model Validation Plots for the Final Model



Notes: Cook's distance cutoff: 0.98. For the calibration plot, the dotted line is apparent, while the continuous line is bias-corrected. Mean Absolute Error = 0.004, Mean Square Error = 0.00006. AUC is 0.83.

Forward selection for AIC chose 18 variables, BIC chose 15 variables, and LASSO selection chose 14 variables. All three models selected the measures of interests. While sharing similar model performance based diagnostics and model validation, LASSO selection had the least amount of predictors. Hence, the LASSO model was chosen as the final model due to its simpler model complexity.

Discussion

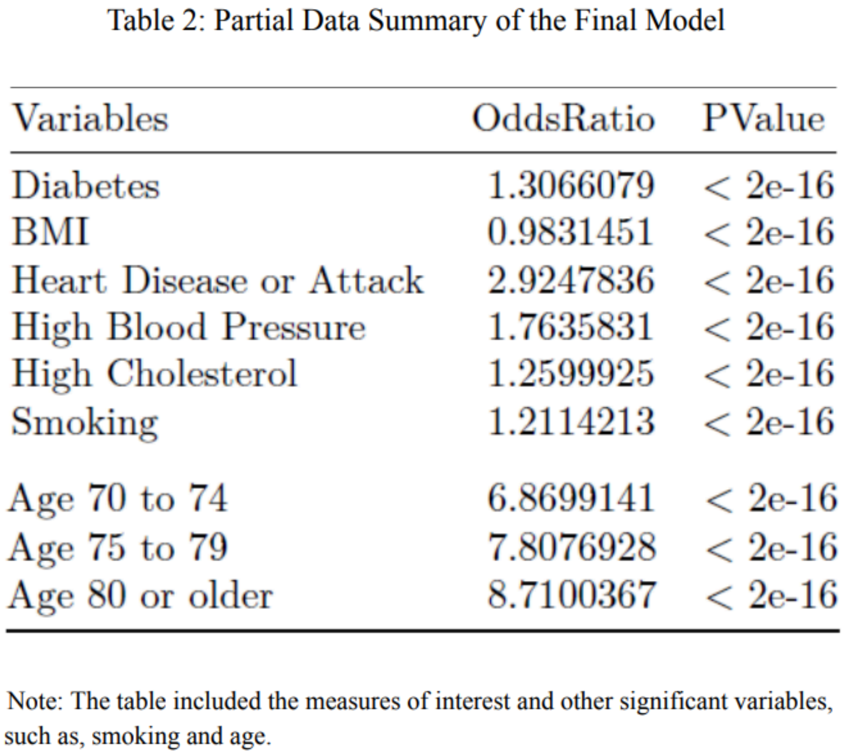


Table 2: Partial Data Summary of the Final Model

Variables	OddsRatio	PValue
Diabetes	1.3066079	< 2e-16
BMI	0.9831451	< 2e-16
Heart Disease or Attack	2.9247836	< 2e-16
High Blood Pressure	1.7635831	< 2e-16
High Cholesterol	1.2599925	< 2e-16
Smoking	1.2114213	< 2e-16
Age 70 to 74	6.8699141	< 2e-16
Age 75 to 79	7.8076928	< 2e-16
Age 80 or older	8.7100367	< 2e-16

Note: The table included the measures of interest and other significant variables, such as, smoking and age.

From Table 2, Odds ratios were used to study the relationship between the measures of interest, such as Diabetes, BMI, Heart Disease or Attack, High Blood Pressure, and High Blood Cholesterol, and Stroke. The odds of having a stroke in the heart disease or attack group was 2.92 times the odds of having a stroke in the heart disease or attack group. Diabetes (OR = 1.31), high blood pressure (OR = 1.76) and high cholesterol (OR = 1.26) also displayed a similar relationship to brain stroke. Again, BMI seemed to have an odds ratio close to 1. The final model showed a higher set of odds ratio than the full model. Overall, the findings in the final model conducted by LASSO demonstrated that the measures of interest are significant and must be taken seriously.

For variable selection limitations, stepwise and backward stepwise elimination did not finish in a reasonable time due to an extremely large number of observations. Only forward stepwise selection for AIC and BIC were considered for stepwise selection.

For LASSO, the misclassification error plot from the algorithm displayed a nearly straight line. The algorithm had a difficult time finding the best lambda with 1 standard error. This penalized the dataset too heavily. To compensate, the minimum lambda was chosen to be the best lambda and the number of folds was increased to 25.

In conclusion, this study showed that diabetes, heart disease, high blood pressure, and cholesterol are associated with stroke. The odds of having a stroke increased when these risk factors were present. Furthermore, obesity had an association with stroke, but not as strong as the other factors. These results indicate that there must be additional studies to further improve the understanding of other health-related factors to brain stroke.

References

Chen, R., Ovbiagele, B., & Feng, W. (2016). Diabetes and Stroke: Epidemiology, Pathophysiology, Pharmaceuticals and Outcomes. The American journal of the medical sciences, 351(4), 380–386. https://doi.org/10.1016/j.amjms.2016.01.011

Kurth T, Gaziano JM, Berger K, et al (2002). Body Mass Index and the Risk of Stroke in Men. Arch Intern Med., 162(22):2557–2562. doi:10.1001/archinte.162.22.2557

Arboix A. (2015). Cardiovascular risk factors for acute stroke: Risk profiles in the different subtypes of ischemic stroke. World journal of clinical cases, 3(5), 418–429. https://doi.org/10.12998/wjcc.v3.i5.418

Taboul A. Diabetes Health Indicators Dataset. diabetes_012_health _indicators_BRFSS2015.csv. Retrieved from https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data