Determinants of Recurrent Stroke Incidence in a Clinical Sample

Critical Thinking Group 3

Data 621

Dr. Jeffrey Edwards

CUNY School of Professional Studies

Coco Donovan

Jean Jimenez

Marjete Vucinaj

Matthew Roland

Determinants of Recurrent Stroke Incidence in a Clinical Sample

Abstract

This study details our efforts to predict the likelihood of experiencing a stroke using a mix of demographic, physical health, and behavioral variables through a logistic regression model. We utilized demographic factors such as age, gender, and race/ethnicity in our model, alongside health indicators including heart health, family history of stroke, smoking status, BMI, and Diabetes status. By analyzing data from the StrokeHealth Outcomes (SHOUT) database, the logistic regression model will identify statistically significant predictors of stroke risk, accounting for potential confounders and interactions between variables. The findings, as they pertain to predictive success and relevant predictor variables, will inform strategic medical/public health interventions that aim to reduce stroke burden by identifying high-risk populations and modifiable risk factors. Ultimately, this research aims to contribute to our understanding of stroke determinants and guide preventive strategies to improve public health outcomes. Our study highlights meaningful relationships between patients' health and demographic info such as age and gender, in addition to the complexity that varied socio-demographics present when trying to understand medical outcomes.

Keywords: *Recurrent Stroke, Logistic Regression, Stroke Risk Factors, Predictive Modeling, Health Disparities*.

Introduction

Stroke incidences are a serious and debilitating phenomenon that is the leading cause of death for Americans and affects an estimated 795,000 Americans each year[14]. In terms of demographics, the risk of stroke tends to increase with age and with the presence of high blood pressure, high cholesterol, obesity, diabetes, and smoking. Furthermore, stroke tends to disproportionately affect individuals based on race and ethnicity, with the likelihood of a stroke in Black adults being twice as high compared to White adults, highlighting a clear systematic disparity[11].

These medical emergencies are characterized by an interference in cranial blood flow, resulting in a variety of symptoms, including lateral numbness, confusion, speech deficits, mobility difficulties, severe headache, the loss of consciousness and even death[7]. In addition, depending on the severity, the presence of a stroke can result in long-term deficits, including permanent disablement. Importantly, the long-term risks associated with stroke are exacerbated by the fact that stroke survivors face double the risk of stroke recurrence within 10 years following the first incidence[12].

Despite the clear importance of recurrent stroke regarding clinical outcomes, there is presently a gap in the literature surrounding the prognosis of recurrent stroke[3]. However, a meta-analysis of 9 studies regarding stroke outcomes by Ferrone et al.[7] indicated that three studies reported on mortality rates; one of which reported a 25.9% mortality rate for in-hospital and hospice occurrences[9]. Two other studies reported mortality rates of 20.6% 30 days following the incidence of a recurrent stroke and 11.6% 4 years following the occurrence of a recurrent stroke[1]. Finally, Albright and colleagues found that only 11.3% of recurrent stroke patients reported disability-free outcomes. Although more research is needed regarding these outcomes,

it is clear that recurrent stroke is a debilitating public health issue that requires attention in terms of delineating factors associated with prognosis, risk, and, importantly, prevention.

Identifying risk-factors and preventative measures for recurrent stroke can be particularly difficult, as strokes are not homogeneous in nature. However, stroke incidences can present as a variety of subtypes with differing risk-factors and, by extension, differing preventative measures to combat onset[9]. For instance, although the presence of diabetes mellitus is presumed to be a primary risk-factor for stroke recurrence, some evidence suggests that it may only be a risk-factor for recurrent cardioembolic (CE) stroke subtypes[7]. On the other hand, white matter hyperintensities (WMH) present as a risk-factor for all stroke types except recurrent CE and lacunar infarct (LI) subtypes. Conversely, other factors, such as smoking, hypertension, age, and a history of prior stroke or transient ischemic attack (TIA) are general risk-factors for all stroke subtypes.

Indeed, recurrent stroke is a chronic and persistent phenomenon that can exacerbate negative stroke outcomes. Clearly, this topic requires further study and documentation regarding both risk-factors and long-term outcomes in terms of disability and mortality. The purpose of this project is to examine the potential risk-factors of recurrent stroke onset within a clinical sample.

Methods

Sample

This study utilized the StrokeHealth Outcomes (SHOUT) database[8]. The SHOUT database is a retrospective dataset collected from patients with acute ischemic stroke (AIS) who were treated at a single comprehensive stroke center. This center serves as the central facility in a substantial, integrated stroke network spanning urban and suburban areas. The database covers a period of ten years, from January 1, 2012, to December 31, 2023. The data used in

this analysis was all adult patients with AIS in 2023. Patient records were transferred into the SHOUT database from the American Heart Association's Get-With-The-Guidelines-Stroke database. The database received institutional review board approval, including a waiver of consent due to its retrospective nature.

Patients included were adult patients that had a diagnosis of AIS during their hospital visit discharged between the dates of January 1st, 2023 and December 31st, 2023. Recurrent stroke was defined as any stroke diagnosis that occurred after the initial visit. Patients that had a stroke during their initial hospital stroke visit were excluded.

Data Collection

2023 data was collected, de-identified (to adhere to HIPPA), and shared with the rest of the team. The dataset was cleaned and an exploratory data analysis was conducted to assess the quality of our data. Missing values were filled in using median imputation, and dummy variables were coded. A logarithmic transformation was conducted on some factors to achieve a more normal distribution.

Analysis

The dataset was split into a training and testing dataset. A correlation matrix was created and factors that were correlated to one another were removed / condensed (**figure 4)**.

Due to the nature of our Target variable, we decided to begin our modeling process with a simple logit model using all possible predictor variables[6]. The result of this model was a significant amount of multicollinearity, which mainly arose from the presence of dummy variables used to include demographic information in the predictor process. To remove the multicollinearity, we turned to Ridge Regression and Lasso Regression. These two regression methods handle multicollinearity in slightly different manners.

Lasso regression handles multicollinearity by introducing a penalty term proportional to the sum of the absolute values of the coefficients of the Least Squares Objective Function. By introducing the penalty term, Lasso Regression can encourage sparsity (the property where many coefficients in the model are zero) bringing the less relevant predictor variables' coefficients to at or around zero[6]. The operative detail here is that some variables will have zero as their coefficient, which means the Lasso Regression has eliminated those variables from the model. The parameter λ (lambda) controls the strength of the penalty. The parameter λ controls the balance between goodness of fit and model complexity, where a large λ yields more shrinkage and sparsity with the inverse true for a small λ.

Ridge Regression handles multicollinearity by introducing a penalty term, represented by the parameter λ, once again. The difference in Ridge regression compared to Lasso Regression is that while the Ridge regression penalty term drives the coefficients towards zero, the coefficients cannot be precisely zero like in Lasso regression[6]. As Ridge regression does not drive the coefficients to exactly zero, Ridge regression can stabilize coefficient estimates by reducing the variance in the coefficient estimates. When a Ridge regression model's λ parameter is larger, there is more shrinkage with potentially higher bias, whereas the inverse is true with a smaller λ.

We also employed a decision tree in this modeling process for good measure. A decision tree operates utilizing split points determined using some decision-making criteria. We used the Recursive partitioning algorithm from the rpart package for the decision tree model. This approach uses what is known as a “greedy” approach, meaning that a locally optimal solution is made at each split point using impurity measures such as Gini impurity or cross-entropy without consideration of the globally optimal solution[5]. After building the initial decision tree, the

Recursive Partitioning algorithm may return to the model and remove splits that may not significantly improve the model's predictive performance.

In theory, we employed these additional regression methods to add sophistication to our eventual prediction method. However, in practice, we compared model performance and weighed the added complexity that these more nuanced methods may contain to inform our model selection decision. Ultimately, the traditional logit model yielded the most ideal MSE measurement, which, when coupled with its level of simplicity in comparison to the other models, made the logit model the logical choice for predictive purposes.

Results

The clinical dataset had 29,662 patients. In the analysis of numerical variables from a dataset focused on recurrent stroke patients, various descriptive statistics have been done, as shown in **Table 1**. The age of the patients ranged from 18 to 121 years, with a mean age of 71.36 years and a median age of 73.00 years. The standard deviation of 14.45 indicates a moderate spread in the ages of the patients. Length of stay in the hospital varied widely among patients, from a minimum of -2,032 hours (likely due to patients transferring from another hospital tracked by a negative value) to a maximum of 9,666 hours, with a mean stay of 167.63 hours and a median of 118.00 hours. This variable displayed a substantial standard deviation of 214.77, reflecting significant variability in hospital stay durations. The Modified Rankin Scale (MRS) scores at discharge, cleaned of errors, averaged 2.07 with a standard deviation of 1.72, suggesting a moderate level of residual post-stroke disability. Scores ranged from 0 (no symptoms) to 8 (severe disability).NIH Stroke Scale (NIHSS) scores upon arrival, which measure the severity of stroke symptoms, also varied considerably, from -7 to 999, with an average score of 6.55. The cleaned NIHSS scores show a similar distribution, indicating that the cleaning process preserved the data's integrity.The proportion of patients who received tissue

plasminogen activator (tPA), a treatment for ischemic stroke, was notably low, with an average of 0.12. Body Mass Index (BMI) averaged at 27.79, suggesting that the majority of the stroke patient population was overweight. Lastly, the TARGET variable, which represents a binary outcome of recurrent stroke, had a mean value of 0.15, indicating a low incidence rate within 2023.

The logistic regression analysis revealed significant predictors with notable implications for clinical outcomes **(see figure 3)**. Age showed a negative coefficient (-0.0144, $p < 0.001$), indicating that the likelihood of recurrent stroke decreases with increasing age. The length of hospital stay also had a slight negative impact on recurrence rates (-0.0003, $p = 0.047$). Clinical severity scores such as the MRS discharge score (0.1053, $p < 0.001$) and the Arrival NIHSS score (0.0113, $p < 0.001$) were positively associated with stroke recurrence, suggesting that more severe initial symptoms are linked to higher risks of recurrence. IVTPA treatment was associated with a reduced likelihood of recurrent strokes (-0.3501, $p < 0.001$), highlighting the effectiveness of this treatment.

Significant predictors also included the absence of hypertension ($b = -0.982$, $p < 0.001$) and hyperlipidemia ($b = -1.100$, $p < 0.001$), both of which significantly predicted decreased risk of recurrent strokes. Similarly, the absence of chronic kidney disease predicted a significantly lower probability of stroke recurrence risk ($b = -0.5044$, $p < 0.001$). The impact of log-transformed BMI and log-transformed Length of Stay hours also emerged as significant, both showing negative associations with stroke recurrence, underscoring the complex relationships between physiological measures and health outcomes.

Model performance was assessed through Mean Squared Error (MSE), with logistic regression showing the lowest MSE (0.1094), suggesting it was the most effective model in predicting stroke recurrence compared to Lasso (MSE = 0.1094), Ridge (MSE = 0.1095), and

decision trees (MSE = 0.1277)(**see figure 1 and Table 2)**. This indicates that despite its simplicity, logistic regression was able to capture the essential dynamics of the predictors without overfitting.

The histogram of predicted probabilities (**figure 2)** from the logistic regression model provides a detailed view of the distribution of the risk of stroke recurrence as modeled on our test dataset. The histogram predominantly features probabilities clustered around lower values, indicating that for most patients, the model predicts a low likelihood of stroke recurrence. The highest frequency of predicted probabilities occurs between 0 and 0.25, with a peak near 0.1. This concentration at lower probabilities suggests that the model generally finds a lower risk of recurrence for the majority of cases.

The distribution is right-skewed, meaning there are fewer cases with higher predicted probabilities, although there is a small tail extending towards higher risk probabilities up to 1. This right skewness could indicate either an underlying feature of the dataset where most individuals are at a lower risk, or it might reflect the model's conservative prediction behavior, particularly if the model has been trained with an imbalanced dataset where fewer instances of recurrence are present.

Discussion

The goal of this report was to explore potential determinants of recurrent stroke risk in a clinical population. This goal was accomplished by applying logistic, lasso, and ridge regression models to a dataset containing stroke patient medical information. Based on our most optimal predictive model–our logistic model–hypertension, hyperlipidemia, high BMI, and increased length of stay predicted an increased likelihood of recurrent stroke risk. These findings are consistent with previous literature detailing risk factors of recurrent stroke[5].

Similarly, our findings support the observations in the literature that Black individuals are more likely to experience instances of recurrent stroke compared to other racial groups, highlighting a significant racial disparity present in our society. Interestingly, we found that the presence of diabetes mellitus type 2, but not type 1, predicted an increased risk of recurrent stroke. This is partially consistent with the extant literature, which has noted that certain stroke subtypes may result in recurrent stroke when DM is present. Finally, and notably, our model suggested that younger individuals (less than 50 years of age) were more likely to experience a recurrent stroke. This finding stands in contrast to literature suggesting that increases in age are a significant risk factor for both stroke and recurrent stroke. This could, perhaps, represent the circumstances of our sample or potential limitations in our model.

Conclusions

Taken together, our analysis underscores the importance of engaging in healthy lifestyle approaches following the incidence of a stroke to reduce the risk of recurrence. As we can see, factors associated with poor diets, stress, and a lack of activity appear to predict the occurrence of recurrent strokes. Thus, physicians should work with the patients to encourage reasonable lifestyle changes and appropriate choices of medication to reduce the presence of recurrent stroke risk factors. In addition, based on our model, this may be especially important for individuals who are younger.

Furthermore, our results highlight important racial disparities in recurrent stroke risk such that African American individuals have a higher likelihood of experiencing recurrent stroke.This could be a result of several factors, including discrimination, which has been shown to increase the risk of poor health outcomes[3] and the risk of cerebrovascular disease[2]. Thus, it is important for future research to explore preventative measures for recurrent stroke in racial and ethnic subgroups.

However, when interpreting these findings, It is important to recognize the limitations inherent in this design. For instance, it is possible that a more appropriate model could have been chosen to represent our data. Also, it is possible that there are additional features not present in our dataset that would be more capable of predicting the presence of a recurrent stroke.

References

1. Albright, K. C., Huang, L., Blackburn, J., Howard, G., Mullen, M., Bittner, V., ... & Howard, V. (2018). Racial differences in recurrent ischemic stroke risk and recurrent stroke case fatality. Neurology, 91(19), e1741-e1750.
2. Beatty Moody, D. L., Taylor, A. D., Leibel, D. K., Al-Najjar, E., Katzel, L. I., Davatzikos, C., ... & Waldstein, S. R. (2019). Lifetime discrimination burden, racial discrimination, and subclinical cerebrovascular disease among African Americans. Health Psychology, 38(1), 63.
3. Davis, B. A. (2010). Discrimination: A Social Determinant Of Health Inequities. *Health Affairs Forefront*.
4. Engel-Nitz, N. M., Sander, S. D., Harley, C., Rey, G. G., & Shah, H. (2010). Costs and outcomes of noncardioembolic ischemic stroke in a managed care population. Vascular Health and Risk Management, 6, 905-913.
5. Faraway, J. J. (2006). Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models. Chapman and Hall/CRC.
6. Faraway, J. J. (2014). Linear Models with R - 2nd Edition. Chapman and Hall/CRC.
7. Ferrone, S. R., Boltyenkov, A. T., Lodato, Z., O'Hara, J., Vialet, J., Malhotra, A., ... & Sanelli, P. C. (2022). Clinical outcomes and costs of recurrent ischemic stroke: A systematic review. Journal of Stroke and Cerebrovascular Diseases, 31(6), 106438.

8. ICD-10-CM Diagnosis Code I63.9: Cerebral infarction. (2024). Retrieved from https://www.icd10data.com/ICD10CM/Codes/I00-I99/I60-I69/I63-/I63.9
9. Kolmos, M., Christoffersen, L., & Kruuse, C. (2021). Recurrent ischemic stroke–a systematic review and meta-analysis. Journal of Stroke and Cerebrovascular Diseases, 30(8), 105935.
10. Sanelli, P., Yang, B., O'Hara, J., Jimenez, J., Gribko, M., Martinez, G., Feizullayeva, C., Adinarayan, K., Hoang, A., Qui, M., Wang, J., Boltyenkov, A., Sangha, K., Sanmartin, M., Elhabr, A., & Katz, J. (2023, April 11). Stroke health outcomes database. Retrieved April 11, 2023.
11. Shah, S., Liang, L., Kosinski, A., Hernandez, A. F., Schwamm, L. H., Smith, E. E., ... & Xian, Y. (2020). Safety and outcomes of intravenous TPA in acute ischemic stroke patients with prior stroke within 3 months: Findings from get with the guidelines–stroke. Circulation: Cardiovascular Quality and Outcomes, 13(1), e006031.
12. Singh, R. J., Chen, S., Ganesh, A., & Hill, M. D. (2018). Long-term neurological, vascular, and mortality outcomes after stroke. International Journal of Stroke, 13(8), 787-796.
13. Tsao, C. W., Aday, A. W., Almarzooq, Z. I., et al. (2022). Heart disease and stroke statistics—2022 update: a report from the American Heart Association. Circulation, 145, e153–e639. https://doi.org/10.1161/CIR.0000000000001052
14. WHO EMRO | Stroke, Cerebrovascular accident | Health topics. (2023). World Health Organization - Regional Office for the Eastern Mediterranean. Retrieved from https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html

Appendices:

Supplemental tables and/or figures.
R statistical programming code.