IJSA (2022), Vol. 6, No. 1, 161-169
DOI: https://doi.org/10.29244/ijsa.v6i1p161-169

Indonesian Journal of Statistics and Its Applications
Available at https://ijsa.stats.id
e-ISSN: 2599-0802

Binary Logistic Regression Model of Stroke Patients: A Case Study of Stroke Centre Hospital in Makassar*

Suwardi Annas[1‡], Aswi Aswi[1], Muhammad Abdy[2], and Bobby Poerwanto[3]

[1,3]Statistics Study Program, Universitas Negeri Makassar, Indonesia
[2]Departement of Mathematics, Universitas Negeri Makassar, Indonesia

‡corresponding author: suwardi_annas@unm.ac.id

Abstract

This paper aimed to determine factors that affect significantly types of stroke for stroke patients in Dadi Stroke Center Hospital. The binary logistic regression model was used to analyze the association between the types of stroke and some covariates namely age, sex, total cholesterol, blood sugar level, and history of diseases (hypertension/stroke/diabetes mellitus). Maximum Likelihood Estimation was used to estimate parameters. Combinations of covariates were compared using goodness-of-fit measures. Comparisons were made in the context of a case study, namely stroke patients (2017-2020). The results showed that a binary logistic model combining the history of diseases and blood sugar level provided the most suitable model as it has the smallest AIC and covariates included are statistically significant. The coefficient estimation of the history of diseases variable is -0.92402 with an odds ratio value exp (-0.92402)=0.4. This means that stroke patients who have a history of diseases experience a reduction of 60% in the odds of having a hemorrhagic stroke compared to stroke patients that do not have a history of diseases. In other words, stroke patients who have a history of diseases tend to have a non-hemorrhagic stroke. Furthermore, the coefficient estimation of blood sugar level is 0.74395 with an odds ratio value exp(0.74395)=2. It means that stroke patients who do not have normal blood sugar levels tend to have a hemorrhagic stroke 2 times greater than stroke patients with normal blood sugar levels. A history of diseases and blood sugar level were factors that significantly affect the types of stroke.

Keywords: hemorrhagic stroke, logistic regression, non-hemorrhagic stroke.

* Received: Apr 2022; Reviewed: Apr 2022; Published: May 2022

1. Introduction

Stroke, cardiovascular disease, is the second leading of death after heart disease with 5.8 million fatal cases annually (Abubakar & Isezuo, 2012; Cai et al., 2019; Truelsen et al., 2007). Strokes accounted for 10% of all total deaths worldwide in 2016. 40% of strokes occur in people who are less than 70 years old. Approximately 15 million new acute strokes occur annually which is two-thirds of these people live especially in developing countries with low income and middle income (Truelsen et al., 2007). In Indonesia, about a third of total deaths are caused by cardiovascular diseases with coronary heart and stroke disease being the leading causes of death (Hussain et al., 2016). Stroke prevalence in rural Indonesia is 0.0017%, in urban Indonesia is 0.022%, 0.5% among urban Jakarta is adults, and overall is 0.8% (Kusuima et al., 2009).

Types of stroke are non-hemorrhagic (ischemic) and hemorrhagic (Hussain et al., 2016). Non-hemorrhagic stroke is a type of stroke that occurs due to the blockage of a blood vessel in the brain. A non-hemorrhagic stroke also called an ischemic stroke, is the most common type of stroke. It is estimated that more than 80% of stroke cases worldwide are caused by non-hemorrhagic stroke.

Regression analysis is a set of statistical procedures for describing the association between the response variable and one or more explanatory variables (covariates). In a linear regression model, it is usually assumed that the dependent variable is continuous. It is not uncommon that the dependent (outcome) variable is discrete with two or more categories (Hosmer et al., 2013). A logistic regression model is a regression analysis used to assess the relationship between categorical response variables and one or more predictor variables which can be either categorical or continuous variables. Logistic regression analysis is often used in epidemiological research, namely the study of patterns of occurrence of disease and the factors that influence it. The logistic regression is distinguished based on the number of categorizations of the response variables, namely binary logistic regression (dichotomous), multinomial regression (polytomous), and ordinal logistic regression (Hosmer et al., 2013).

In this study, we used the binary logistic regression model to analyze the association between the types of stroke and some covariates, namely age, sex, total cholesterol, blood sugar levels, and the history of diseases (hypertension/stroke/diabetes mellitus) for stroke datasets in a Dadi Stroke Center Hospital in Makassar, Indonesia. Although stroke has been the subject of considerable research effort, there appears to have been little research so far into modeling the types of stroke including non-hemorrhagic and hemorrhagic strokes.

Several studies have used logistic regression for modeling stroke (Cai et al., 2019; Han et al., 2019; Kim & Kim, 2018; Yang et al., 2020). Binary logistic regression has been used to investigate factors that affect antidepressant prescriptions with acute ischemic stroke (Kim & Kim, 2018). They found that the length of stay and the mechanical ventilation usage were associated with escitalopram prescriptions. Logistic regression was used to evaluate factors related to knowledge and poor understanding of cardiovascular disease (stroke and heart attack symptoms) (Han et al., 2019). They concluded that poor understanding of cardiovascular disease warning signs was related with the lower education level,

male gender, older age, lack of physical activity, unemployment, unmarried, poor economic, and poor psychological status, poor health behaviors, and the presence of hypertension. Another study used binary logistic regression models to investigate the relationship between different dimensions of physical activity (frequency, volume, duration, intensity) and stroke risk (Yang et al., 2020). They found that the different dimensions of moderate regular exercise are significantly associated with stroke risk. Multilevel logistic regression models and zero-inflated Poisson were used to investigate factors associated with social participation among stroke survivors (Cai et al., 2019). They stated that communicating with friends and going to social clubs to play games were the most popular social activities. Furthermore, limited social participation causes a high risk of stroke. However, these papers implemented only one model and focused only on important factors associated with stroke. These papers did not use both two types of stroke (non-hemorrhagic and hemorrhagic stroke). None of these papers appeared to compare different combination models with the inclusion of some covariates. The overall objective of this study is to compare different combination models and determine factors that affect significantly the types of stroke for stroke patients in Dadi stroke center hospital. Combinations of covariates were compared using goodness-of-fit measures, such as *Akaike Information Criterion* (AIC).

2. Methodology

2.1 Study Area

Data on admitted stroke were gathered from patient medical records from a South Sulawesi Regional Special Hospital (Stroke Center), namely Dadi Hospital in Makassar, South Sulawesi Province. The data include inpatient ischemic stroke (non-hemorrhagic stroke) and hemorrhagic stroke cases admitted from 2017 to 2020. The initial stroke patient from Stroke Centre Hospital comprised 329 patients. Of these, there were 133 (40.4%) stroke cases with incomplete clinical variables. As a result, the final stroke patient dataset, 196 patients (59.6%), was included in the analysis.

The information collected was stroke type which consists of two types, namely non-hemorrhagic (0) and hemorrhagic (1), sex, age, total cholesterol, history of diseases, and blood sugar level. The response variable is the type of stroke. Only patients who stay overnight were included in the study. Any problems found were justified by the medical record officer. The dataset was cleaned before performing the analysis. The explanation of the variables used in this study is given in Table 1.

2.2 Research Method

The binary logistic regression model was used to analyze the association between the types of stroke and some covariates. Age, sex, total cholesterol, blood sugar level, and history of diseases were included as covariates. This model allows the variable outcomes (stroke types) to be categorized into two categories, namely non-hemorrhagic and hemorrhagic. We computed the odds of the event versus no event.

In general, the binary logistic regression model is given as follows:

$$p(x_i) = \frac{\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_i x_i)}{1+\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_i x_i)} \quad (1)$$

Table 1. The variables used in the study

Variables	Description
Stroke type (Y)	0: Non-hemorrhagic 1: Hemorrhagic
Sex (X1)	0: Male 1: Female
Age (X2)	year
Total Cholesterol (X3)	0: Normal (< 200 mg/dl) 1: Not normal
History (X4)	0: no history of the disease 1: there is a history of hypertension/stroke/diabetes mellitus
Blood Sugar Levels (X5)	0: Normal (70-100) mg/dl 1: Not normal

where $p(x_i)$ is the probability of the binary outcome being present.
The logit function is its linearizing transformation from equation (1) and is given as follows:

$$g(x) = \ln\left(\frac{p(x_i)}{1-p(x_i)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i \quad (2)$$

where $g(x)$ is the logit also known as the log odds of the outcome variable. The term β_0 is the intercept that reflects the log odds or logit estimate of the response variable when model independents are evaluated at zero. There is one intercept estimate in binary logistic regression and there are *i* independent variables. The $\beta_1, \beta_2, \ldots \beta_i$ are the logistic regression coefficients. Exp(β) is the "odds ratio" for an explanatory variable or the natural log base *e* raised to the power of β. The odds ratio of an explanatory variable is the factor by which the explanatory variable increases (if positive) or decreases (if negative) the log odds of the dependent variable. Binary logistic models were run in R version 4.1.2 (R Core Team, 2019).

The analysis used in this study is started with a descriptive analysis of both dependent and independent variables, then followed by inferential statistics. The selection of the best model was evaluated based on the AIC value (Akaike, 1974) with the inclusion of the statistically significant covariates. AIC is a method that can be used to select the best logistic regression model found by Akaike and Schwarz. According to the AIC method, the best regression model is the regression model that has the smallest AIC value. To address the second aim, for the selected model, variables were considered to be important if the p-value is less than or equal to 0.05.

3. Results

3.1 Descriptive analysis

Dadi Stroke Center Hospital is situated in the Mamajang district, Makassar, South Sulawesi Province, Indonesia. The number of non-hemorrhagic stroke patients and hemorrhagic stroke patients is 101 (52%) and 95 (48%) patients respectively. The

number of stroke patients who have no history of diseases (107 patients) is more than the number of stroke patients who have a history of diseases (89 patients). The number of non-hemorrhagic stroke patients of males (56 patients) is more than females (45 patients). On the other hand, the number of hemorrhagic stroke patients in males (44 patients) is less than in females (51 patients). 50% of stroke patients have less than 60 years old. The types of stroke disease distribution based on sex are given in Figure 1.

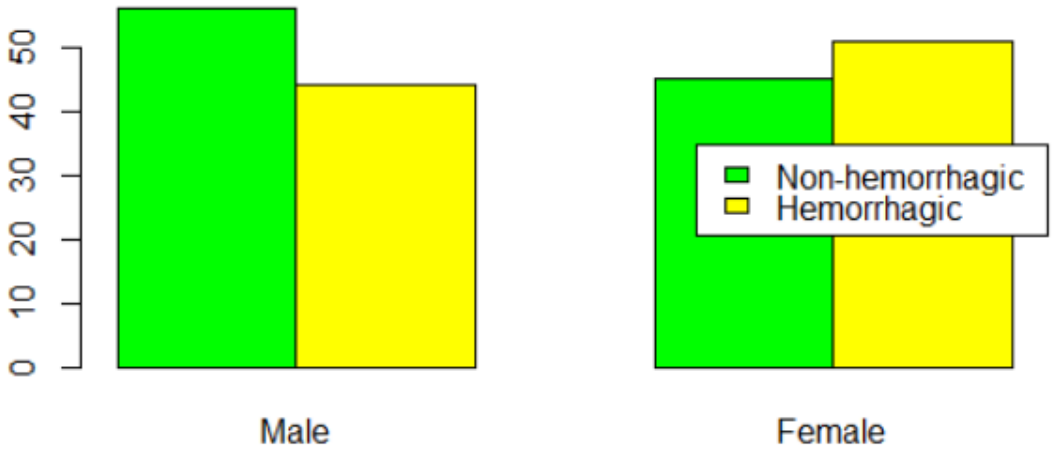

Figure 1. The types of stroke disease distributed based on sex

3.2 Model comparison using Binary Logistic Regression Models

The results of different combinations of covariates for the stroke dataset from 2017 to 2020 as well as the AIC values of possible combination models are given in Table 2. We started for full models, that is, the model with the inclusion of all covariates (age + sex + total cholesterol + blood sugar level + history of diseases) (Model 1). Then, the parameter significance test (overall test) is carried out to see the overall parameter coefficient on the dependent variable. The null hypothesis test is all parameters = 0. In another word, there is no independent variable that influences the dependent variable. *G* is the test statistics used for the likelihood ratio test for simultaneous testing which equals minus two times the difference in the log-likelihoods. Based on the results of simultaneous testing, it shows that the value of $G = 12.79 >$ chi-square table = 11.07. Therefore, the conclusion is to reject the hypothesis null which means that at least one parameter is not equal to zero. In another word, there is at least one independent variable that influences the dependent variable.

Based on Table 2, it can be concluded that among these 12 models, Model 4 with the inclusion of history of diseases and blood sugar level has the smallest AIC (182.72) and all covariates included in the model have a p-value less than 0.05.

Table 2. The AIC values for all combination models

Models	Covariates	AIC
1	Sex + age + total cholesterol + history* +blood sugar level*	187.63
2	Sex + total cholesterol + history* +blood sugar level*	185.65
3	Total cholesterol+ history* + blood sugar level*	183.82
4	History* + blood sugar level*	**182.72**
5	History*	184.81
6	Blood sugar level*	187.53
7	Total cholesterol	192.16
8	Total cholesterol + history*	186.27
9	Total cholesterol + blood sugar*	188.96
10	Sex +blood sugar level*	189.37
11	Sex + history*	186.30
12	Age + history*	186.73

*95% posterior confident interval for the coefficient does not contain zero.

The partial testing for the full model (Model 1) is carried out to determine the effect of each independent variable on the dependent variable using the Wald test. If the p-value < alpha 5%, then the null hypothesis is rejected which means that the logit coefficient is significant to the model. The partial test of full models (Model 1) is given in Table 3.

Table 3. The partial test of full models (Model 1)

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-0.093700	1.060325	-0.088	0.92958
Sex (Female)	0.155783	0.376995	0.413	0.67944
Age	0.002246	0.015887	0.141	0.88756
Total cholesterol (1)	-0.361708	0.373507	-0.968	0.33284
History (1)	-0.948185	0.362150	-2.618	0.00884 **
Blood Sugar level (1)	0.749014	0.384043	1.950	0.05114 .

From Table 3, we can see that covariates that significantly affect the types of stroke are history of diseases and blood sugar level. This is in line with the results of AIC values (see Table 2) which showed that the model with the inclusion of covariates history of diseases, and blood sugar level provided the best model in terms of AIC values. Therefore, the preferred model in modeling the types of stroke is Model 4 with the inclusion of covariates of history of diseases and blood sugar level. The partial test of the preferred model is given in Table 4.

Table 4. The partial test of Model 4

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-0.06399	0.34526	-0.185	0.85295
History(1)	-0.92402	0.35853	-2.577	0.00996 **
Blood Sugar level (1)	0.74395	0.37153	2.002	0.04524*

Based on Table 4, it can be concluded that the coefficient estimation of the variable of history of diseases is -0.92402 with the reference category being no history of the disease (0). The odds ratio value is exp (-0.92402) = 0.4 which means

that stroke patients who have a history of diseases experience a reduction of 60% in the odds of having a hemorrhagic stroke compared to stroke patients that do not have a history of diseases. In other words, stroke patients who have a history of diseases tend to have a non-hemorrhagic stroke. Furthermore, the coefficient estimation of the variable of blood sugar level is 0.74395 with the reference category being normal (0). The odds ratio value is exp (0.74395) = 2 which means that stroke patients who do not have normal blood sugar levels tend to have a hemorrhagic stroke 2 times greater than stroke patients with normal blood sugar levels.

3.3 Discussion

We have presented a binary logistic regression model to analyze the association between the types of stroke and some covariates (age, sex, total cholesterol, blood sugar level, and the history of diseases). In this study, we also presented the comparison of the combinations of covariates using the goodness-of-fit measure namely AIC in the context of a case study of stroke patients and considered the significance of the included covariates. Based on the AIC and the significance of the included covariates, a binary logistic model with the inclusion of blood sugar level and history of diseases is preferred.

Our results found that blood sugar level and history of diseases are important covariates that affect the types of stroke. Diabetes is a chronic disease characterized by elevated levels of blood sugar. These results are in line with the previous research that found that diabetes is a crucial modifiable risk factor for especially ischemic strokes and it elevates ischemic stroke incidence for all age groups (Chen et al., 2016) Another study reported a somewhat similar result to that found in our study (Elnady et al., 2020; Krajicek, 2016). They found that approximately 50% of patients who survive from an ischemic stroke have an elevated risk of recurrent stroke within a few days or weeks of the initial insult where the first week is the highest risk. On the other hand, we found that the total cholesterol does not significantly affect the types of stroke. This result is in line with the other research that stated that Cholesterol levels are inconsistently correlated with the risk of hemorrhagic stroke (Wang et al., 2011).

To our knowledge, we provide the first report using the binary logistic regression models on modeling the stroke types using the combination covariates in stroke data in Dadi stroke center hospital in Makassar, Indonesia. However, our results only used data from 2017 to 2020 and some covariates. We acknowledge that adding some more data and some other covariates as well as using other models may have different results. Our future work aims to add some more data, some covariates, and used other models such as Kernel logistic regression model. Another limitation of this study is that the estimation of parameters only used one method namely the Maximum Likelihood estimation (MLE). It is acknowledged that using other estimation methods such as the Bayesian method may affect the results. Our future work aims to compare the MLE method and the Bayesian method.

4. Conclusions

In summary, the key finding is that different combination models provided different results for this case study. Factors that significantly affect the types of stroke are blood sugar level and history of diseases. Stroke patients who have a history of diseases experience a reduction of 60% in the odds of having a hemorrhagic stroke compared to stroke patients that do not have a history of diseases. Stroke patients who do not have normal blood sugar levels tend to have a hemorrhagic stroke two times greater than stroke patients with normal blood sugar levels. Factors that affect the types of stroke patients need to be tackled by adopting a healthier lifestyle. The MLE method was used to estimate the parameter in the model which may have different results when we use other methods such as the Bayesian method. The comparison of the MLE method and the Bayesian method could be a potential future work.

Acknowledgments. The authors acknowledge the Dadi stroke center hospital, South Sulawesi Province for providing the stroke dataset. The authors also acknowledge the Kemdikbudristek. This work was supported by Kemdikbudristek through the PDUPT research grant scheme.

References

Abubakar, S., & Isezuo, S. (2012). Health related quality of life of stroke survivors: experience of a stroke unit. *International Journal of Biomedical Science*, *8*(3): 183--187.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6): 716–723.

Cai, Y., Towne, S. D., & Bickel, C. S. (2019). Multi-Level Factors Associated with Social Participation among Stroke Survivors: China's Health and Retirement Longitudinal Study (2011–2015). *International Journal of Environmental Research and Public Health*, *16*(24): 5121.

Chen, R. M., Ovbiagele, B. M., & Feng, W. M. (2016). Diabetes and stroke: epidemiology, pathophysiology, pharmaceuticals and outcomes. *The American Journal of the Medical Sciences*, *351*(4): 380–386.

Elnady, H. M., Mohammed, G. F., Elhewag, H. K., Mohamed, M. K., & Borai, A. (2020). Risk factors for early and late recurrent ischemic strokes. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery*, *56*(1): 1–7.

Han, C. H., Kim, H., Lee, S., & Chung, J. H. (2019). Knowledge and poor understanding factors of stroke and heart attack symptoms. *International Journal of Environmental Research and Public Health*, *16*(19): 3665.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Hussain, M. A., Mamun, A. A., Peters, S. A., Woodward, M., & Huxley, R. R. (2016). The Burden of Cardiovascular Disease Attributable to Major Modifiable Risk Factors in Indonesia. *Journal of Epidemiology*, *26*: 515–521. https://doi.org/10.2188/jea.JE20150178

Kim, J., & Kim, Y. (2018). Which Patients Are Prescribed Escitalopram?: Predictors for Escitalopram Prescriptions and Functional Outcomes among Patients with Acute Ischemic Stroke. *International Journal of Environmental Research and Public Health*, *15*(6): 1085. https://doi.org/10.3390/ijerph15061085

Krajicek, S. (2016). Prediction of Early Recurrence After Acute Ischemic Stroke: Arsava EM, Kim GM, Oliveira-Filho J, et al. JAMA Neurology. 2016; 73 (4): 396-401. *Journal of Emergency Medicine*, *51*(1): 91–92.

Kusuima, Y., Venketasubramanian, N., Kiemas, L., & Misbach, J. (2009). Burden of stroke in Indonesia. *International Journal of Stroke*, *4*(5): 379–380.

R Core Team. (2019). *R: A language and environment for statistical computing*. Austria: R Foundation for Statistical Computing.

Truelsen, T., Heuschmann, P. U., Bonita, R., Arjundas, G., Dalal, P., Damasceno, A., Nagaraja, D., Ogunniyi, A., Oveisgharan, S., Radhakrishnan, K., Skvortsoya, V. I., & Stakhovskaya, V. (2007). Standard method for developing stroke registers in low-income and middle-income countries: experiences from a feasibility study of a stepwise approach to stroke surveillance (STEPS Stroke). *The Lancet Neurology*, *6*(2): 134--139.

Wang, X., Guo, P., & Huang, X. (2011). A review of wind power forecasting models. *Energy Procedia*, *12*: 770–778.

Yang, D., Bian, Y., Zeng, Z., Cui, Y., Wang, Y., & Yu, C. (2020). Associations between intensity, frequency, duration, and volume of physical activity and the risk of stroke in middle-and older-aged Chinese people: a cross-sectional study. *International Journal of Environmental Research and Public Health*, *17*(22): 8628.