Prediction Of Mortality Using Logistic Regression Analysis In Female Patients Suffering From Myocardial Infarction

Geliza Marie I. Alcober, Ace C. Lagman, Teodoro F. Rivano Jr., Rossana T. Adao

Abstract— This study aims to generate a predictive model that can be used in determining the mortality of female patients suffering from Myocardial Infarction (MI) or simply heart attack. Predictive modeling is now considered as one of the best techniques in identifying the outcome of a certain event by determining patterns based on the different parameters that are considered significant. This paper includes several steps in developing a Logistic Regression model based on patient's age, previous Myocardial Infarction event, smoking status, if a patient has diabetes or high blood pressure, or if the patient has previously experienced stroke. This logistic regression model could be applied in addressing health-related issues like formulating a prognostic guide for practicing a more personalized way of treating patients who have experienced heart attack. Generating a medical prediction model is truly time-consuming and cumbersome however through applying predictive modeling in medicine, medicine professionals can now formulate better ways to accommodate patients especially women who have a history of Myocardial Infarction. This study can identify what are the possible characteristics that made up the clusters or groups of Myocardial Infarction patients that survived and characteristics that made up the clusters or groups of MI patients. The accuracy rate according to the given values attained the value of 0.805 or 81%. Therefore, the logistic regression model can be implemented and embedded to a decision-based software application.

Index Terms— Logistic Regression, Myocardial Infarction, Predictive Model, Heart Attack, Regression Analysis.

1 INTRODUCTION

ANALYTICS bridges the gap of what is unknown and an assumption that is based on historical data. Through all the several predictive modeling techniques available today, any organization can now answer the question, what will happen? Allowing any field like medicine, create prognostic guides that can enhance personalized medications. Medicine professionals can now gain meaningful insights from the different data available in the field in order to develop new technologies for curing several diseases. At present, predictive models are shaping the way organizations accommodate the needs of the people. Myocardial Infarction (MI) or also known as heart attack happens when fatty deposits build up over time blocking the vein that supplies the heart with blood and oxygen. Through the years, heart attack is considered as an old man disease because older men have more historical records of experiencing heart attack than women. However, doctors are starting to see the increasing number of women who are experiencing heart attack, and the main factor that the doctors is now considering is stress. Women are usually doing several part-time jobs to fulfill the needs of their families. There are also few researchers that identified that the age of female patients who are experiencing heart attack is now decreasing, therefore younger women are not excused from this disease. This research study also aims to raise awareness for women to take good care of themselves to avoid heart attack – a disease that is considered traitor because of its unpredictable nature. Not knowing what might happen in the future is a big disadvantage for anyone especially in relation with health. Through utilizing predictive modeling in identifying what might happen next to the journey of a patient is big help to avoid undesirable events. Hospitals can now make use of patients' information in identifying patterns that can be a big help in treating diseases. This research paper used the data from Newcastle (Australia) centre of Monica project with a total of nine risk indicators that were tested in order to build a logistic regression model. The dataset downloaded from Rdatasets website contained 509 records of patients. The World Health Organization (WHO) released the dataset in the early 1980s as part of the MONICA (Multinational MONItoring of trends and determinants in Cardiovascular disease) Project. Heart disease is often thought to be more of a problem for men. However, it's the most common cause of death for both women and men in the United States. Because some heart disease symptoms in women can differ from those in men, women often don't know what to look for. Fortunately, by learning their unique heart disease symptoms, women can begin to reduce their risks [10]. The MONICA project aims to see trends in cardiovascular diseases. The data provided are still currently being analyzed to further improve ways of treating diseases mainly myocardial infarction and stroke and enhance the understanding of the population who have experienced the said diseases.

2 LITERATURE REVIEW

Predictive modeling is defined as the process of utilizing both data and statistics to predict outcomes with data models [4]. Through this process, medicine practitioners can now have a basis of what possible actions to perform in order to help female patients suffering from myocardial infarction survived.

Logistic regression is used to obtain a predicted outcome based on several explanatory variables. The underlying method of logistic regression is comparable to linear regression. Both linear and logistic regressions are well known techniques in solving classification problem or a problem that deals with a dependent variable that is categorical in nature.

Several data have been acquired and processed with a view to generate reliable predictions from the different formulas, algorithms, and models developed. The patient is considered as the top beneficiary of the predictive models to be generated. Medicine practitioners will be able to map out the root causes of the disease like myocardial infarction itself, as well as, map out the patterns that make the patient's mortality rate even higher [1]. The most frequently reported symptoms didn't include chest pain. Instead, women reported unusual fatigue, sleep disturbances, and anxiety. Nearly 80 percent reported experiencing at least one symptom for more than a month before their heart attack [11]. There are 11 factors discovered on the study [2] conducted by Mayo clinic. According to the article, the narrowing of arteries throughout a person's body are caused of the fatty deposits that have accumulated through time.

These fatty deposits (atherosclerosis) are caused of the following risk factors:

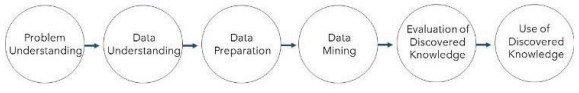
- Age
- Tobacco
- High blood pressure
- High blood cholesterol
- Obesity
- Diabetes
- Metabolic Syndrome
- Family history of heart attack
- Lack of physical activity
- Stress
- Illicit drug use
- A history of preeclampsia
- An autoimmune condition

The article released by Mayo clinic identified that older men and women are prone to experience myocardial infarction. Obesity, which is also caused of lack of physical activity, can also be the root of high blood pressure, an incident that causes damage to a person's heart. Smoking also contributes to the condition of the heart. People with high sugar levels have higher risk to experience heart attack. According to [6], in a health care system where the predominant subjects are men, women had better short- and long-term survival than men after an acute myocardial infarction. Multiple risk factors were predictive of greater perceived diabetes risk, whereas, only family history of heart attack, high blood pressure and increases in BMI significantly contributed to perceived risk of heart attack among ethnically diverse at risk middle-aged adults [7]. Metabolic syndrome is also considered as a risk factor for attaining diabetes or experiencing stroke or even heart attack. Having a history about heart attack in your family also depicts a higher chance of experiencing heart attack. All the factors mentioned above were taken to considerations in producing a predictive model for this study.

3 METHODOLOGY

The techniques and analysis used in this study are discussed in this section. For the purpose of formulating a model with high accuracy to predict the mortality of female patients suffering from myocardial infarction, the researchers used Knowledge Discovery in Databases (KDD) as its reference. KDD is referred to as a systematic process which highlights the key steps including problem understanding, modeling, evaluation, and deployment.

Fig. 1.



Knowledge Discovery in Database Model

Figure 1 indicates the six different phases that narrate the process of developing this study. The six-step model is done through understanding the problem, performing analysis on the data gathered, transforming those data into a more consistent format so that the data can be easily used in building models, evaluating the model, and through applying the model generated. KDD must be done thoroughly in order to avoid false analysis.

A. Problem Understanding

The first step of KDD is to define the main objective of performing the analyzation. It is the phase where the researchers collected enough data in order to formulate the problem that needs to be addressed.

B. Data Understanding

This step involves the selection of significant parameters that could be utilized to attain the predictive model that can predict with higher accuracy rate.

TABLE 1
DESCRIPTION OF DATASET FEATURES

Attributes	Descriptions	Values	Feature
Outcome	Mortality Outcome	Live dead	Nominal
Age	Age at onset	40+	Numeric
Yronset	Year of onset	1985	Numeric
premi	Previous Myorcardialin farction event	A factor w/ levels	Nominal
smstat	Smoking Status	C – current x-ex smoker n – non- smoker	Nominal
diabetes	A factor w/ levels	Y – Yes N - No	Nominal
highbp	High blood pressure	Y – Yes N - No	Nominal
hichol	High cholesterol	Y – Yes N - No	Nominal
Angina	A factor with levels	Y – Yes N - No	Nominal
Stroke	A factor level	Y – Yes N - No	Nominal

The attribute outcome is the target variable, the target variable is said to be dichotomous, with value either 0 or 1. The value of the outcome is based on the constant for each significant attribute that SAS tools generated. It has eight attributes namely: (a) age; (b) previous Myocardial Infarction event; (c) smoking status; (d) diabetes; (e) diabetes; (f) high blood pressure; (g) high cholesterol; (h) angina; and stroke, contribute to the development of heart diseases specifically heart attacks. It examines one-way frequencies of selected variables.

C. Data Preparation

The third step in the KDD model is considered as the hardest step. Data preparation requires a lot of modifications in order to select the set of parameters that could produce a model with higher accuracy.

TABLE 2
TRANSFORMED DATASETS FEATURES

Attributes	Descriptions	Values	Legend

Outcome	Mortality Outcome	Live dead	1
			0
premi	Previous Myorcardial infarction event	Yes	1
		No	2
smstat	Smoking Status	C – current	1
		x- ex smoker	2
		n – non-smoker	3
diabetes	A factor w/ levels	Y – Yes	1
		N - No	2
highbp	High blood pressure	Y – Yes	1
		N - No	2
hichol	High cholesterol	Y – Yes	1
		N - No	2
Angina	A factor with levels	Y – Yes	1
		N - No	2
Stroke	A factor with levels	Y – Yes	1
		N - No	2

In order to improve the accuracy of the logistic model, the researchers transformed some information from the original dataset downloaded from WHO (see Table 2). Since logistic regression deals with binary as dependent variable, the outcome variable is categorized as 0 or 1. If the patient is identified as 0, it signifies the algorithm that the patient is dead and if the patient is identified as 1, it means that the patient was able to survive previous myocardial infarction. There are seven categorical attributes that contribute to the disease, and in order to produce a mathematical formula, categorical attributes were transformed into numerical distribution and put a scale for each attribute.

D. Data Mining

The researchers performed the analysis using SAS Enterprise Guide (EG).

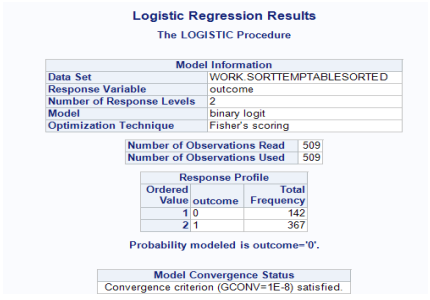


Fig. 2. Logistic Regression results for Female Patients

Figure 2 displays the partial results generated after applying logistic regression analysis to the dataset which is consists of records for female patients. SAS EG was able to successfully identify the number of target levels, the response variable, and the total frequency computed per each response level. As observed in figure 2, there are more female patients who survived from myocardial infarction than those who didn't.

4 RESULTS AND DISCUSSION

4.1 Logistic Regression Model

A logit model is another term for logistic regression. It is a method preferably used in dealing with data with dichotomous outcome or for those data with a target variable with a binary data type. The table below displays the results generated by SAS Enterprise Guide (EG). Everything starts with the concept of probability. Let's say that the probability of success of some event is .8. Then the probability of failure is 1 – .8 = .2. The odds of success are defined as the ratio of the probability of success over the probability of failure. In our example, the odds of success are .8/.2 = 4. That is to say that the odds of success are 4 to 1. If the probability of success is .5, i.e., 50-50 percent chance, then the odds of success is 1 to 1 [15].

TABLE 3
ANALYSIS OF MAXIMUM ESTIMATES

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	4.8495	1.4342	11.4339	0.0007
age	1	0.0388	0.0196	3.9019	0.0482
premi	1	-1.3902	0.2540	29.9487	<.0001
smstat	1	-0.3531	0.1469	5.7796	0.0162
diabetes	1	-1.7968	0.2481	52.4638	<.0001
highbp	1	-0.2909	0.2814	1.0685	0.3013
highcol	1	-0.5439	0.2449	4.9320	0.0264
angina	1	0.0968	0.2636	0.1348	0.7135
stroke	1	-0.7402	0.2818	6.9015	0.0086

The table above helps in determining the significant parameters that will be used for formulating the logit model. As observed, the table identifies the coefficients (labeled as Estimate) and the other columns together with the calculated p-values. SAS sets the default of 0.05 as the significant level, meaning if the coefficient of the variable is greater than 0.05, the variable isn't significant to the model, and if the variable has a coefficient less than or equal to 0.05, it is considered as significant. Therefore, all the parameters except angina, with a coefficient equal to 0.09, are considered as statistically significant. Women who have historical records of MI can be guided by considering the parameters that were successfully identified as significant by logistic regression, those factors are the age, previous myocardial event, smoking status, diabetes record, if the patient has high blood pressure, due to high cholesterol, and had experienced stroke.

The logit model coefficients that are shown on table 3 describes the change of log odds in every unit increase of independent variable.

- If the variable **age** changes with one unit, the odds of surviving myocardial infarction (versus not surviving MI) increases by 0.038.
- If the variable **premi** changes with one unit, the odd of surviving myocardial infarction (versus not surviving MI) increases by -1.39.
- If there's a unit change in **smstat**, the odds of surviving myocardial infarction (versus not surviving MI) increases by -0.35.
- If **diabetes** changes with one unit, the odd of surviving myocardial infarction (versus not surviving MI) increases by -1.17.
- If **highbp** changes with one unit, the odd of surviving myocardial infarction (versus not surviving MI) increases by -0.03.

- If the variable **highcol** changes with one unit, the odd of surviving myocardial infarction (versus not surviving MI) increases by -0.54.
- For a unit change in stroke, the odd of surviving myocardial infarction (versus not surviving MI) increases by -0.74.

The logit model generated by SAS Enterprise Guide can also be interpreted as,

$$P=\frac{1}{1+e^{(\hat{b}_0+0.0388x_1-1.3902x_2-0.3531x_3-1.7968x_4-0.2909x_5-0.5439x_6-0.7402x_7)}} \quad (1)$$

Equation 1 defines the predictive model generated by SAS EG. The result of the logit model is no longer surprising, all attributes present in the model are truly identified as risk indicators for MI. The assumption is that when a patient has a diabetes, suffers from high blood pressure, has a high cholesterol, or suffered from stroke has a high probability to die, consequently, patient with this type of record must have a full-time monitor and high level of treatment. Medical practitioners could use this model in order to predict the mortality of the patient by simply substituting the values for the x_n which is according to the significant parameters shown in Table 3.

Effect	Point Estimate	95% Wald Confidence Limits	
age	1.040	1.000	1.080
premi	0.249	0.151	0.410
smstat	0.703	0.527	0.937
diabetes	0.166	0.102	0.270
highbp	0.748	0.431	1.298
highcol	0.581	0.359	0.938
angina	1.102	0.657	1.847
stroke	0.477	0.275	0.829

TABLE 4
ODDS RATIO ESTIMATES

The table 4 shows the Exp(Est) show the estimated multiplicative effect of the corresponding IVs on the estimated odds, controlling for the other predictors [Jaccard, 2001]. The odds ratio for **age** of 1.040 signifies than an increase of 1 point on the scale measuring age increases the odds characterized to have an outcome 0 is by 1.040 times. For the variable **premi** which signifies whether the patient has a past record of heart attack, the odds of 0.249 times. The variable **smstat** that connotes for the smoking status of a patient has an equivalent odd of 0.703. The log odd ratio for variable diabetes is 0.166. The variable **highbp** has 0.748 odd ratio. The variable **highcol** has 0.581 times to come up with an outcome 0 or dead. The parameter **stroke** that identifies whether a patient has experienced stroke has 0.477 times log odd of dying.

4.2 Model Validation

A logistic regression model has been built and the coefficients have been evaluated. Nonetheless, the critical question remains, how good is the model generated? And how accurate is the prediction? The rest of this section will discuss how to validate the accuracy of the logit model generated. There are various critical metrics that deal with how the model contributes in predicting the output of the target variable. SAS Enterprise Guide offers different techniques in testing the accuracy of the model such as Likelihood Ratio Test, Hosmer-LemeshowTes, Wald Test, ROC Curve, and Classification test. Likelihood Ratio Test is done through comparing the likelihood of the data under the full model against the probability of the data under a model with fewer indicators.

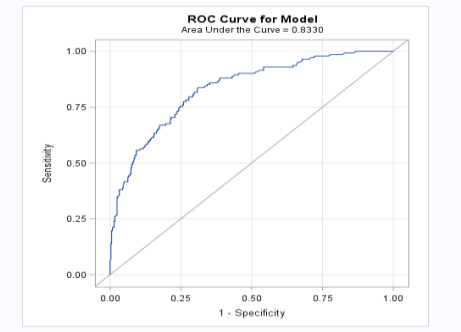


Fig.3.ROC Curve of the generated Logit Model

Figure 3 shows the area under the ROC curve which is equal to 0.8330. The rule of thumb says that the closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. A Receiver Operating Characteristic Curve (ROC) is a method mostly used to summarize the performance over a range of trade-offs between true positive (TP) and false positive (FP). In dealing with logistic regression models, as shown in Figure 3, the curve is equivalent to 0.8330 which indicates that the predictive power of the logit model is good.

$$Accuracy = \frac{TN + TP}{TP + FP + TN + FN} \quad (2)$$

Equation 2 describes another way of validating the logit model generated by SAS Enterprise guide. Equation 2 is most commonly used to determine the accuracy of the algorithm being used in any analysis. The true positive (TP) refers to the number of outcomes that are correctly predicted in terms students with regular status while true negative (tn) refers to the number of outcomes that are correctly predicted in terms of students with irregular status. TN represents the number of instances that the model's predicted results matches the observation but negatively identified. FN indicates the number of predicted results that is actually positive according to the standards. FP indicates the number of predicted outcomes with the false positive result that is actually negative. After processing the dataset in SAS Enterprise Miner, below are the values for each variable:

Event Classification Table

Data Role=TRAIN Target=outcome Target Label=' '

False Negative	True Negative	False Positive	True Positive
31	74	68	336

If the values generated from the tool will be substituted to the

formula previously identified, this will be the output,

$$Accuracy = \frac{336 + 74}{336 + 74 + 68 + 31} = 0.805$$

Based on the given results, 336 instances were positively identified by the model create out of 509 observations. The negatively identified instances is equal to 74. False positive instances have a total of 68 and 31 instances were identified as false negative. The accuracy rate according to the given values attained the value of 0.805 or 81%. Therefore, the logistic regression model formulated through SAS Enterprise Guide is a good model.

5 CONCLUSION

Using the historical data of the female patients suffering from Myocardial Infraction is useful in developing a predictive model. A multiple logistic regression model, relating age, previous Myocardial Infarction event, smoking status, diabetes, diabetes, high blood pressure, high cholesterol, and stroke to predict the patient's suffering from MI probability of surviving, was estimated. The area of ROC curve is equivalent to 0.8330. The calculated accuracy rate of the model based on misclassification rate formula is 81%. The results gathered only verify the validity of the model developed. Therefore, the variables that were identified as significant in predicting the outcome of heart attack mortality must be observed by many. Acquiring a prognostic model is very helpful in securing one's capability to survive from heart attack. However, it is more helpful if every individual today will observe a healthy lifestyle. Prevention is truly better than cure. The indicators were effectively identified in this research paper, anyone could now start checking their current heart condition and be able to have a big chance of treating Myocardial Infarction. In every individual's health patterns change in rapid time, for example, a new indicator of heart attack is proven by medical professionals, the model needs to be continually updated and validated yearly to improve the predictive power. Consequently, the model presented in this research paper could be no longer effective in future usage.

ACKNOWLEDGMENT

The researchers would like to express their deep gratitude to FEU Institute of Technology for all the support and continuous encouragement to pursue academic development through participating in different research conferences. Special thanks should be given to all the faculty members for the valuable insights they recommended to make this research paper possible.

REFERENCES

[1] As, G. R., & Dong, N. (2015). Developing Prediction Models from Results of Regression Analysis: Woodpecker™ Technique. Journal of Biometrics & Biostatistics, 07(01). doi: 10.4172/2155-6180.1000276

[2] Stroke: Risk factors, symptoms and the importance of time.(n.d.).Retrievedfromhttp://medprofvideos.mayoclinic.org/videos/stroke-risk-factors-symptoms-and-the-importance-of-time.

[3] Smoking and Heart Disease. (2019, July 1). Retrieved fromhttp://www.webmd.com/heart-disease/guide/smoking-heart-disease#1.

[4] Predictive Modeling: The Only Guide You Need. (n.d.). Retrievedfromhttps://www.microstrategy.com/us/resources/introductory-guides/predictive-modeling-the-only-guide-you-need.

[5] Sperandei, S. (2014, February 15). Understanding logistic regressionanalysis.Retrievedfromhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3936971/.

[6] Ajam, T., Devaraj, S., Fudim, M., Ajam, S., Soleimani, T., & Kamalesh, M. (2019). Lower Post Myocardial Infarction Mortality among Women Treated at Veterans Affairs Hospitals Compared to Men. The American Journal of the Medical Sciences. doi: 10.1016/j.amjms.2019.12.005

[7] Fukuoka, Y., Choi, J. W., Bender, M. S., Gonzalez, P., & Arai, S. (2015, April 20). Family history and body mass index predict perceived risks of diabetes and heart attack among community-dwelling Caucasian, Filipino, Korean, and Latino Americans-DiLH Survey. Retrieved from https://www.sciencedirect.com/science/article/pii/S0168822715001977.

[8] Facts and figures. (n.d.). Retrieved from http://www.bloodpressureuk.org/microsites/kyn/Home/Media/Factsandfigures.

[9] Griffin, R. M. (2012, September 14). High Cholesterol Risks: Heart Attack and Stroke. Retrieved from http://www.webmd.com/cholesterolmanagement/features/high-cholesterol-risks-top-2-dangers#1.

[10] How heart disease is different for women. (2019, October 4). Retrieved from https://www.mayoclinic.org/diseases-conditions/heart-disease/in-depth/heart-disease/art-20046167.

[11] Story, C. (2018, September 29). Heart Attack Symptoms in MenandWomen.Retrievedfromhttps://www.healthline.com/health/heart-disease/heart-attack-symptoms.

[12] Acute Myocardial Infarction in Women | Circulation. (n.d.). Retrievedfromhttps://www.ahajournals.org/doi/abs/10.1161/cir.0000000000000351.

[13] Jurgens, C. Y., Moser, D. K., Armola, R., Carlson, B., Sethares, K., Riegel, B., & Heart Failure Quality of Life Trialist Collaborators. (2009, October). Symptom clusters of heart failure. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3234105/.

[14] ModelValidation.(n.d.).Retrievedfromhttps://www.sciencedirect.com/topics/earth-and-planetary-sciences/model-validation.

[15] HOME.(n.d.).Retrievedfromhttps://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/.