RESEARCH ARTICLE

An application of Logistic Regression Modeling to Predict Risk Factors for Bypass Graft Diagnosis in Erbil

Azhin M. Khudhur, Dler H. Kadir

Department of Statistics and Informatics, College of Administration and Economics, Salahaddin University-Erbil, Kurdistan Region - F.R. Iraq

ABSTRACT

In the medical world, predictive models for assessing operative risk using patient risk factors have gained appeal as a useful tool for adjusting surgical outcomes. The goal of this study was to see if there was a link between the severity of atherosclerosis as determined by angiography and changes in several key biochemical, hormonal, and hematological variables in patients who had coronary artery bypass graft (CABG) surgery. This study included 100 adult patients who had coronary angiography, as well as a standardized case–control study of acute myocardial infarction that included 60 healthy people. In addition, not all investigations of heart attack disorders were concerned with modeling; rather, they were all concerned with classification. A family of Generalized Linear Models called Binary Logistic Regression was used. Because most phenomena' outcomes have only two values (alive/dead, exposed/not exposed, presence/absence, etc.), logistic regression analysis is a common method and plays an important role in health science. Overall, 62.5% of individuals were grouped into surgical bypass grafts, while 37.5% were healthy people. Hemoglobin A1c was wisely significant, and the odds of one unit increase led to roughly 7.488 times higher. Age and body mass index had quite high and substantial effect parameters with a 1.2 times higher likelihood than those who have smaller BMI and younger. According to the study, smokers were more likely to be at risk of undergoing bypass surgery by 4.18 times. However, there was no significant link between gender, screening creatinine, cholesterol, triglycerides, high-density lipoprotein levels, lower density level, systolic blood pressure, and diastolic blood pressure with the outcome variable.

Keywords: Bypass graft, GLMs, risk assessment of coronary artery bypass graft surgery, logistic regression model, explanatory analysis, odds ratio

AIM OF THE STUDY

The aim of this study is to examine factors affecting undergoing bypass surgery using multiple logistic regression. This can be done by assessing the contribution of factors that are linked to the probability of being in surgeries between patients and healthy people.

RESEARCH QUESTIONS

1- Are the factors significantly correlated to bypass graft?
2- What is the greatest procedure to investigate the connection between those factors and the probability of receiving surgery due to bypass graft?

BACKGROUND TO THE STUDY

Coronary heart disease can be defined as a disease that plaque (a waxy substance) constructs inside the coronary arteries. Oxygen blood is supplied to the heart by these arteries. Bypass grafting surgery is a kind of surgery that enhances the amount of blood flow to the heart. Multiple coronary arteries can be bypassed by the surgeon in just one surgery.[1]

From several of each specific trial or meta-analysis of various trials identifying whether CABG is better than percutaneous coronary intervention (PCI) and larger result for diabetic patients with multivessel coronary artery disease (CAD) at 5 years or longest follow-up, incorporating evidence from several studies, in diabetic patients with multivessel CAD. The basic notion that CABG revascularization improves survival over PCI is instantly supported by Bayesian approaches. In diabetic patients with multivessel CAD, CABG is recommended that advantage clearly exceed risk in most, if not all, diabetic

Corresponding Author:

Azhin M. Khudhur, Department of Statistics and Informatics, College of Administration and Economics, Salahaddin University-Erbil, Kurdistan Region - F.R. Iraq. E-mail: azhin.khudhur@su.edu.krd
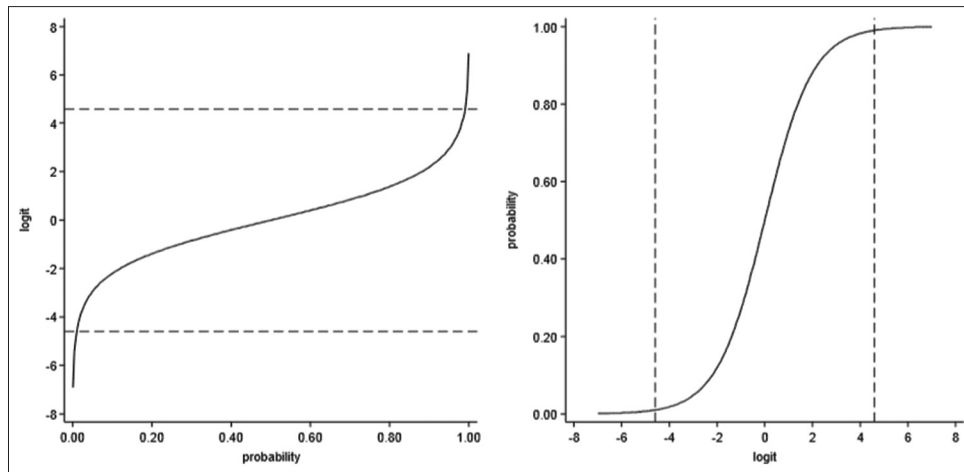
Figure 1: Logit and inverse logit functions

patients with multivessel CAD. A suggestion of positive results in general survival in FREEDOM declares that CABG appeared to negotiate greater survival than did PCI in a single supporter of patients at an individual point in time.[2]

Marshall *et al*.[3] as used Bayesian logit model for the risks associated with the coronary disease of bypass. He stated that Bayesian approach is more popular nowadays among other approaches to predict mortality between patients undergoing bypass surgery. The proposed new model uses data from 12,712 patients undergoing bypass grafting procedures collected between 1987 and 1990. The authors have provided a comparative analysis between Bayes theorem, logistic regression, discrimination analysis, and Bayesian logit models. Their data are divided into a train data set and test data set. The collected data were from various factors measured during bypass surgery and there were missing data in some variables. The data have been obtained from more than 6000 CABG surgery between 1987 and 1992 and they found outstanding risk factors that were significant.

Researchers gathered data at several sites across the United States depending on the number of CABG cases performed each year and surgeon experience with beating-heart procedures at each location. The study was limited to 3 years, from 1999 to 2001, to avoid any bias resulting from the centers' early learning experiences. The database comprised patients who had pre-operative or original surgery. The authors compared various groups to determine selection criteria for mortality and morbidity. Then, to eliminate selection bias, computer programming was utilized. Data on patients who had heart surgery were eventually acquired and the date was then entered into a logical database.

The data were evaluated using multivariate logistic regression to predict risk factors affecting mortality in 17401 isolated CABG patients. The data include 7283 (41.9%) off-pump coronary arteries and 10118 (58.1%) routine CABG with cardiopulmonary bypass procedures. Missing data were also considered as it accounted for more than 10%.

To find predictive risk factors for mortality, researchers used the multivariate logistic regression approach. Sub-groups that are almost certain to benefit were identified. The Parsonnet

risk model is a logistic regression model that was frequently used to characterize variation among treatment groups which can affect patient selection and results, with 47 potential risk factors investigated preoperatively to detect risk. This model is used to compare anticipated and observed death rates. With cardiopulmonary bypass as an independent variable, backward elimination techniques were employed to find significant predictors of operation mortality from a collection of 20 pre-operative risk factors. A series of alternate regression models were investigated with mortality and extensive morbidity as outcome variables to discover subgroups that may benefit from the beating-heart procedure.[4]

Risk improvement plays an increasing role in assessing health-care quality. Through the evaluation of hospitals, the implementation has become a necessary issue that needs a selection of convenient score systems. The study's goal was to see if administrative data could be used to predict mortality in patients undergoing CABG in the hospital. Between 2000 and 2001, data were gathered from the administrative database of hospital discharge abstracts in the Italian region of Emilia Romagna. The method used for the present study was multiple logistic regression to assess the efficiency with which mortality for patients can be predicted by conforming to each various risk adjustment procedure used.[5]

Menard[6] as used statistical analysis on bypass graft data to investigate the relation of angiography to hematological and some biochemical variables in coronary artery patients. The study aimed to examine the relation between harshness of coronary artery disease due to angiography and alteration of some necessary biochemical, hormonal, and hematological variables in underwent coronary artery bypass graft (CABG) patients. They selected 80 adults who underwent coronary angiography patients.

The present study of Menard[7] used statistical modeling procedures including multivariate regression analysis on the demographic status of the patient that is used to examine the corporation of different risk factors influencing individual medical techniques and predict the effect of any risk factor on the outcome. This would commonly imply a logistic regression analysis. In 1985, the Parsonnet scoring system was utilized to adjust mortality rates for the 1st time. The Parsonnet scoring

system identified 16 risk indicators that could affect the risk of a patient undergoing heart surgery. The considered factors were gender and age of the patients. Patients with a history of diabetes or blood pressure or both were separated as the independent variables.

GENERALIZED LINEAR MODELS (GLMS)

To make GLMs more relevant than general linear models, we must demonstrate some key properties, the most essential of which is that the least square estimate is no longer valid and maximum likelihood estimation is used instead. The most important elements are listed below:[8]

1- Random Element: This element is related to the response variable's probability distribution and is to be of the exponential family (Normal, Binomial, Poisson, Gamma, and Negative Binomial). Each defines the response variable's level values as well.

2- Systematic Element: This part denotes the model's independent (covariates) variables, which can be continuous or categorical, as well as interaction terms between predictors and polynomial functions of predictors.

3- Function of the Link: It is generally concerned with the relationship between the random (response variable) and systematic components (predictors). The factors influence the predicted value of Y:

$$g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon_i$$

The link function $g(x)$ and the parameters β_0, β_1 and so on must be approximated. The three most typical link functions are shown below.

1. Identity Link: It is represented as $g(m) = m$ in mathematics and is utilized in the typical linear models.

2. Log Link: It is represented as $g(m) = log(m)$ in math, and it is usually utilized in log-linear models for count data (non-negative numbers).

3. Logit Link: It is written as $g(m) = ln\left[\dfrac{m}{1-m}\right]$ in math. As well as for binary data (Logistic Regression).

LOGISTIC REGRESSION MODEL

In medicine and health research, one of the most typical applications of GLMs is to create a responsive model between response variables with only two outcome values (e.g., alive/dead, presence/absence, etc.) and explanatory factors, which can be any sort of data, continuous or categorical, as from Figure 1, its shape is seen as S-Shape and most of the time it is recognized as S-Shape. Due to the structure of the response variable, there were two scenarios when GLM application was required, as indicated in (binary type).[9] The first was to forecast the likelihood of survival (whether it had survived or not), and the second was to predict the likelihood of burrowing. Rees and Dineschandra[10] also looked at the association between the presence/absence of species in forest shrubs as a response variable versus a series of predictors including distance to nearest woodland, stand age, stand area, and so on.

LOGISTIC MODEL AND PARAMETERS

The logistic regression model can be expressed as follows:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}}$$

Where $\pi(X)\pi(X)$ is the probability that $y_i = 1$ (i.e., the event is present) for a collection of specified x_i's and βs are parameters that need to be estimated. To make the parameters easier to comprehend, we can simply convert them to odds and interpret them the same way we did in the 2×2 contingency table, which is the likelihood of an event occurring compared to its inverse.

$$Odds = \frac{\pi(x)}{1 - \pi(x)}$$

If the odds are more than one, the chance of $y_i = 1$ is greater than the reverse, and vice versa. By putting the coefficient values on an exponential function, we can simply transform logistic regression odds to odds and odd ratio. The natural logarithm can then be applied to the odds formula to make modeling easier.[11]

$$g(x) = ln(Odds) = ln\left[\frac{\pi(x)}{1 - \pi(x)}\right]$$

Then $g(x)$ can be controlled as follows:

$$g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

It is also worth noting that the maximum likelihood estimation is processed and the values are produced through computing due to their complexity. A test called the likelihood ratio test (LRT) is used to determine whether an overall model is significant, as it compares the likelihood of the null model (i.e., a model with only intercept).[12] The LRT functionality is almost similar to the F-test in linear regression and its formula is as follows:

$$G_0^2 = 2ln\frac{L}{L_0} = 2lnL - lnL_0$$

Where L represents the entire model's probability and (L_0) represents the null model's likelihood. The statistic (G_0) has a k degree of freedom (df) with an approximate two distribution (k=number of predictors in the entire model). However, we evaluate the complete and saturated model employing deviance (SSE in linear regression) using a special type of LRT.[13,14] The deviance is an unexplained measure of variation in the data.

$$D = 2ln\frac{L_{sat}}{L_{full}} = 2(lnL_{sat} - lnL_{full}) = -2(lnL_{full})$$

Furthermore, when it comes to determining the significance of parameters, the Wald test[4] is a method for evaluating the null hypothesis.

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

MATERIALS AND METHODS

Dataset

This study is a case–control study so 100 adult patients underwent coronary angiography at Erbil Cardiac Center

and data were obtained from them. Patients were followed up until the end of December 2020. Data were also obtained from 60 healthy people who underwent the same coronary angiography.

The obtained data were collected from both patients and healthy people. The data include gender, age, smoke, blood sugar, creatinine, weight, length, body mass index (BMI), hemoglobin A1c (HbA1C), blood urea nitrogen, cholesterol, lower density level, high-density lipoprotein levels (HDL), triglycerides (TG), systolic blood pressure (SBP), and diastolic blood pressure (DBP) in Erbil City.

ILLUSTRATING AND REPORTING LOGISTIC REGRESSION ANALYSIS RESULT

It is preferable to construct descriptive statistics for the data and examine how the variables connect to the outcome variable. Table 1 indicates that some independent factors and the dependent binary variable may have a relationship with the test value. We first demonstrate the continuous variables in accordance to control and patient groups, it is clear that age has quite an influence to get affected with the bypass graft. The result shows that the mean age for the control group is 44.87-years-old, while the patient people tend to be 59.51 and this tells us more into age, more chance to be suffering, and the P-value of the T-test can confirm that with a value <0.05.

Moreover, the high value of blood sugar is more likely to put an individual at risk of heart problems, as from the initial output the mean values are far away from each other among the control and patient people with 105.87 and 180.08, respectively. To support this, the driven T-test was found to be statistical significance since its $P < 0.05$.

Another interesting result is for BMI measurement, as it is recorded at 24.7 for people who are in the control group whereas a 28.61 mean value is found for the patient group.

Table 1: Descriptive statistical analysis of the associated attributes with the outcome variable

Variables	Control	Patient	T. Test	*P*-value
	Mean±SD	Mean±SD		
Age	44.87±14.31	59.51±8.51	−8.12	0.0000
Blood Sugar	105.87±14.97	180.08±99.57	−5.73	0.0000
HBA1C	5.07±0.55	7.21±2.25	−7.23	0.0000
Blood Urea	28.13±5.44	36.88±15.33	−4.26	0.0000
Creatinine	0.7±0.27	0.899±0.32	−4.05	0.0001
CHO	142.62±27.19	160.61±40.78	−3.04	0.0028
TG	140.55±65.5	175.4±82.32	−2.79	0.0059
HDL	39.02±8.106	36.51±10.03	1.64	0.1030
LDL	84.38±21.6	100.57±36.38	−3.13	0.0021
Weight	69.72±12.56	77.87±15.82	−3.4	0.0009
Length	1.67±0.07	1.65±0.1	1.59	0.1134
BMI	24.7±3.09	28.61±5.59	−4.98	0.0000
SBP	127.85±10.61	128.99±19.86	−0.41	0.6800
DBP	74.13±6.09	78.1±12.92	−2.23	0.0271

Based on this, T-test reported that the mean values differed from each other statistically significantly.

On the other hand, HDL and SBP were highlighted to be non-statistically significant with *P*-values (0.1030 and 0.6800) greater than 0.05, respectively.

Likewise, to assess the relationship of the response variable as being in surgery due to bypass graft definition with the categorical variables in the study, results in Table 2 are found for gender as well as smoking status.

The percentage of males in the patient group is about double toward of females with 64% and 36% constantly according to Figure 2, while in the control group, their differences are only 13%. This leads us to confirm that males are in the position to get an operation due to bypass graft concern and the Chi-square test shows that there is an association between these two variables. Similarly, smoking seemed to have a quite strong influence as shown in Figure 3.

LOGISTIC REGRESSION ANALYSIS

The dataset was subjected to a multiple logistic regression analysis to determine the likelihood of some biographic, health, and other parameters having an impact on being into surgery. RStudio version 1.4.1717 was used to conduct the analysis. We first implemented the simple logistic regression plot for each of the
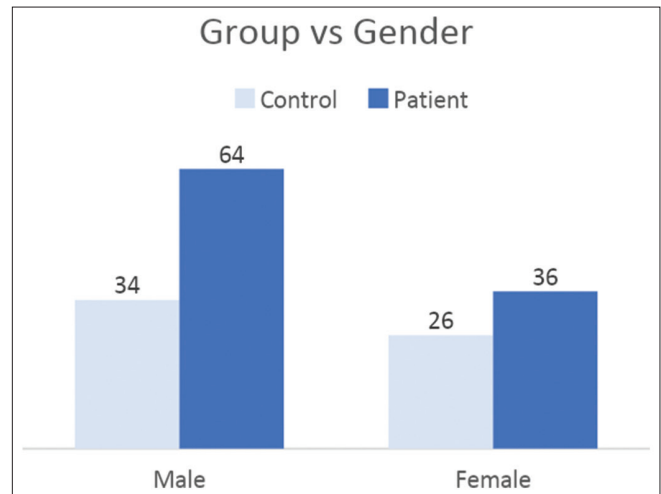

Figure 2: Bar-chart demonstration between response and Gender variables

Table 2: Cross-tabulation table of Chi-square for gender and smoking versus outcome variable

Variables	Control		Patient		Chi-square	*P*-value
	N	%	N	%		
Gender						
Male	34	56.67	64	64	0.84	0.0357
Female	26	43.33	36	36		
Smoke						
Yes	10	16.67	35	35	6.24	0.013
No	50	83.33	65	65		

variables individually with the curve plot as shown in Figure 4, to get a clue about the nature association among the response and covariates. It is important to spot some potential points here as seen that age, BMI, creatinine, blood sugar, and HBA1C variables showed a relatively interesting link with the dependent variable.

Considering the descriptive statistics along with Figure 4, the final model has been taken into account for further discussion. First, it was preferred to enter only the significant variables according to Wald test statistics called the reduced model compared to the full model as shown in Table 3.
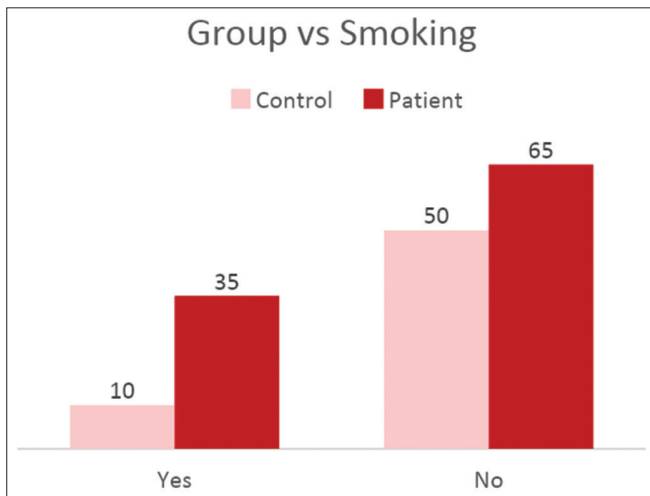

Figure 3: Bar-chart demonstration between response and smoking variables

Table 3: AIC of two fitted models

	Model Fitted	AIC
Model 1:	Surgery~Age+BMI+Blood_ Sugar+HBA1C+Blood Urea+factor (Smoke)	81.871
Model 2:	Surgery~Age+BMI+Blood_ Sugar+HBA1C+ Blood Urea+Creatinine+CHO+TG+HDL+LDL +SBP+DBP+factor (Gender) + factor (Smoke)	83.833

Table 4: Chi-square test of two fitted models

S. No.	Residual Dfa.	Residual Deviance	Df	Deviance	Pr(>Chi)
1	153	67.871			
2	145	53.833	8	14.039	0.08077

From Table 4, the ANOVA test was carried out using Chi-square test to see if the model with fewer variables improved the model. Consequently, it turned out that the reduced model is much better than the full model referencing AIC, and another important feature is that all parameters were statistically significant.

Table 5's result is perhaps the most essential in logistic regression output since it shows how the outcome variable and covariates are related. As seen, all the calculated parameters (Wald statistics) were found to be statistically significant with $p < 0.05$. The odds ratio was calculated by exponentiation B's value to grasp and interpret the parameter value.

The most effective attribute was found to be HBA1C, as per the output, it shows a significant and positive impact on the response variable holding other variables constant. The result provides us with (7.488) more chances to be selected for bypass graft operation with one unit increase in this variable in contrast to a low unit.

Referencing participants who smoke, the odds (4.418) indicate that according to the model, smokers have a 4.418 times higher chance of being classified as a patient than non-smokers. This is in fact enables us that people, who smoke have a higher chance to get into bypassing graft monitoring.

In addition, age and BMI covariates were also found to be effective with similar strength in raising people to be considered for bypass graft surgery, and the model predicts that one unit increase in age and BMI is resulting in an increase in the odds of being suffering from bypass graft with 1.180 chance.

The odds ratio for blood sugar was computed to be 1.050, with a 95% confidence interval of 0.005–0.107. A case has a 1.050 likelihood of having surgery compared to not having surgery for every one-unit rise in blood sugar, a 50% increase in the probabilities of having surgery.

Blood urea also played an important role to be effective in increasing risks for the response variable as with one unit increase in blood urea measurement this would lead to having odds (1.092) times chance to face bypass graft surgery compared to low blood urea unit.

When comparing low blood urea units to high blood urea units, blood urea had a major role in increasing risks for the response variable, as one unit increase in blood urea measurement resulted in an odds (1.092) times likelihood of having bypass graft surgery.
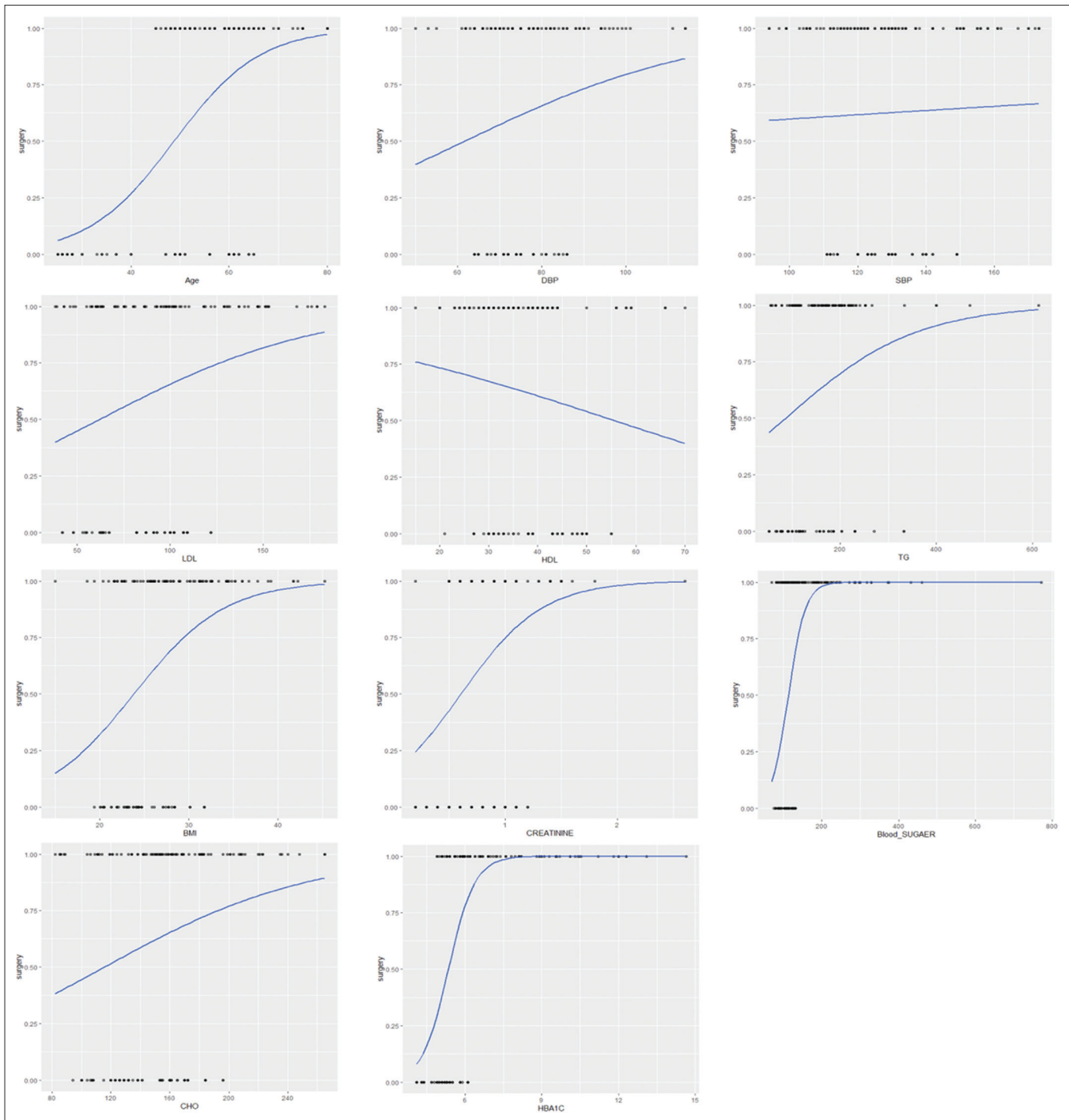
Table 5: Result of estimated parameters and odds ratio in logistic regression

	B	S.E.	Wald	Sig.	Exp (B)	95% C.I.for EXP (B)	
						Lower	Upper
(Intercept)	−31.446	6.414	−4.903	0.000	0.000	−51.569	−17.926
Age	0.169	0.038	4.399	0.000	1.184	0.087	0.294
BMI	0.167	0.081	2.049	0.040	1.182	−0.030	0.401
Blood sugar	0.048	0.020	2.484	0.013	1.050	0.005	0.107
HBA1C	2.013	0.691	2.912	0.004	7.488	0.567	4.180
Blood Urea	0.088	0.039	2.244	0.025	1.092	−0.004	0.205
factor (Smoke) 2	1.486	0.714	−2.082	0.037	4.418	−3.525	0.257

Figure 4: Logit model fit of surgery on predictors

Table 6 shows that the full model reduces the log-likelihood value, resulting in a substantial difference between the two models.

The goodness of fit test was used to evaluate the fitted model, as shown in Table 7. There was no actual evidence for lack of fit because the statistical test of Hosmer-Lemeshow was not significant, as it was for Pearson Chi-square or G_0^2 (deviance). Furthermore, the Nagelkerke R Square suggests a positive impression of the model, which explained 80.8% of the uncertainty in going through bypass graft diagnosis using the factors incorporated in the model.

Finally, yet importantly, it is worth mentioning how the classification percentage improved after fitting the model with and without the influential factors. The overall percentage of truly predicted cases was 62.5% where there was only a constant in the model, and this had been significantly increased to 93.8%.

Table 6: Log-likelihood value of the full model and reduced model

Models	Chi-square
Full model log-likelihood	75.798
Reduced model (constant only) log-likelihood	143.829
−2 Log likelihood	67.871
P-value	0.000

Table 7: Goodness of fit output for the full model

Statistic	Value	Df	P-value
Hosmer and Lemeshow	14.337	8	0.073
Nagelkerke R Square	0.808		
Cox and Snell R Square	0.593		

CONCLUSION AND RECOMMENDATION

Several essential criteria were discovered to be highly successful in increasing high risk among patients undergoing coronary artery bypass grafting, and they may not be the same in different places. Our model would be able to properly forecast the instances with an accuracy of 80.8%, which is substantial, and it is worth noting that we arrived at the final model after eight steps of the model selection approach, which is a decent fit to explain the dataset.

HBA1C had a significant impact on arising the risk, and increasing one unit at HBA1C increased the odds of being into surgical admission by 7.488 times, holding other variables constant. Age and BMI had quite high and substantial effect parameters, with individuals who are quite an in age and had relatively high BMI led to be at risk with a 1.2 times higher likelihood than those who have smaller BMI and younger. Furthermore, blood sugar and blood urea played a significant role in being effective contributors, with 1.05 and 1.09 odds ratios in favor of increasing the risk. According to the study, smokers were more likely to be at risk of undergoing bypass surgery than non-smokers, with an odds ratio of 4.418.

Gender, screening creatinine, CHO, TG, HDL, LDL, SBP, and DBP, on the other hand, were not found to be statistically significant. It is worth noting that this study had its own limitations, such as the small number of cases. To determine how effective the covariates influenced the outcome variable, create a classification table and calculate sensitivity, specificity, and display the receiver operating characteristic curve.

REFERENCES

1. J. Alexander and P. Smith. Coronary-artery bypass grafting. *The New England Journal of Medicine*, vol. 374, no. 20, pp. 1954-1964, 2016.
2. C. D. Lang, Y. He and J. A. Bittl. Bayesian inference supports the use of bypass surgery over percutaneous coronary intervention to reduce mortality in diabetic patients with multivessel coronary disease. International Journal of Statistics in Medical Research, vol. 4, no. 1, pp. 26-34., 2015.
3. G. Marshall, A. L. W. Shroyer, F. L. Grover and K. E. Hammermeister. Bayesian-logit model for risk assessment in coronary artery bypass grafting. *The Annals of Thoracic Surgery*, vol. 57, no. 6, pp. 1492-1500, 1994.
4. D. Kadir. Likelihood approach for bayesian logistic weighted model. *Cihan University-Erbil Scientific Journal*, vol. 4, no. 2, pp. 9-12, 2020.
5. G. Q. Othman, R. S. Saeed, D. H. Kadir and H. J. Taher. Relation of angiography to hematological, hormonal and some biochemical variables in coronary artery bypass graft patients. *Journal of Physics*, vol. 1294, no. 6, p. 062110, 2019.
6. S. Menard. *Applied Logistic Regression Analysis*. Vol. 106. New York: Sage.
7. S. Menard. Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, vol. 54, no. 1, pp. 17-24, 2000.
8. M. J. Mack, A. Pfister, D. Bachand, R. Emery, M. J. Magee, M. Connolly and V. Subramanian. Comparison of coronary bypass surgery with and without cardiopulmonary bypass in patients with multivessel disease. *The Journal of Thoracic and Cardiovascular Surgery*, vol. 127, no. 1, pp. 167-173, 2004.
9. C. Ugolini and L. Nobilio. Risk adjustment for coronary artery bypass graft surgery: an administrative approach versus EuroSCORE. *International Journal for Quality in Health Care*, vol. 16, no. 2, pp. 157-164, 2004.
10. M. Rees and J. Dineschandra. Risk stratification in assessing risk in coronary artery bypass surgery. In: 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06), pp. 303-308, 2006.
11. J. Hilbe. Generalized additive models softwar. *The American Statistician*, vol. 47, pp. 59-64, 1993.
12. M. Beck. Size-specific shelter limitation in stone crabs: A test of the demographic bottleneck hypothesis. *Ecology*, vol. 76, pp. 968-980, 1995.
13. G. Matlack. Plant species migration in a mixed history forest landscape in eastern North America. *Ecology*, vol. 75, pp. 1491-1502, 1994.
14. N. H. Mahmood, R. O. Yahya and S. J. Aziz. Apply binary logistic regression model to recognize the risk factors of diabetes through measuring glycated hemoglobin levels. *Cihan University-Erbil Scientific Journal*, vol. 6, no. 1, pp. 1-7, 2022.