Predicting risk factors for postoperative coronary artery bypass grafting using logistic regression and CHAID

A Nabilah, D Lestari, S Mardiyati and S Devila

Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia

Corresponding author’s email: dian.lestari@sci.ui.ac.id

Abstract. Non-fatal postoperative complications are postoperative morbidity that can affect the patient's functional status and quality of life. Evaluation of postoperative morbidity is the step needed to assess and improve the quality of patient care. Therefore, a method is required in order to predict risk factors in evaluating a patient's postoperative morbidity. After this the results will be used to determine the insurance premiums. In this research, the Logistic Regression are used to know the risk factors that would occur in patients who had undergone Coronary Artery Bypass Grafting (CABG) surgery. Then we use the CHAID method to classify readmission based on patient characteristics. Based on the two analyzes, it can be concluded that the CHAID analysis supports the Logistics analysis, there are two risk factors significantly influence the complications of patients after Coronary Artery Bypass Graft (CABG), namely Sex and Ejection Fraction.

Keywords: Maximum likelihood method, morbidity, risk factors

1. Introduction

Based on data from the Sample Registration System (SRS) in 2014, coronary heart disease was the highest cause of death, which reached 12.9 % [1]. One way that can be done to reduce deaths from coronary heart disease (CHD) is by performing Coronary Artery Bypass Grafting (CABG). Coronary Artery Bypass Grafting is an action taken to overcome coronary heart disease by eliminating coronary artery blockage and maximizing the flow of blood vessels [2].

Before performing the Coronary Artery Bypass Grafting action, an analysis of risk factors before the CABG action needs to be done to determine the postoperative risk. Data used for the analysis of factors before surgery are derived from medical records of CABG patients [3]. Risk factors before surgery are one of the things needed to determine the morbidity that will occur postoperatively. CABG postoperative morbidity determines the length of time the patient is hospitalized. Therefore, the risk factor prediction model plays an important role that can be used to predict morbidity that will occur after CABG surgery.

There have been several studies regarding risk prediction models for CABG operations that have been conducted. In 2002, Hakala, et al. discussed about predicting the risk of atrial fibrillation prediction after CABG operation with logistic regression in Finland, and found that ejection fraction (EF) influence the complication of patient [4]. In 2003, Huijskes, et al. discussed the risk prediction that occurred after underwent CABG and valve surgery in Netherland, using multivariate logistic regression, and found that sex and injection fraction influence the complication of patient [5]. In 2006, Ghotkar, et al. proposed
logistic regression with forward stepwise to identify the risk factors that occurred after underwent CABG surgery in England, and found that renal dysfunction, unstable angina, ejection fraction, peripheral vascular disease, obesity, age, smoking, diabetes, priority, hypercholesterolaemia, and hypertension affect the complication of patient [6]. In 2009, Antunes, et al. discussed the risk prediction that occurred after underwent CABG operations in Portugal, using the development of logistic regression and bootstrap analysis models, and found that Diabetes, Obesity, and peripheral vascular disease affect the complication of patient [7]. In this study, risk factors for CABG surgery will be predicted using the logistic regression and CHAID methods to provide information to patients about the risk following CABG operation that influence the complication of patient.

2. Methodology

2.1. Logistic regression model

Logistic regression is a regression model used to analyze the relationship between predictor variables and response variables, where the response variable has only two possibilities, namely success, and failure expressed by 0 and 1. The following is a general form of logistic regression [8]:

\[
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i,
\]

or can be written as:

\[
y_i = x_i^T \beta + \varepsilon_i,
\]

where \( i = 1, 2, 3, \ldots, n \), and \( n \) denotes the number of observations, \( x_i^T = [x_{i1}, x_{i2}, \ldots, x_{ik}] \) and

\[
\beta = \begin{bmatrix} 
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_k 
\end{bmatrix}.
\]

It is assumed that \( Y_i \) has a Bernoulli distribution (\( Y_i \) is the response variable) and has the following probability distribution (table 1).

Since \( E(\varepsilon_i) \), then the expected value for the response variable is given as follows:

\[
E(Y_i) = (1 - \pi_i)0 + \pi_i, 1 = \pi_i
\]

which means:

\[
E(Y_i) = E(x_i^T \beta + \varepsilon_i) = E(x_i^T \beta) + E(\varepsilon_i) = x_i^T \beta.
\]

Since \( Y \) is a binary response variable, then the value of \( E(Y_i) \) has the following restrictions:

\[
0 \leq E(Y_i) = \pi_i \leq 1.
\]

In this logistic regression, we use the link function, in the form of a log link that is defined as follows:

\[
x_i^T \beta = \ln \left( \frac{\pi_i}{1 - \pi_i} \right),
\]

\[
\pi_i = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}.
\]
Table 1. The probability of response variable

| $Y_i$ | Probability |
|-------|-------------|
| 1     | $P(Y_i = 1) = \pi_i$ |
| 0     | $P(Y_i = 0) = 1 - \pi_i$ |

This function is used to guarantee that no expectation value of the response variable belongs to the interval $(-\infty, \infty)$. We use the maximum likelihood method to estimate the parameter of $\beta$. Each observation follows the Bernoulli distribution. Hence, the probability distribution of observations is

$$f(y_i) = \pi_i y_i (1 - \pi_i)^{1-y_i}, \quad i = 1, 2, ..., n.$$  \hspace{1cm} (8)

The observation $y_i$ has a value of 0 or 1. Since the observations are mutually independent, then the likelihood function is given as follows:

$$L(\pi_1 | y_1, y_2, ..., y_n) = \prod_{i=1}^{n} f(y_i | \pi_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}. \hspace{1cm} (9)$$

By substituting equation 7 to equation 9, the likelihood function of $\beta$ is obtained as follows:

$$L(\beta_0, \beta_1, \beta_2, ... \beta_k) = \prod_{i=1}^{n} \left( \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right)^{y_i} \left( 1 - \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right)^{1-y_i} = \prod_{i=1}^{n} \frac{e^{x_i^T \beta y_i}}{1 + e^{x_i^T \beta}}. \hspace{1cm} (10)$$

Therefore, the log-likelihood function is

$$\ln L(\beta_0, \beta_1, \beta_2, ... \beta_k) = \ln \prod_{i=1}^{n} \frac{e^{x_i^T \beta y_i}}{1 + e^{x_i^T \beta}} = \sum_{i=1}^{n} (x_i^T \beta) y_i - \sum_{i=1}^{n} \ln \left( 1 + e^{x_i^T \beta} \right). \hspace{1cm} (11)$$

The next step is to take the derivative of equation 11 with respect to $\beta_0, \beta_1, \beta_2, ..., \beta_k$. The resulting reduction is accompanied by zero to get the optimal estimated value. Hence, we get the following general equation:

$$\sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij} \hat{\pi}(x_i) = 0, j = 0, 1, 2, ..., k \hspace{1cm} (12)$$

with $\hat{\pi}(x_i) = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}$.

Furthermore, the estimator $\hat{\beta}$ is obtained by solving equation 12, but equation 12 is still in the implicit form so that a numerical method is needed in its solution. In this research, Newton-Raphson method will be used. The principle of Newton-Raphson method is to find $\hat{\beta}$ that satisfies $f(\hat{\beta}) = 0$, with $f(\hat{\beta})$ is given in equation 13. Next, we will look for the second derivative of the log-likelihood function for each parameter, and the general equation will be compressed as follows:

$$\frac{\partial^2 l(\beta)}{\partial \beta_a \partial \beta_b} = - \sum_{i=1}^{n} x_{ia} x_{ib} n_i \pi(x_i) (1 - \pi(x_i)) = f(\hat{\beta}); a, b = 0, 1, 2, ... k. \hspace{1cm} (13)$$
From the equation 13, then an iteration will be performed to find the estimated value of $\hat{\beta}$. After obtaining the estimated value of $\hat{\beta}$, $j = 0, 1, 2, ..., k$, the next step is testing the Goodness of fit to find out whether the model obtained is feasible or not to be used. The hypothesis used in this test is as follows:

$H_0 =$ Model is feasible to use.
$H_1 =$ Model is not suitable to use.

The test statistic used in this study is deviance, which is formulated as follows:

$$D = -2 \sum_{i=1}^{n} \left[ y_i \ln \left( \frac{\hat{\beta}_i}{\hat{\gamma}_i} \right) + (1 - y_i) \ln \left( \frac{1 - \hat{\beta}_i}{1 - \hat{\gamma}_i} \right) \right] \sim \chi^2_{(n-k)}.$$  

The principle of the decision is to reject the $H_0$ if $D > \chi^2_{(n-k)}$. After testing using deviance then Wald tests will be performed to determine the effect of risk factors on postoperative readmission. The results of this test indicate whether a risk factor is feasible to enter into the logistic regression model [9].

The hypothesis used in this test is as follows:

$H_0 =$ There is no influence of risk factors on patient readmission.
$H_1 =$ There is an influence of risk factors on patient readmission.

The test statistics used in this test is:

$$W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim \chi^2, \quad j = 0, 1, 2, 3, ... k$$  

The rule to reject the $H_0$ if $W < \chi^2_{(1-a)^2}$ or $W > \chi^2_{(1-a)^2}$.

Interpretation of Logistic Regression Model Parameters needs to be done by looking at the odds ratio value. The odds ratio is a measure to see how likely the influence of risk factors is on post-surgery patient readmission and is defined as follows [10]:

$$\hat{\delta}_R = \frac{\text{odds}_Y}{\text{odds}_X} = e^{\hat{\beta}_1}, \quad (16)$$

with $\hat{\delta}_R$ represents the ratio between the odds of a binary response variable $Y$ and a binary predictor variable $X$.

2.2. Chi-square automatic interaction detection

The chi-square automatic interaction detection (CHAID) method was first introduced by Kass in 1980. This method is a method for classifying categorical data to divide data sets into sub-groups. In this study, the CHAID method is used to classify readmission based on patient characteristics. The results of the CHAID classification will be displayed in a tree diagram.

The CHAID method is divided into three stages [11], namely:

a. Merging Phase, at this phase, the significance of each category of independent variables will be examined for the dependent variable. The first step that must be done is to form a two-way contingency table for each independent variable with the dependent variable. Then calculate the Chi-Square statistics for each pair of categories that can be chosen to be combined into one, here are the test statistics used [12]:

$H_0$ is rejected if $\chi^2 > \chi^2_{table}$
$H_1$ is accepted if $\chi^2 < \chi^2_{table}$
\[ X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}, \]

(17)

with \( X^2_{table} \) is obtained from Chi-Square table, \( n_{ij} \) represents the number of observation, and \( E_{ij} \) represents the expected frequency of observations given by the following formula

\[ E_{ij} = \frac{n_i n_j}{N}, \]

(18)

with

\( n_i \) : The number of \( i \)-observations.
\( n_j \) : The number of \( j \)-observations.
\( N \) : \( \sum_{j=1}^{c} \sum_{i=1}^{r} n_{ij} \).

b. Then for each paired chi-square value, calculate the paired \( p \)-value together. Among insignificant pairs, combine a pair of the most similar categories. For couples who have the smallest paired chi-square value and the largest \( p \)-value becomes a single category. Next, re-examine the significance of the new class after combining it with other class in the independent variable. If there are still significant pairs, repeat the previous steps. If all pairs are significant, then calculate Bonferroni’s corrected \( p \)-value based on the merged table.

c. Separation Phase, this phase is the phase of selecting variables that are used as the best node separator. The risk factor variable that has the smallest \( p \)-value is used as the best node separator, and if no risk factor variable has a \( p \)-value \( \leq \alpha \), then the separation is not performed [13].

d. Termination Phase, this phase is done if there is no real relationship between the risk factor variable with the response variable and the size of the child node is less than the minimum node child-size specification. If there is a merger in the explanatory variable, then a Bonferroni correction is performed for each type of free variable as follows. Bonferroni correction is a correction process for several statistical tests when independence is carried out together.

Independent type Bonferroni multipliers for nominal origin variables [14]:

\[ B = \sum_{i=0}^{r-1} \frac{(-1)^i (r - i)^c}{i! (r - 1)!}, \]

(19)

Monotonic type Bonferroni multiplier for ordinal origin variables is:

\[ B = \binom{c - 1}{r - 1}, \]

(20)

with

\( B \) : Bonferroni multiplier
\( c \) : The number of origin categories
\( r \) : The number of new categories

3. Results and discussion

The data used in this study came from the European Society of Cardiology (ESC) guidelines recorded in 2010–2014. There were 388 patients observed where 82 patients had disease complications after CABG, and 306 patients did not have disease complications after CABG. The percentage of each class
of patients that occurred complications and complications did not occur after the action of CABG can be seen in Table 2.

Risk factors observed included age, sex, weight, height, Body Mass Index (BMI), Hypertension, Diabetes mellitus (DM), Hyperlipidemia, Chronic Obstructive Pulmonary Disease (COPD), Peripheral arterial disease (PAD), stroke, Canadian Cardiovascular Society (CCS), Congestive Heart Failure (CHF), Myocardial Infarction (MI), Posterior Cruciate Ligament (PCL), Chronic Kidney Disease (CKD), Atrial fibrillation (AF), Diseased Vessels, Left Main Coronary Artery (LM), and Ejection Fraction (EF).

In this study, the observed risk factors are used to construct a prediction model with logistic regression. Using the Newton-raphson method, an estimated value of the coefficient $\beta$ for each risk factor is obtained as shown in Figure 1.

After obtaining the estimated coefficient $\beta$ for the prediction model, the next step is to carry out the Wald test to determine the effect of risk factors on the occurrence of complications after carrying out CABG individually. Based on Wald's value in Figure 1, with the value of $\alpha$ is 10%, five risk factors significantly influence complications after CABG, namely Sex, NYHA, and EF. Thus, the prediction model used is as follows.

$$\pi_i = \frac{e^{0.054 + 0.665(Sex) + 0.363(NYHA) - 0.018(EF)}}{1 + e^{0.054 + 0.665(Sex) + 0.363(NYHA) - 0.018(EF)}}$$ (21)

Table 2. The number of patients in every class.

| Class                      | The number of patients | Percentage |
|----------------------------|------------------------|------------|
| No (There are no complications) | 306                    | 78.87 %    |
| Yes (There are complications) | 82                     | 21.13 %    |

Figure 1. The result of logistic regression
From the odds ratio in the figure 1, we found that men are more at risk of 1.944 times more than women, NYHA is more at risk of 1.438 times more than not experienced, and EF is more at risk of 0.982 times than not experienced. Then, CHAID analysis will be carried out with the value of $\alpha = 10\%$. The result obtained from the CHAID analysis can be seen in figure 2.

As shown in figure 2, the risk factor that has the strongest influence on the occurrence of morbidity is the sex variable. Out of 388 respondents, around 25% of male sex experienced post-operative morbidity, and 12.9% of female patients experienced morbidity. Variable sex and male sex around 43.6% having morbidity while sex and female having morbidity around 20.1%.

4. Conclusion
Based on the analysis using logistic regression that has been done on patient data after CABG surgery from the European Society of Cardiology, it obtained three risk factors that significantly influence post-patient complications (CABG), namely sex, NYHA, and ejection fraction. Whereas, if using CHAID analysis, there are two risk factors significantly influence the complications of patients after Coronary Artery Bypass Graft (CABG), namely sex and ejection fraction.

Based on the data analysis carried out, it can be concluded that the CHAID analysis supports the Logistics analysis in predicting risk factors for postoperative CABG.

Acknowledgments
This work was financially supported by Universitas Indonesia under research grant PITTA with grant contract number NKB-0669/UN2.R3.1/HKP.05.00/2019.

References
[1] Depkes RI 2013 *Riset Kesehatan Dasar* (Jakarta: Badan Penelitian dan pengembangan Kesehatan Kementrian Kesehatan RI)

[2] Hardian R, Satoto H and Soenarjo 2014 *Jurnal Anestesiologi Indonesia* **4** 6435

[3] Vercellis C 2009 *Business Intelligence: Data Mining and Optimization for Decision Making* (Chichester: John Wiley and Sons Ltd)

[4] Hakala T, Pitkanen O and Hippelainen M 2002 *Scand. J. Surg.* **91** 339-44

[5] Huijskes R V H P, Rosseel P M J and Tijsen J G P 2003 *Eur. J. Cardiothorac. Surg.* **24** 741-9

[6] Ghotkar S V, Grayson A D, Fabri B M, Dihmis W C and Pullan D M 2006 *J. Cardiothorac. Surg.* **1** 1-8

[7] Antunes P E, Oliveira J F d and Antunes M J 2009 *Eur. J. Cardiothorac. Surg.* **35** 760-8

[8] Hosmer D W and Lemeshow S 2000 *Applied Logistic Regression, 2nd Ed.* (New York: John Wiley and Sons Ltd)

[9] Kass G V 1980 *Appl. Statist.* **29** 119-27

[10] Ritschard G 2013 CHAID and Earlier Supervised Tree Methods *Contemporary Issues in Exploratory Data Mining in Behavioral Sciences* ed. McArdle JJ and Ritschard G (New York: Routeledge)

[11] Diaz-Perez F M and Bethencourt-Cejas M 2016 *J. Dest. Mark. Manage.* **5** 275-82

[12] Gallagher C A 2000 *J. Appl. Statist.* **29** 256-66

[13] Sharp R A, Romaniuk J T and Ciepicki S A 1998 *Proc. ANZMAC Conf.* **1** 2398-411

[14] Bagozzi R 1994 *Advanced Methods of Marketing Research* (Oxford: Blackwell Publishers)