TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC QUY NHƠN

Mô hình tiên đoán bệnh tim mạch vành sử dụng hồi quy logistic dựa vào tập dữ liệu Evans

Lê Thanh Bính*

Khoa Toán và Thống kê, Trường Đại học Quy Nhơn, Việt Nam

Ngày nhận bài: 24/02/2020; Ngày nhận đăng: 22/04/2020

TÓM TẮT

Bệnh tim mạch, bao gồm cả bệnh tim mạch vành (CHD), nằm trong số những bệnh phổ biến ở cả những nước phát triển và đang phát triển và được xem là nguyên nhân chủ yếu gây tử vong trên toàn thế giới. Chỉ riêng bệnh tim mạch vành, bệnh này tiếp tục là nguyên nhân hàng đầu gây nên bệnh tật và tử vong ở người trưởng thành tại châu Âu và Bắc Mỹ. Sự tiên lượng sớm bệnh tim mạch vành có thể giúp đưa ra các quyết định thay đổi lối sống ở những bệnh nhân có nguy cơ cao và từ đó làm giảm các biến chứng của bệnh. Vì vậy, sử dụng các thuật toán khai phá dữ liệu có thể hữu ích trong dự đoán bệnh tim mạch vành. Nghiên cứu này nhằm xây dựng một mô hình dự đoán bệnh tim mạch vành sử dụng hồi quy logistic, với sự trợ giúp của phần mềm thống kê R, dựa vào tập dữ liệu Evans về bệnh tim.

Từ khóa: *Bệnh tim mạch, bệnh tim mạch vành (CHD), mô hình hồi quy logistic, tập dữ liệu Evans.*

**Tác giả liên hệ chính.*
Email: *lethanhbinh@qnu.edu.vn*

JOURNAL OF SCIENCE
QUY NHON UNIVERSITY

A coronary heart disease prediction model using logistic regression based on Evans dataset

Le Thanh Binh[1,*]

[1]*Faculty of Mathematics and Statistics, Quy Nhon University, Vietnam*

Received: 24/02/2020; Accepted: 22/04/2020

ABSTRACT

Cardiovascular diseases, including coronary heart disease (CHD), are among the common diseases in both developed and developing countries and regarded as the main cause of death throughout the world. Coronary heart disease itself has been still being the leading cause of morbidity and mortality among adults in Europe and North America. The early prognosis of coronary heart disease can help making decisions in changing lifestyle of high-risk patients, thereby reducing complications of the disease. Therefore, the use of data mining algorithms could be useful in predicting coronary heart disease. This study aimed to create a coronary heart disease prediction model using logistic regression, with the help of statistical software R, based on the Evans heart disease dataset.

Keywords: *Cardiovascular diseases, coronary heart disease (CHD), logistic regression model, Evans dataset.*

1. INTRODUCTION

Cardiovascular diseases (CVDs) are caused by disorders of the heart and bood vessels. CVDs include *coronary heart disease* (CHD, heart attacks), *cerebrovascular disease* (stroke)*, raised blood pressure (hypertension), peripheral artery disease, rheumatic heart disease, congenital heart disease and heart failure.* CVDs are the leading cause of morbidity and mortality worldwide, with 80% of total deaths occurring in developing countries.[1] Based on the report by WHO, in 2017, more than half (54%) of the dealths around the world were caused 10 leading causes, and CVDs which led to 15 million dealths in 2015 constituted the largest group of fatal diseases.[2] CVDs kill millions of people annually and this value may be increased up to 24.8 million by 2020 if preventive measures are not taken.[3]

In Vietnam, CVDs was among the top 10 leading causes of death in 2006, 2007 and 2009.[4,5] Around 32% of deaths from non-communicable diseases in rural areas are caused by CVDs.[6,7] For just coronary heart disease (CHD), according to the WHO data published in 2017 CHD deaths in Vietnam reached 58452 or 11.58% of total deaths. Evidently, the burden of CHD will continue to rise unless effective interventions for addressing its underlying risk factors are put in place.

Early detection of complications helps to treat CHD patients in a comprehensive way. Therefore, medical communities attempt to find a way for the accurate and timely prediction of CHD by using new statistical techniques, such as data mining techniques.[8] These techniques can help to recognize the patterns and factors influencing diseases.

The novel science of data mining is among the 10 developing sciences which have made

[*]*Corresponding author.*
Email: *lethanhbinh@qnu.edu.vn*

the next decade face enormous technological evolutions. Using specialized knowledge, it will have extensive applications in the domain of medicine.[9,10] Predictive modeling of the risk of CVDs can render valuable information for planning of health care interventions. Over the last decade, several models have been developed and validated. The first well-known was developed in the United States using data from the Framingham study. This model, however, might be less reliable in other populations and has overestimated or underestimated the CVDs risks in specific settings.[11,12] Accordingly, other models were developed, such as the Systematic Coronary Risk Evaluation model in Europe[13]; a risk model based on QRESEARCH database in the United Kingdom[14]; the Prospective Cardiovascular Munster study in Germany[15] and a risk model based on database of CUORE Cohorts Project in Italy.[16]

In Asia, models to predict CVDs risk have been developed in Thailand, China, Japan, Malaysia, and Singapore. For all Asian populations, a tool was developed on the basis of six cohorts in this region (Asian model).[11,17-22] In Vietnam, with its own environment of biological, behavioral, and social characteristics, there is not yet a specific model to predict CVDs. The largest survey on risk factors for CVDs done in Vietnam applied the Framingham model only.[23]

The literature review showed that different algorithms such as clustering, logistic regression, decision trees, Bayesian network, neural network, scaled conjugate gradient (SCG) and support vector machine (SVM) have been used for predicting CVDs.[24-28] Among these algorithms, logistic regression has some advantages, such as high speed, simplicity. Logistic regression belongs to a family, named *Generalized Linear Model* (GLM), developed for extending the linear regression model to other situations. It is a widely used technique because it is very efficient and does not require too many computational resources.[29] Logistic regression model output is very easy to interpret compared to other classification methods. In addition, because of its simplicity it is less prone to overfitting than flexible methods such as decision trees. For this reason, we would like to use logistic regression in the present article in order to create a prediction model of coronary heart disease based on the Evans heart disease dataset available in R (*version 3.6.1*).

2. METHODS AND DATA

2.1. Logistic regression model

The logistic regression model is a type of predictive model that can be used when the response variable is binary, meaning that there are only two possible outcomes such as *live/die, disease/no disease, purchase/no purchase, and win/lose.*[29,30] In short, we want to model the probability of getting a certain outcome, in effect modeling the mean of the variable (which is the same as the probability in the case of binary variables). A logistic regression model can be applied to response variables with more than two categories; however, those cases are less common.

As the responses are not on a continuous measure and as such is not continuous, the use of logistic regression differs somewhat from the well-known linear regression, because, while in both cases we are modeling the mean, the mean in linear regression lies anywhere between $(-\infty, +\infty)$ whereas the mean (or the probability) in logistic regression lies between [0, 1]. Thus, we are predicting the probability that Y is equal to 1 (rather than 0) given certain cases of the predictors $X_1,..., X_p$.[30] It is important to make the distinction between these and linear regression models so we can think about how the observed data may be 0 or 1, but the predicted value may lie between [0, 1]. For example, we might try to predict the probability of whether a patient will live or die based on the patient's age as well as the number of years of experience his or her operating physician has.

The general form of the the logistic regression model is[29,30]

$$\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n.$$

where p_1 is the probability that Y = 1 (the event), given $X_1,\dots, X_n$ are the predictors (covariates), and β_i , i = 1, 2, …, n are known as the regression coefficients, which have to be estimated from the data. One can see that logistic regression model forms a linear combination of the explanatory variables to impact the logit, which is *log{probability of event/ probability of nonevent}.*

2.1.1. Probability

Let probability p_1 denote success and $(1 - p_1)$ denote failure with the results constrained to lie between 0 and 1. On the probability scale, we define[29,30]

$$p_1 = \frac{exp[\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n]}{1 + exp[\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n]}$$

The constraints of $0 \le p_1 \le 1$ make it impossible to construct a linear equation for predicting probabilities.

2.1.2. Odds

Odd is the ratio of the probability of an event to the probability of a nonevent. For example, flipping a coin and getting a head as an event versus getting tail as the nonevent. On the odds scale, we define[29, 30]

$$\text{odds} = \frac{p_1}{1-p_1}$$
$$= \exp[\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n].$$

They are constrained by $0 \le \frac{p_1}{1-p_1} < \infty$, with 1 as the point for which both outcomes are equally likely.

2.1.3. Logits

Logit is the natural logarithm of the odds. On the logit scale, we define[29,30]

$$\text{logit}(p_1) = log\ odds = \log\left(\frac{p_1}{1-p_1}\right)$$
$$= \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n.$$

On this scale, we have linearity. The logits are symmetric. They lie in the range $-\infty$ to $+\infty$. The value that is equally likely for both outcomes is 0. If the identification of the two outcomes are switched, the log odds are multiplied by -1, since log(a/b) = – log(b/a). The log odds of an event relays equally the same message as the probability of the event, so if a certain predictor has a positive impact on the logit then it has the same directional effect on the odds. When the log odds take on any value between -∞ and +∞, the coefficients form a logistic regression equation that can be interpreted in the usual way, meaning that they represent the change in log odds of the response per unit change in the predictor.[30]

2.2. Description of dataset

Data on coronary heart disease risk factors in Vietnam are currently limited, whereas the default installation of R comes with several relevant built-in datasets. For this reason, to illustrate the application of logistic regression analysis in the current study, we'll be working on the *Evans dataset* in the *lbreg package*. The data are from a cohort study in which 609 while males were followed for 7 years, with coronary heart disease as the outcome of interest. The variables are defined as in Table 1.[29]

Table 1. Description of Evans dataset.

CHD	A dichotomous outcome variable indicating the presence (coded 1) or absence (coded 0) of coronary heart disesase.
CAT	A dichotomous predictor variable indicating high (coded 1) or normal (coded 0) catecholamine level.
AGE	A continuous predictor variable for age.
CHL	A continuous predictor variable for cholesterol, mg/dl.
SMK	A dichotomous predictor variable indicating whether the subject has ever smoked (coded 1) or never smoked (coded 0)
ECG	A dichotomous predictor variable indicating the presence (coded 1) or absence (coded 0) of electrocardiogram abnormality
DBP	A continuous variable for diastolic blood pressure, mmHg

SBP	A continuous variable for systolic bood pressure, mmHg
HPT	A dichotomous variable indicating the presence (coded 1) or absence (coded 0) of high blood pressure. HPT is coded 1 if the diastolic blood pressure is greater than or equal to 160 or the systolic bood pressure is greater than or equal to 95.

Furthermore, the readers should notice that there is a slight difference about the notations. Namely, the outcome variable (indicating coronary heart disease) in the Evans dataset in R is denoted by “CDH” instead of “CHD” as described before. Accordingly, when executing the R codes which will appear in this article, we have to use “CDH” for the outcome variable.

3. RESULTS AND DISCUSSION

3.1. Data preparation

When working with a real dataset we need to take into account the fact that some data might be missing or corrupted, therefore we need to prepare the dataset for our analysis. As a first step we load the data.

```
# Load the required R package
library(lbreg)
# Load the dataset
data(Evans)
# Give the dimensions of the dataset
dim(Evans)
# Give the names of the variables
names(Evans)
```

Then we get the following:

```
[1]  609  9
[1] “CDH” “CAT” “AGE” “CHL” “SMK”
“ECG” “DBP” “SBP” “HPT”
```

We introduce the *is.na( )*[31,32] function as a tool for finding missing values. By *is.na(Evans),* we can verify that the Evans dataset has no missing values. Moreover, a visual take on the missing values might be helpful: the *Amelia package* has a special plotting function *missmap()*[31,32] that will plot dataset and highlight missing values. With the following R code,

```
# Load the required R package
library(Amelia)
# Draw a map of the missingness in the dataset
missmap(Evans, main=“”, col=c(4,6))
```

we have the output as Figure 1. The “*col”* parameter is to choose the colors we want.

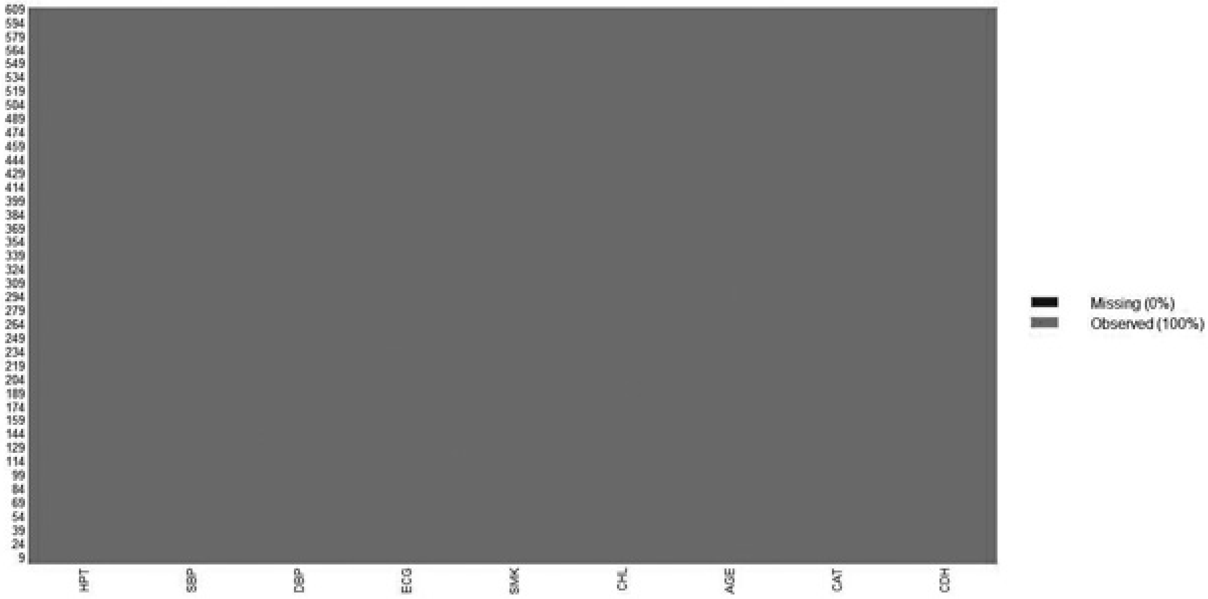


Figure 1. Missing values and observed values.

3.2. Model fitfing

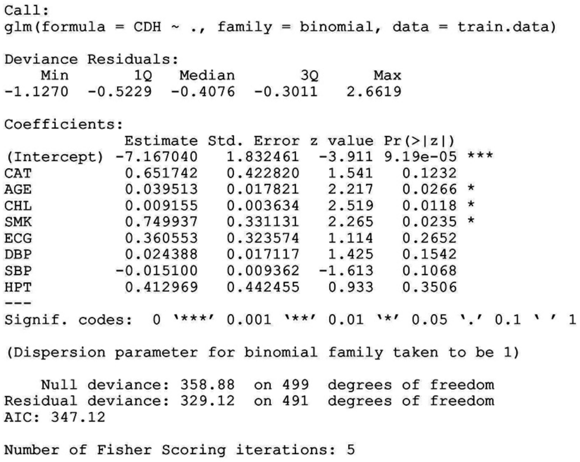
R programing language makes it very easy to fit a logistic regression model with the *glm( )* function and the fitting process is not so different from the one used in linear regression. In this section, we would like to fit a binary multiple logistic regression model corresponding to the Evans dataset and explain each step. We'll now split the data into two parts: *training set* for building a predictive model and *testing set* for evaluating the model.

```
# Split the data into two parts
train.data <-  Evans[1:500,]
test.data <- Evans[501:609,]
```

The function *glm()*[31,32], for generalized linear model, is used to compute logistic regression. We need to specify the option *"family = binomial*", which tells to R that we want to fit logistic regression.

The following R code is to build a model to predict the probability of being coronary heart disease-positive (CHD-positive) based on all the predictor variables available in the dataset. This is done using "~.". Using the *summary()*[31,32] function, we obtain the results of our model as described in Figure 2.

```
# Fit the model
model <- glm(CDH ~ ., data = train.data,
family = binomial)
summary(model) # Summarize the model
```



```
Call:
glm(formula = CDH ~ ., family = binomial, data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1270  -0.5229  -0.4076  -0.3011   2.6619

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.167040   1.832461  -3.911 9.19e-05 ***
CAT          0.651742   0.422820   1.541   0.1232
AGE          0.039513   0.017821   2.217   0.0266 *
CHL          0.009155   0.003634   2.519   0.0118 *
SMK          0.749937   0.331131   2.265   0.0235 *
ECG          0.360553   0.323574   1.114   0.2652
DBP          0.024388   0.017117   1.425   0.1542
SBP         -0.015100   0.009362  -1.613   0.1068
HPT          0.412969   0.442455   0.933   0.3506
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 358.88  on 499  degrees of freedom
Residual deviance: 329.12  on 491  degrees of freedom
AIC: 347.12

Number of Fisher Scoring iterations: 5
```

Figure 2. The results of the logistic regression.

From the output above, the coefficients table shows the beta coefficient estimates and their significance levels. Columns are:[31,32]

- *Estimate*: the intercept (β_0) and the beta coefficient estimates associated to each predictor variable.
- *Std. Error*: the standard error of the coefficient estimates. This presents the accuracy of the coefficients. The larger the standard error, the less confident we are about the estimate.
- *z value*: the z-statistic, which is the coefficient estimate (column 2) divided by the standard error of the estimate (column 3).
- *Pr(> |z |):* The *p-value* corresponding to the z-statistic. The smaller the p-value, the more significant the estimate is.

3.3. Interpretation

Now we can analyze the fitting and interpret what the model is telling us. First of all, it can be seen that only 3 out of 8 predictors are significantly associated to the outcome. These includes: AGE, CHL and SMK. As for the statistically significant predictors, CHL has the lowest p-value suggesting a strong association of the cholesterol concentration of the individuals with the probability of being coronary heart disease-positive (CHD-positive).

The coefficient estimate of the variable AGE is 0.039513, which is positive. This means that an increase in AGE is associated with increase in the probability of CHD-positive. Similarly for the variables CHL, SMK. We also see that the coefficient estimate of the variable SBP is negative (note that SBP is not statistically significant), therefore an increase in SBP will be associated with a decreased probability of being CHD-positive.

On the other hand, recall that in the logit model the response variable is given as

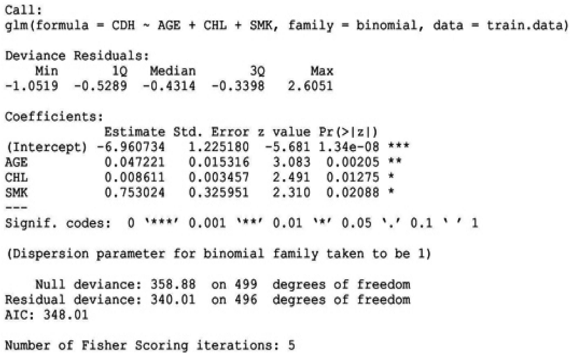
$$log\ odds = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$$

The coefficients for CHL and SMK are 0.009155 and 0.749937, respectively. Since SMK is a dummy variable, being having ever smoked increases the log odds by 0.749937, while one unit increases in the cholesterol concentration increases the log odds by 0.009155. As for the predictor AGE, one unit increases in AGE will increase the log odds by 0.039513.

From the logistic regression results, one should notice that some variables (including CAT, ECG, DBP, SBP and HPT) are not statistically significant. Keeping them in the model may contribute to overfitting. Consequently, they should be eliminated to obtain an optimal model with a reduced set of variables, without compromising the model accuracy. Here, we select manually the most significant variables:

```
# Fit the reduced model with only three
predictor variables
model1 <- glm(CDH ~ AGE + CHL +
SMK, data = train.data, family = binomial)
# Summarize the reduced model
summary(model1)
```

The reduced model is named *model1* and the results of this model, by *summary(model1),* is given as Figure 3:



```
Call:
glm(formula = CDH ~ AGE + CHL + SMK, family = binomial, data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0519  -0.5289  -0.4314  -0.3398   2.6051

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.960734   1.225180  -5.681 1.34e-08 ***
AGE          0.047221   0.015316   3.083  0.00205 **
CHL          0.008611   0.003457   2.491  0.01275 *
SMK          0.753024   0.325951   2.310  0.02088 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 358.88  on 499  degrees of freedom
Residual deviance: 340.01  on 496  degrees of freedom
AIC: 348.01

Number of Fisher Scoring iterations: 5
```

Figure 3. The results of the reduced model.

3.4. Making predictions

The reduced logistic model can be written as

$$p_1 = \exp(-6.9607 + 0.0472\times AGE + 0.0086\times CHL + 0.753\times SMK)/ [1 + \exp(-6.9607 + 0.0472\times AGE + 0.0086\times CHL + 0.753\times SMK)],$$

where p_1 is *the probability of being CHD-positive*. Using this formula, we'll make predictions using the testing data in order to evaluate the performance of our logistic regression model. The proceduce is twofold:

- Predict the class membership probabilities of observations based on predictor variables;
- Assign the observations to the class with highest probability score (i.e. above 0.5).

The function *predict()*[31,32] can be used for predicting the probability of being CHD-positive, given the values of the predictors. Using the option *type = "response"* to directly get the probabilities.

Predict the probabilities of being CHD-positive:

```
# For easy data manipulation and visualization
library(tidyverse)
#Predict the probability of being CHD-positive
probabilities <- model1 %>% predict(test.
data, type = "response")
# See some first probabilities
head(probabilities)
```

The output is as follows:

```
        501          502          503
 0.12714064   0.10954876   0.13142426
        504          505          506
 0.07268374   0.14418855   0.36544538
```

Which classes do these probabilities refer to? In our case, the output is the probability that the coronary heart disease test will be positive.

Predict the class of individual: The following R code categorizes individuals into two groups based on their predicted probabilities of being coronary heart disease-positive. Individuals with probability above 0.5 (random guessing) are considered as CHD-positive.

```
# Predict the class of individual
predicted.classes <- ifelse(probabilities > 0.5,
"pos", "neg")
head(predicted.classes)
```

The output is as follows:

```
  501   502   503   504   505   506
"neg" "neg" "neg" "neg" "neg" "neg"
```

3.5. Assessing model accuracy

The model accuracy is measured as *the proportion of observations that have been correctly classified.*[29,30] Inversely, the classification error is defined as the proportion of observations that have been misclassified. We now use the following R code to compute the proportion of correctly classified observations:[30,31]

```
# Classify the individuals in the testing set.
test.data_CDH <- if.else(test.data$CDH ==
1, “pos”, “neg”)
# Compute and display the model accuracy
accuracy <- mean(predicted.classes ==
test.data_CDH)
print(paste(‘Accuracy:’, accuracy))
```

Then the output is as follows:

```
[1] “Accuracy: 0.880733944954128”
```

From above, the *0.88 accuracy* on the testing set is quite a good result. However, you should keep in mind that this result is somewhat dependent on the manual split of the data that we made earlier.

In conclusion, with the ROCR package[30,31], we are going to plot the ROC curve and calculate the AUC (area under the ROC curve) which are typical performance measurements for a binary classifier. The ROC[29,30] is a curve generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings while the AUC[29,30] is the area under ROC curve. As a rule of thumb, a model with good predictive ability should have an AUC closer to 1 than to 0.5.

```
# Load the ROCR package
library(ROCR)
#Predict the probability of being CHD-
positive
probabilities <- model1 %>% predict(test.
data, type = “response”)

pr <- prediction(probabilities, test.data$CDH)
perf <- performance(pr, measure = “tpr”,
x.measure = “fpr”)

# Plot the ROC curve
plot(perf, xlab = “False positive rate”, ylab
= “True positive rate”, col = “blue”)

# Calculate and display the AUC
auc <- performance(pr, measure = “auc”)
auc <- auc@y.values[[1]]
print(paste(‘AUC:’, auc))
```

When excuting the above R code, we will get the ROC plot as shown in Figure 4 and the AUC:

```
[1] “AUC: 0.729967948717949”
```

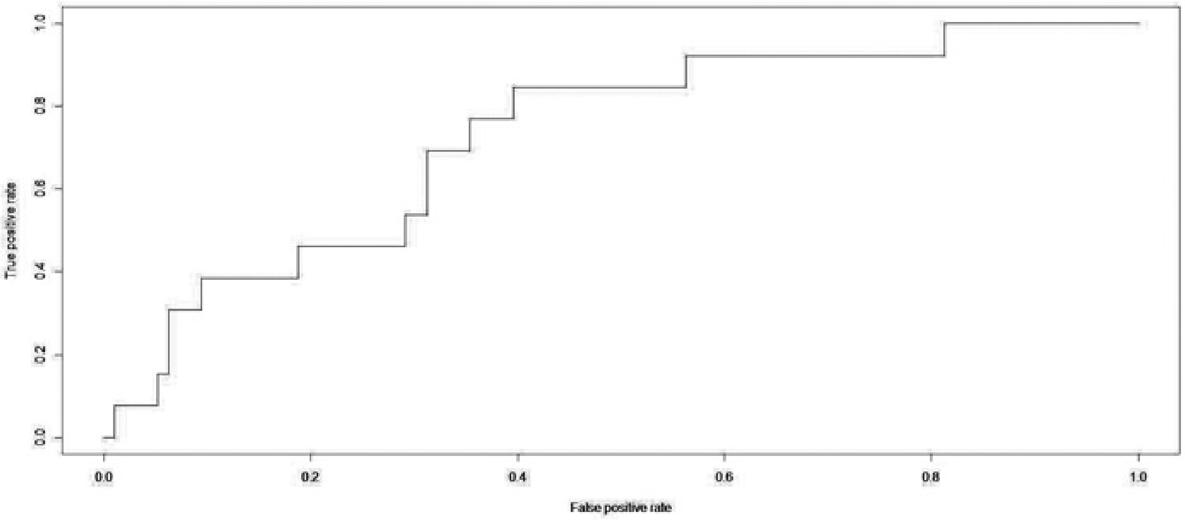


Figure 4. The ROC plot.

4. CONCLUSIONS

The process of disease prediction in medical sciences is as a very important process for decision-making and physicians need to know the risk factors for different diseases. This process can be facilitated by using statistical methods and data mining algorithms, such as logistic regression. In this study, we have described how logistic regression works and provided R codes to compute logistic regression. Moreover, we demonstrated how to make predictions and to assess the model accuracy. The results of the current study showed that the use of logistic regression can be very useful in predicting coronary heart disease and a simple coronary heart disease prediction model was proposed, having the 0.88 accuracy on the testing set, which allows physicians to predict CHD risk in patients. As described before, several CVDs (including CHD) prediction models are available around the world in general and in Vietnam in particular. These models vary in different aspects such as the time horizon used, characteristics of study population included, predictor variables, and outcome, they may all produce different results. Nevertheless, due to its advantages, logistic regression is still a powerful tool, especially in epidemiologic studies, allowing multiple predictor variables being analyzed simultaneously, meanwhile reducing the effect of confounding factors. With the gradual rise in CHD events among Vietnamese, the prevention and treatment of CHD risk factors are important public health concerns. Vietnam should of course ideally have its own cohort to feed models to predict CHD risk and validate models for the national setting. Even building an own Vietnamese model could be considered, may be an adaptation of one of the existing models. Because of the long time needed for that, and because of resource constraints, the suggestion that for Vietnamese populations, our logistic regression model can provide valid and reliable results should however be investigated in further analyses using real-life data for potential confirmation. Also, the current results of the study may be useful to health planners and will provide a basis for a further study on cost-effectiveness modeling of CHD management. In conclusion, we have to emphasize that further research is needed to compare the performance of different models, thereby identifying which models are best for Vietnamese populations.

Acknowledgement

This study is conducted within the framework of science and technology projects at institutional level of Quy Nhon University under the project code T2018.554.03.

REFERENCES

1. WHO. 2002. The World Health Report: Reducing Risks, Promoting Healthy Life, <https://www.who.int/whr/2002/en/>, accessed 05/02/2020.
2. WHO. 2017. The top ten causes of death, < https://www.who.int/news-room/fact-sheets/detail-/the-top-10-causes-of-death>, accessed 05/02/2020.
3. WHO. 2018. Physical activity and older adults, <https://www.who.int/dietphysicalactivity/factsheet_olderadults/en/>, accessed 07/02/2020.
4. N. P. Hoa, C. Rao, D. Hoy, N. D. Hinh, N. T. K. Chuc, N. D. Anh. Mortality measures from sample-based surveillance: evidence of the epidemiological transition in Vietnam, *Bulletin of the World Health Organization,* **2012**, *90*(10), 764-772.
5. Ngo D. A, C. Rao, N. P. Hoa, T. Adair, N. T. K. Chuc. Mortality patterns in Vietnam, 2006: findings from a national verbal autopsy survey, *BMC Research Notes,* **2010**, *3*(1).
6. H. V. Minh, D. L. Huong, S. Wall, N. T. K. Chuc. Cardiovascular disease mortality and its association with socioeconomic status: findings from a populationbased cohort study in rural Vietnam, 1999 - 2003, *Preventing Chronic Disease,* **2006**, *3*(3), 1-11.
7. D. L. Huong, H. V. Minh, U. Janlert, D. D. Van. Socio-economic status inequality and major causes of death in adults: a 5-year follow-up study in rural Vietnam, *Public Health,* **2006**, *120*(6), 497-504.
8. P. Rezaei, M. Ahmadi, S. Alizadeh, F. Sadoughi. Use of data mining techniques to determine and predict length of stay of cardiac patients, *Healthc. Inform. Res.,* **2013**, *19*(2), 121-129.

9. R. Bellazzi, F. Ferrazzi, L. Sacchi. Predictive data mining in clinical medicine: A focus on selected methods and applications, *Wiley interdisciplinary reviews: WIREs data mining and knowledge discovery*, **2011**, *1*(5), 416-430.

10. F. Amato, A. López, E. Pena-Méndez, P. Vanhara, A. Hampl, J. Havel. Artificial neural networks in medical diagnosis, *J. Appl. Biomed.,* **2013**, *11*, 47-58.

11. J. Liu, et al. Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial Cohort Study, *JAMA*, **2004**, *291*, 2591-9.

12. S. Kanjilal, *et al.* Application of cardiovascular disease risk prediction models and the relevance of novel biomarkers to risk stratification in Asian Indians, *Vasc. Health. Risk. Manag.,* **2008**, *4*, 199-211.

13. RM. Conroy, *et al.* Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project, *Eur. Heart. J.*, **2003**, *24*, 987-1003.

14. J. Hippisley-Cox, et al. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study, *BMJ*, **2007**, 335-136.

15. G. Assmann, P. Cullen, H. Schulte. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the Prospective Cardiovascular Munster (PROCAM) study. *Circulation*, **2002**, *105*, 310-5.

16. M. Ferrario, et al. Prediction of coronary events in a low incidence population: assessing accuracy of the CUORE Cohort Study prediction equation, *Int. J. Epidemiol.,* **2005**, *34*, 413-21.

17. P. Sritara, *et al.* Twelve-year changes in vascular risk factors and their associations with mortality in a cohort of 3499 Thais: the Electricity Generating Authority of Thailand Study, *Int. J. Epidemiol.*, **2003**, *32*(3), 461-8.

18. Y. Wu, *et al.* Estimation of 10-year risk of fatal and nonfatal ischemic cardiovascular diseases in Chinese adults, *Circulation*, **2006**, *114*(21), 2217-25.

19. I. Saito, *et al.* A low level of C-reactive protein in Japanese adults and its association with cardiovascular risk factors: the Japan NCVC-Collaborative Inflammation Cohort (JNIC) study, *Atherosclerosis*, **2007**, *194*(1), 238-44.

20. H. Arima, et al. Development and validation of a cardiovascular risk prediction model for Japanese: the Hisayama study, *Hypertens. Res.*, **2009**, *32*(12), 1119-22.

21. C. Purwanto Eswaran, *et al*. Prediction models for early risk detection of cardiovascular event, *J. Med. Syst.*, **2012**, *36*(2), 521-31.

22. I. Baik, *et al.* Dietary information improves cardiovascular disease risk prediction models, *Eur. J. Clin. Nutr.*, **2013**, *67*(1), 25-30.

23. N. Q. Ngoc, *et al*. Cardiovascular disease risk factor patterns and their implications for intervention strategies in Vietnam, *Int. J. Hypertens.*, **2012**.

24. S. Desai, S. Giraddi, P. Narayankar, S. Sulegaon, N. Pudakalakatti. Back-propagation neural network versus logistic regression in heart disease classification, *J. Adv. Comput. Commun. Technol.*, **2019**.

25. N. Kausar, *et al*. Ensemble clustering algorithm with supervised classification of clinical data for early diagnosis of coronary artery disease, *J. Med. Imaging Health Inform.*, **2016**, *6*(1),78-87.

26. L. Guner, N. Karabacak, O. Akdemir, P. Karagoz, S. Kocaman, A. Cengel, M. Unlu. An open-source framework of neural networks for diagnosis of coronary artery disease from myocardial perfusion SPECT, *J. Nucl. Cardiol.,* **2010**, *17*(3), 405-13.

27. K. Orphanou, A. Stassopoulou, E. Keravnou. DBN-extended: a dynamic Bayesian network model extended with temporal abstractions for coronary heartdisease prognosis, *IEEE J. Biomed. Health Inform.*, **2016**, *20*(3), 944-52.

28. J. Kim, J. Lee, Y. Lee. Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree, *Healthc. Inform. Res.*, **2015**, *21*(3), 167-74.

29. D. Kleinbaum, M. Klein. *Logistic Regression: A Self-Learning Text*, Springer, 2010.

30. J. Wilson, K. Lorenz. *Modeling Binary Correlated Responses using SAS, SPSS and R,* Springer, 2015.

31. J. Ledolter. *Data Mining and Bussiness analytics with R,* John Wiley & Sons, 2013.

32. Y. Li, J. Baron. *Behavioral Research Data Analysis with R,* Springer, 2012.