World Academy of Science, Engineering and Technology
International Journal of Biomedical and Biological Engineering
Vol:9, No:11, 2015

Applying the Regression Technique for Prediction of the Acute Heart Attack

Paria Soleimani, Arezoo Neshati

Abstract**—**Myocardial infarction is one of the leading causes of death in the world. Some of these deaths occur even before the patient reaches the hospital. Myocardial infarction occurs as a result of impaired blood supply. Because the most of these deaths are due to coronary artery disease, hence the awareness of the warning signs of a heart attack is essential. Some heart attacks are sudden and intense, but most of them start slowly, with mild pain or discomfort, then early detection and successful treatment of these symptoms is vital to save them. Therefore, importance and usefulness of a system designing to assist physicians in early diagnosis of the acute heart attacks is obvious.

The main purpose of this study would be to enable patients to become better informed about their condition and to encourage them to seek professional care at an earlier stage in the appropriate situations. For this purpose, the data were collected on 711 heart patients in Iran hospitals. 28 attributes of clinical factors can be reported by patients; were studied. Three logistic regression models were made on the basis of the 28 features to predict the risk of heart attacks. The best logistic regression model in terms of performance had a C-index of 0.955 and with an accuracy of 94.9%. The variables, severe chest pain, back pain, cold sweats, shortness of breath, nausea and vomiting, were selected as the main features.

Keywords**—**Coronary heart disease, acute heart attacks, prediction, logistic regression.

Open Science Index, Biomedical and Biological Engineering Vol:9, No:11, 2015 waset.org/Publication/10002699

I. Introduction

CARDIOVASCULAR DISEASE is now along with the top three causes of death and disability worldwide and has become the major cause of mortality or morbidity in most countries. Each year, approximately 32 million heart attacks and strokes occur in the world which is causing the deaths of more than 17 million people. 60% of all deaths worldwide in 2000 occurred due to non-communicable diseases and is estimated to reach 73% by 2020. Share of the cardiovascular disease by more than 48 percent. So that more than 20 million of the 64 million deaths in 2015 will be related to cardiovascular disease. And if the effective measures are not taken, it is expected deaths due to chronic diseases increased 17% from 2005 to 2015 and the 35 million deaths rise to 41 million deaths [1]. In recent years, due to preventive measures and effective interventions, deaths due to cardiovascular diseases in developed countries has been declining trend, in contrast in developing countries is still rising. Unfortunately, the patients often delay substantially before seeking care that this is due to several factors, including lack of understanding by the patients of the symptoms of MI (Myocardial infarction), the emotional factors, and the inadequate advice by the health care workers when the patients develop symptoms.

Early identification of these diseases is critical to successful treatment. There is evidence that the patients with better knowledge of the symptoms of MI will request help earlier. If the patients could get rapid and accurate advice on whether their symptoms were likely to be serious, then it is possible that delays in seeking treatment could be further reduced.

The purpose of this study is to determine how well a predictive model would perform based only upon patient-reportable clinical history factors, without using diagnostic tests or physical exam findings. Even though we would not expect such a model to perform as well as one using these strong predictors, however the model may have important practical applications. This type of prediction model might have application outside of the hospital setting to give accurate advice to patients to influence them to seek care in appropriate situations. For example, such a system might be used directly by patients in a patient-oriented software application, or might be used by healthcare workers as a decision support aid in telephone nurse triage.

We reviewed some works of previous related researches as:

Reference [2] shows that they compared performance of the prediction models, logistic regression, decision tree and neural network for the diagnosis of acute cardiac ischemia in the emergency department patients with clinical data (include: patient-reportable history, physical exam findings, and electrocardiogram (ECG)). All these methods could predict very well however the choice between these methods should be based on the specific application requirements and not on the assumption that naturally one stronger than the others.

Reference [3] shows that they paid on the early diagnosis of acute myocardial infarction based on clinical data and electrocardiographic data by applying regression models.

Reference [4] shows that they provided a consensus approach to diagnosing coronary artery disease based on the clinical and exercise tests.

Reference [5] shows that they compared the performance of a logistic regression model and an artificial neural network to predict the risk of myocardial infarction in the patients based on the patient-reportable clinical history factors. They concluded that the performance of both logistic regression and artificial neural network model were good and acceptable and there was no statistically significant difference between them. The best performing logistic regression model and neural

Paria Soleimani, Assistant Professor, is with the Department of Industrial Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran (phone: 982177638853; fax: 982177638845; e-mail: p_soleimani@azad.ac.ir).

Arezoo Neshati is with the Department of Industrial Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran (e-mail: arezoo_neshati@yahoo.com).

Open Science Index, Biomedical and Biological Engineering Vol:9, No:11, 2015 waset.org/Publication/10002699

network model had C-index of 0.8444 and 0.8503, respectively. In our study, the best logistic regression model in terms of performance had a C-index of 0.955 and with an accuracy of 94.9% to predict the risk of heart attacks.

Reference [6] shows that they offered a model using artificial neural networks for the automatic detection of the acute myocardial infarction in patients on 12- lead ECGs[1].

Reference [7] shows that they presented an artificial neural network model to predict acute coronary syndrome using clinical data from one survey. This study confirms that artificial neural networks can be useful for the development of a diagnostic algorithm for patients with chest pain.

Reference [8] shows that he studied to predict the risk of heart disease by using weighted fuzzy rules, as the clinical decision support system.

Reference [9] shows that they conducted a study to feature selection in ischemic heart disease identification using feed forward neural networks, when the input features were 12 items. The predicted accuracy during training was high as 89.4% and during testing was high as 82.2%. Further removal of the features lowered the accuracy and hence the interesting features selected for prediction was concluded to be as 12 for that IHD (Ischemic Heart Disease) data set.

Reference [10] shows that they designed a model using data mining techniques to predict myocardial infarction.

Reference [11] shows that they conducted a study with the goal of making an efficient system with fully automated classifier to detect the presence of ischemic heart disease. They applied the artificial intelligence techniques.

Reference [12] shows that they developed an artificial neural networks-based (ANNs) diagnostic model for coronary heart disease (CHD) using a complex of traditional and genetic factors of this disease.

Reference [13] shows that they compared the performance of methods from the data-mining and machine-learning literature with that of conventional classification trees to classify patients with heart failure (HF) according to the following subtypes: HF with preserved ejection fraction (HFPEF) and HF with reduced ejection fraction. They also compared the ability of these methods to predict the probability of the presence of HFPEF with that of conventional logistic regression.

Reference [14] shows that they compared performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease.

This paper is organized as follows: In Section II, the proposed models along with its assumptions is presented. The performance of methods is evaluated in Section III. Our concluding remarks and discussion are presented in the final section.

II. The Proposed Model

A. Data Collection

This is a diagnostic study that deals with predicting the probability of acute myocardial infarction. The data set consisted of 711 patients who presented to the emergency room in Hamedan Ekbatan hospital in Iran from 2013 to 2014. After identifying and removing outliers, the final set consisted of 663 patients with a mean age of 63.29 years and the standard deviation of 14.37. Data set contained 28 potential prediction attributes based on the experts' considerations. A binary outcome variable indicates the presence or absence of acute heart attack. The 28 potential prediction attributes contained the clinical patient-reportable history factors only. Table I shows the list of 28 patient-reportable history factors that were included as potential covariates in our models. All variables were binary (0= absent or 1= present) except for 'Age' (years), 'Sex' refers to the gender of the patient, where 0= female and 1= male.

TABLE I
The Patient-Reportable Clinical History Factors That Were Candidates for Inclusion as Predictor Covariates in the Models

Variable	Type
1Age (years)	Numeric
2Sex (1=male)	Categorical
3Smokes	Categorical
4Previous MI	Categorical
5Diabetes	Categorical
6Hypertension (HTN)	Categorical
7Hyperlipidemia (HLP)	Categorical
8Severe chest pain (CP)	Categorical
9Left chest pain (LCP)	Categorical
10Right chest pain (RCP)	Categorical
11Back pain (BP)	Categorical
12Left Arm pain (LAP)	Categorical
13Right Arm pain (RAP)	Categorical
14Sweats	Categorical
15Shortness of breath (SOB)	Categorical
16Nausea	Categorical
17Vomiting	Categorical
18Syncope	Categorical
19Palpitations	Categorical
20Epigastric pain	Categorical
21History of heart disease	Categorical
22Edema	Categorical
23Drowsiness	Categorical
24Dizziness	Categorical
25Weakness	Categorical
26Cough	Categorical
27Anxiety	Categorical
28Headache	Categorical

B. Logistic Regression

Logistic regression analysis used to analyze the data according to the response variable [15]. The IBM SPSS Statistics was used for building the logistic regression (LR) models. We built three types of LR models with all the 28 patient history factors as covariates for comparison. The models were built using the automatic stepwise, forward, and

[1] 12- lead ECGs: The standard 12-lead electrocardiogram is a representation of the heart's electrical activity recorded from electrodes on the body surface. This section describes the basic components of the ECG and the lead system used to record the ECG tracings.

backward variable selection algorithms using α = 0.05 as the entry and exit criteria [16]. Three models based on three algorithms called model (1), model (2) and model (3), respectively.

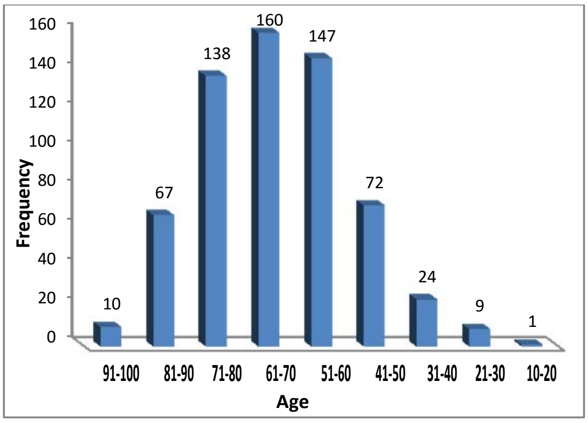


Fig. 1 Chart based on the age distribution of patients with acute heart attack

For logistic regression models building, the data set of 663 cases was randomly split into 80% train and 20% test sets.

1) The first model:

$$\text{logit(P)} = \ln\frac{\text{P}}{1-\text{P}} = -3.125 + 2.996\,CP + 4.552\,BP + 4.030\,Sweats + 4.985\,SOB - 5.668\,Nausea + 5.329\,Vomiting \quad (1)$$

where the variables are: severe chest pain, back pain, cold sweats, shortness of breath, nausea and vomiting.

2) The second model:

$$\text{logit(P)} = \ln\frac{\text{P}}{1-\text{P}} = -1.708 + 0.38\,Age + 1.736\,CP + 1.665\,BP + 2.183\,Sweats + 2.898\,SOB - 1.946\,Cough \quad (2)$$

where the variables are: age, severe chest pain, back pain, cold sweats, shortness of breath and cough.

3) The third model:

$$\text{logit(P)} = \ln\frac{\text{P}}{1-\text{P}} = -1.225 + 2.481\,HTN + 4.167\,HLP + 2.926\,CP + 3.848\,BP + 2.914\,Sweats + 4.771\,SOB - 5.892\,Nausea + 3.487\,Vomoting - 2.553\,Coug \quad (3)$$

where the variables are: hypertension, hyperlipidemia, severe chest pain, back pain, cold sweats, shortness of breath, nausea, vomiting and cough.

Note that some predictive factors had negative coefficient in all three models. For factors such as Nausea, Cough and Vomiting, this is not surprising, because, according to the cardiologists, these factors are not specific risk factors for the acute heart attack. For example, the occurrence of severe chest pain and cough symptoms in a patient with a history of lung disease reduces the risk of the acute heart attacks. Therefore in the medicine, this negative coefficient is reasonable and justified.

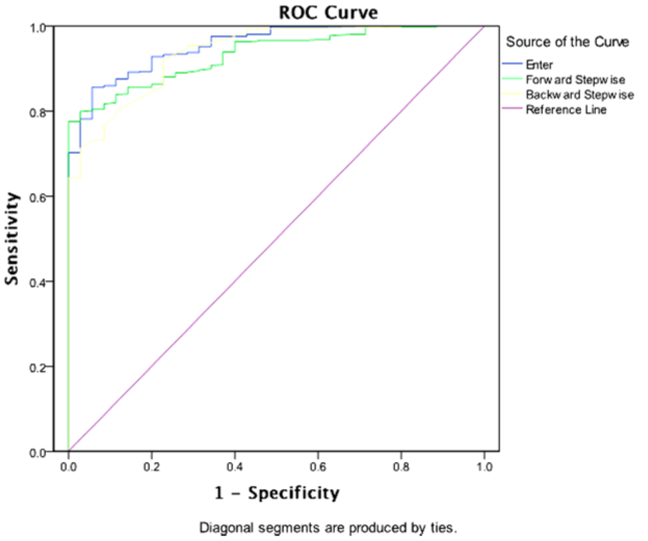
III. Calculation and Analysis

The purpose of this study is to determine how well a predictive model would perform based only upon patient-reportable clinical history factors, without using diagnostic tests or physical exam findings. The discrimination performance of the prediction models is typically measured in terms of the area under the receiver operating characteristic (ROC) curve [17]. The ROC curve for each three type of built regression model is presented in Fig. 2. (C-index is equivalent to the area under the ROC curve). By comparing the C-index for models (1), (2) and (3) that are shown in Table II, it is clear that the model (1) with variable selection algorithm Enter(automatic stepwise) and C-index of 0.955 has the best performance in comparison with the other models.

TABLE II
The Area Under the ROC Curve (Equivalent C-Index) for Models (1), (2), and (3)

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Predicted probability **(Enter)**	**.955**	**.012**	.000	.931	.979
Predicted probability **(Forward)**	**.934**	**.013**	.000	.908	.960
Predicted probability **(Backward)**	**.938**	**.017**	.000	.905	.972

a. Under the nonparametric assumption
b. Null hypothesis: true area = 0.5



Diagonal segments are produced by ties.

Fig. 2 The area under the ROC curve (equivalent C-index) for models (1), (2), and (3)

Summary results of the regression model (1), (2) and (3) are presented in Table III. By comparing these results also can realize that the model (1) with 94.9% accuracy of prediction, Pseudo R-squared between %23 to %67 and chi-square 2.096, is the best model between the other models.

To compare the effect of different randomization splits on the models building, we repeated the randomization process 10 times, and each variable selection algorithms was run on the 10 randomization splits [5]. Table IV shows the results for the average values and range of C-index for the 10 different randomization splits in each of the variable selection methods.

TABLE III
SUMMARY RESULTS OF MODEL (1), (2), AND (3)

Model type	Accuracy	Chi-Square	Cox & Snell R Square	Nagelkerke R Square
Model (1)	**94.9**	**2.096**	**0.229**	**0.673**
Model (2)	**93.7**	**4.225**	**0.157**	**0.476**
Model (3)	**93.9**	**5.797**	**0.218**	**0.644**

TABLE IV
AVERAGE VALUES AND RANGE OF C-INDEX FOR THE 10 DIFFERENT RANDOMIZATION SPLITS IN EACH OF THE VARIABLE SELECTION METHODS

Models	C-index Average	C-index Range
Model (1)Enter	**0.956**	**(0.942 ، 0.964)**
Model (2)Forward	**0.930**	**(0.896 ، 0.953)**
Model (3)Backward	**0.947**	**(0.939 ، 0.957)**

Number of times that each of the independent variables appeared as an effective variable in 10 random repetitions of the models (1), (2), and (3) are shown in Table V. These results also confirm the accuracy of the effective independent variables that have appeared in the models (1), (2), and (3).

IV. DISCUSSION AND CONCLUSION

In this paper, we proposed three logistic regression models to predict the risk of heart attacks on basis of the 28 patient-reportable clinical history factors. We compared performance of the models by the C-index criteria.

The original motivation for this study was to determine if a clinical software application could be built that could successfully predict the likelihood of a myocardial infarction based on clinical history factors alone. As expected, our models do not perform as well as those that used physical findings and ECG (electrocardiogram) data, but they still performed remarkably well even without this objective information.

Our results lead us to believe that these models could be used in real software applications in a clinical setting. If the performance of these models holds up in further studies, software applications could be written using these models that could have important utility in settings outside of a hospital when a healthcare provider may not yet be available. For example, a software application could be designed to assist nurses doing telephone triage when they are assessing a patient's risk of MI over the phone. Alternatively, an application could be designed for direct use by patients to assist them in determining the seriousness of their chest pain symptoms. For example, a standalone software application that assesses chest pain symptoms could be designed to run on a home desktop computer or personal digital assistant (PDA) device. Patients could enter information about their chest pain symptoms and get back an estimate of the likelihood that they are experiencing symptoms of a heart attack and obtain advice on whether or not they should seek professional care.

We made a subjective determination as to which prediction factors could be easily reported by a patient. Further evaluation will be needed to determine if patients can accurately report these data items. The accuracy of these factors will affect obviously the performance of the prediction model.

TABLE V
NUMBER OF APPEARANCES AS AN EFFECTIVE VARIABLE IN 10 RANDOM REPETITIONS OF MODELS (1), (2), (3)

Independent variables	Number of appearances as an effective variable in 10 random repetition of model (1)	Number of appearances as an effective variable in 10 random repetition of model (2)	Number of appearances as an effective variable in 10 random repetition of model (3)
Age (years)	**3**	**3**	**2**
Sex (1=male)			
Smokes			
Previous MI			
Diabetes			
HTN	**4**	**5**	**4**
HLP			**1**
CP	**10**	**10**	**10**
LCP	**5**	**2**	**4**
RCP			
BP	**8**	**7**	**5**
LAP			
R AP			
Sweats	**10**	**9**	**10**
SOB	**10**	**10**	**10**
Nausea	**9**	**6**	**10**
Vomiting	**5**	**3**	**3**
Syncope			
Palpitations		**1**	
Epigastric pain			
History of heart disease			
Edema			
Drowsiness			
Dizziness			
Weakness			
Cough		**3**	
Anxiety			
Headache			

REFERENCES

[1] T. Samavat, A. Hojjatzadeh, M. Shams, A. Afkhami, A. Mahdavi, Sh. Bashti, H. Pouraram, M. Ghotbi, A. Rezvani, Prevention and control of cardiovascular disease (for government employees). Second edition (2012).

[2] H. P. Selker, J. L. Griffith, S. Patil, W. J. Long, R. B. D'Agostino, A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients, J. Investig. Med, 43 (1995) 468-476.

[3] R. L. Kennedy, A.M. Burton, H.S. Fraser, L.N. McStay, R.F. Harrison, Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: derivation and evaluation of logistic regression models, Eur. Heart J. 17 (1996) 1181-1191.

[4] D. Do, J. A. West, A. Morise, E. Atwood, V. Froelicher, A consensus approach to diagnosing coronary artery disease based on clinical and exercise test data, Chest 111 (1997) 1742- 1749.

[5] S. J. Wang, L. Ohno-Machado, H. S. F. Fraser, R. Lee Kennedy, Using patient-reportable clinical history factors to predict myocardial infarction: Computers in Biology and Medicine, 31 (2001) 1-13.

[6] H. Haraldsson, L. Edenbrandt, M. Ohlsson, Detecting acute myocardial infarction in the 12- lead ECG using Hermite expansions and neural networks, Artificial Intelligence in Medicine, 32 (2004) 127-136.

[7] R. F. Harrison, R. L. Kennedy, Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation, Ann Emerg Med, 46 (2005) 431-439.

[8] P. K. Anooj, Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules, Journal of King Saud University – Computer and Information Sciences, 24 (2012) 27-40.

[9] K. Rajeswari, V. Vaithiyanathan, T. R. Neelakantan, Feature Selection in Ischemic Heart Disease Identification using Feed Forward Neural Networks, Procedia Engineering 41 (2012) 1818 – 1823 .

[10] R. Safdari, M. GhaziSaeedi, G. Arji, M. Gharooni, M. Soraki, M. Nasiri, A model for predicting myocardial infarction using data mining techniques, Iranian journal of medical informatics, (2013) vol. 2, issue 4.

[11] Suchithra, P. U. Maheswari, Survey on Clinical Decision Support System for Diagnosing Heart Disease, IJCSMC, (2014) vol. 3, Issue 2, 21-28 .

[12] O. Y. U. Atkov, S. G. Gorokhova, A. G. Sboev, E. V. Generozov, E. V. Muraseyeva, S.Y . Moroshkina, N. N. Cherniy, Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters, Journal of Cardiology, 59 (2012) 190-194.

[13] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, D. S. Lee, Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes, Journal of Clinical Epidemiology 66 (2013) 398-407.

[14] Kurt I., Ture M., Kurum A. T. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert SystAppl (2008) 34(1) 366-374.

[15] M. Scott, Applied logistic Regression Analysis, Second Publication, Sage Publication (2001).

[16] S. Dreiseitl, L. Ohno-Machado, S. Vinterbo, Evaluating variable selection methods for diagnosis of myocardial infarction, Proceedings of AMIA Annual Fall Symposium (1999) pp. 246-250.

[17] M. H. Zweig, G. Campbell, Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine (published erratum appears in Clin. Chem. 39(8) (1993) 1589), Clin. Chem. 39 (1993) 561-577.