Early screening of incidence coronary heart disease based on risk factor using fuzzy rule system

Wiharto Wiharto [1, 2, *], Herianto Herianto [2, 3], Hari Kusnanto [2, 4]

[1]Department of Informatics, Sebelas Maret University, Surakarta, Indonesia
[2]Department of Biomedical Engineering, Gadjah Mada University, Yogyakarta, Indonesia
[3]Department of Mechanical and Industrial Engineering, Gadjah Mada University, Yogyakarta, Indonesia
[4]Department of Medicine, Gadjah Mada University, Yogyakarta, Indonesia

ARTICLE INFO

ABSTRACT

A thorough examination in the context of the diagnosis of coronary heart disease requires a relatively high cost. To reduce the cost of the diagnosis can be done gradually, preceded by the initial screening based on risk factors. Initial screening can be done using a model that has been developed. Unfortunately the development of a model screening, reference to specific populations, such as the Framingham risk scores (FRS), so sometimes does not correspond to other populations. This study proposes a model of screening by combining FRS with artificial intelligence techniques, for initial screening on the risk of coronary heart disease. The research method is divided into several stages, firstly the selection of an attribute of risk factors. The second, modeling of the attribute into a fuzzy rule with reference to a standard FRS. The third, modeling of inference using Mamdani method, and the last of analyzing system performance. The test results show that the model proposed system has the ability, if positive patients tested, capable of producing a really positive output of coronary heart disease in the amount of sensitivity that is 91.37%. The performance is relatively better than a number of previous studies, with only requires an examination of the four attributes, namely age, sex, cholesterol and systolic blood pressure.

1. Introduction

Coronary heart disease (CHD) is one of the biggest causes of death in both developed and developing countries. WHO estimates that approximately 17.5 million people die each year from cardiovascular disease, especially coronary heart disease which is 7.4 million people. Unhealthy lifestyle such as eating fatty foods, lack of exercise, smoking, hypercholesterolemia, and hypertension may increase the likelihood of developing the disease(Lee and Kim, 2016). Lifestyle factors have an impact on the risk of a person against coronary heart disease, in particular, the risks of the factors that can be modified. In addition, risk factors in the modification of, heart disease are also affected by the risk factors that can not be modified, such as age, gender, and heredity (Popovic et al., 2013).

The prediction model has been developed heart disease using risk factor attributes. The prediction models such as the Framingham Risk Score (FRS) (Wilson et al., 1998) and Prospective cardiovascular Munster (PROCAM) (Assmann et al., 2002). In addition to the two standards, also used the Systematic Coronary Risk Evaluation (SCORE), a standard commonly used in European populations (Graham et al., 2007). Standard models that have been developed for predicting coronary heart disease events refer to a particular population, so the resulting model is sometimes not appropriate for different populations. A study conducted by Versteylen et al. (2011) explained that the standard FRS and SCORE provide better results than another standard. It is also reinforced in research Selvarajah et al. (2014) showed that the FRS and SCORE also be used to identify a high-risk cardiovascular population of Malaysia or Asia. The development of information technology brings many changes, especially in the medical world. It has been proven that the use of information technology can provide many benefits, both from the clinicians and patients (Anderson et al., 2013; Garg et al., 2005). The impact of these developments included in the model system

predictive of coronary heart disease events. The system has been developed using artificial intelligence techniques so that the system is capable of being used as a support clinicians in making decisions in the initial screening. Kim et al. (2015) proposed a model system predictive of coronary heart disease, with reference to the Korean dataset using the decision of tree algorithm to perform the extraction rule. Rule-based is then converted into fuzzy rules, while the membership function refers to the Framingham risk score (Kim et al., 2015). The process of decision-making is done by fuzzy inference system. Not much different from Kim et al. (2014) also proposed a model of coronary heart disease prediction system using data from a hospital in Korea using a procedure similar to Kim et al. (2015), which is using fuzzy inference system (FIS). Similar research has also been conducted, which combines Framingham and PROCAM risk score for prediction of coronary heart disease. Merging the two standards using the method of fusion with Dempster-Shafer algorithm (Khatibi and Montazer, 2010). Based on previous studies, this study will propose an initial screening system model to predict the incidence of coronary heart disease based on fuzzy inference system. Extraction is done with the fuzzy rule-based menag the attribute table and score in the FRS, thus gathering of the membership function. Fuzzy rule based generated, used to make decisions by using fuzzy inference system. Inforensi method used is mamdani. System performance measured by using a number of parameters, in particular the sensitivity, specificity, accuracy, positive prediction value (PPV), negative prediction value (NPV) and area under the Curver (AUC).

2. Material and method

2.1. Data

This study uses the dataset is available at the University of California Irvine, which can be accessed online (Detrano et al., 1988). The dataset there are four types of data cleveland with 303, hongarian 294 data, switzerland 123 data and the long beach 200 data. The dataset has 13 attributes of coronary heart disease and one attribute output, with an output value of 0, for normal and 1 for abnormal. The attribute of coronary heart disease consisting of a number of test results that can be grouped into several groups. The group is the risk factors, symptoms, examination electrocardiography (ECG), fluoroscopy and scintigraphy. The focus in this paper emphasizes on risk factors. The risk factors that used a total of four, namely age, gender, total cholesterol and systolic blood pressure at rest.

2.2. Fuzzy inference system

The main concept of the theory of fuzzy logic is mapped an input space into the output space by using the IF-THEN rules. Mapping is done by using a

FIS. FIS evaluate all rule simultaneously to produce results. The fuzzy logic system has a structure consisting of, fuzzification, rule-based, database, inference, and defuzzification. Set of rule-based and database called a knowledge based.

Fuzzification is a process of converting nonfuzzy variables into fuzzy variables. Rule-based is a collection of fuzzy rule extracted by the database, which is expressed in the form "*IF antecedent THEN consequent*".

The antecedent of a fuzzy rule is the "input" while the consequents are the "output". The inference is the process of reasoning using fuzzy inputs and fuzzy rules that have been determined (knowledge based), resulting in a fuzzy output. The method used in the process of inference can use a number of methods, including Mamdani. The Mamdani method represented in a rule as follows:

IF x_1 is A_1 AND x_2 is A_2 … … AND x_n is A_n THEN y is B

where in variables A1, A2….,An, and B is the set of fuzzy values (fuzzy sets), and " x1 is A1" declared a value of variable x1 is a member of the fuzzy set A1. Mamdani method has several advantages, namely intuitive, widely accepted and highly suitable given human input.

Defuzzification is the process of changing into a crisp output value of fuzzy membership function on the terms specified. Defuzzification method used is the centroid, with crisp value calculation as follows:

$$y^* = \frac{\sum y \mu_R(y)}{\sum \mu_R(y)}$$

where, y is the value of crisp and μR (y) is the degree of membership of y.

2.3 Method

This research was conducted through several stages. First do attribute selection, by considering attributes that exist in the dataset and standard of FRS. Both modeling each attribute into a linguistic variable and membership functions. Third rule-based fuzzy modeling by reference to the standard FRS (Wilson et al., 1998). The fourth stage, with modeling for the screening process with reference to the rule-based fuzzy generated. Early screening system model in predicting coronary heart disease events, in general, can be shown in Fig. 1. The model also can provide an overview of the stages used in the developer and the process of using the system for screening. Processing of the data at the time is screening divided into two, namely data categories and continuous. If the category data can be input directly fuzzy inference engine, whereas continuous data fuzzification process needs to be done first. Attribute with the type of continuous data in this empirically is cholesterol, systolic blood pressure, and age, whereas attributes with data type category are gender. Fuzzification process and defuzzification using membership functions for each attribute are

shown in Fig. 2, 3, and 4. After fuzzification stage, the inference process is then performed using Mamdani, output in the form of the fuzzy system by using the membership function as shown in Fig. 5.
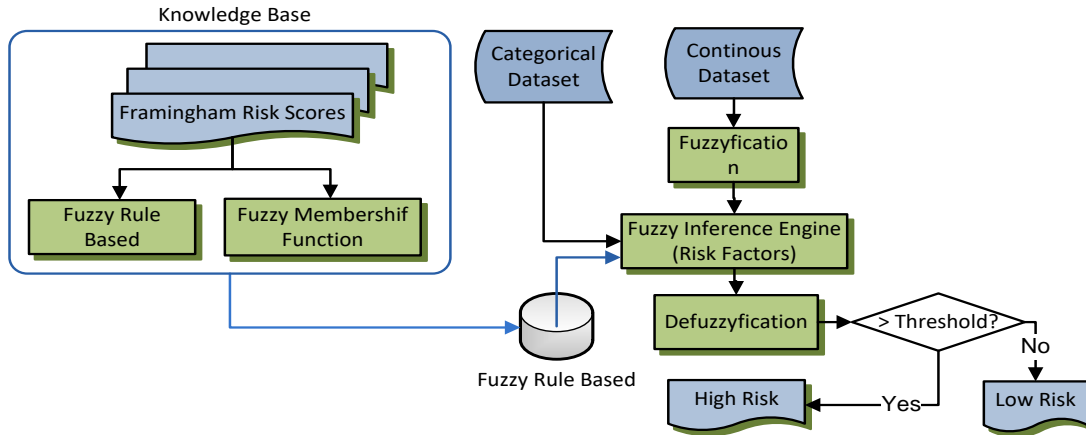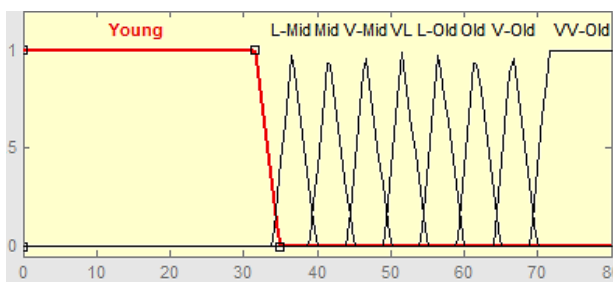

Fig. 1: Model of screening

Fig. 2: Membership function of age
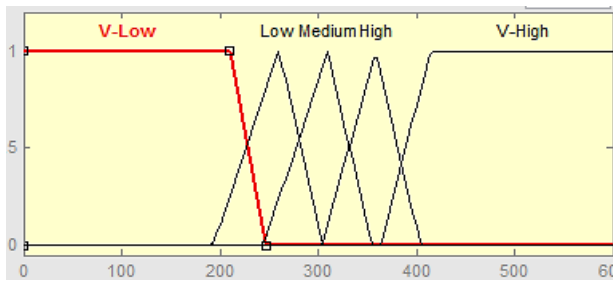

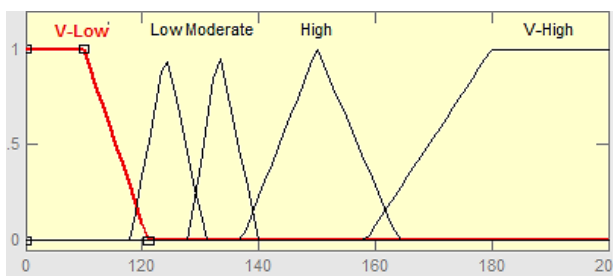**Fig. 3**: Membership function of cholesterol

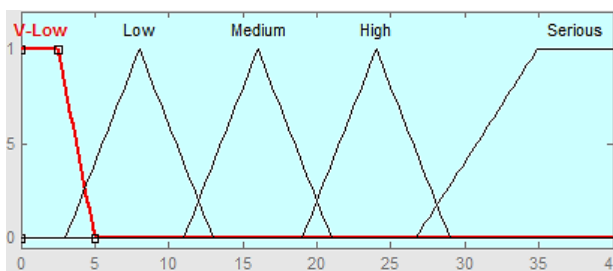**Fig. 4**: Membership function of blood pressure

Fig. 5: Membership function of output fuzzy system

Stages latter is defuzzification process is carried out, which will give an output in the form percentage

occurrence of coronary heart disease, with refers to the standard output FRS. The output system is designed into two namely the high and the low risk, so the required value of the threshold value. If the percentage of the occurrence of coronary heart disease exceeds the threshold, then say high risk, and vice versa low.

2.4. Performance of parameters

Analysis of system performance confusion matrix using a table as shown in Table 1. Parameters in the analysis include sensitivity, specificity, accuracy, an area under the curve, positive prediction value (PPV) and negative prediction value (NPV).

Table 1: Confusion matrix

Actual Class	Predictive Class	
	Positive	Negative
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative)

Sensitivity = TP/ (TP+FN)
Specificity = TN/(TN+FP)
True Positive Rate(TPR)=Sensitivity
NPV = TN/(TN+FN)
Accuracy = (TP+TN)/(TP+TN+FP+FN)
AUC = (1+TPR-FPR)/2(Ramentol et al., 2012)
False Positive Rate(FPR)=FP/(FP+TN)
PPV = TP/(TP+FP)

3. Result and discussion

Model system initial screening on the incidence of coronary heart disease is developed using fuzzy approach. The results of modeling with fuzzy rule-based, produced a number of rule 38. From this number of modeling 4 attributes are used. Here are shown some examples of rule-based fuzzy generated in Table 2. The number of rules will be a knowledge-based system of early screening for coronary heart disease events. The process of screening done by inference using the knowledge based Mamdani.
Output screening system refers to FRS that is prediction percentage incidence of coronary heart disease. The next step conversion, into two output

states of high and low risk. The level of risk is determined by using a threshold value which is the percentage value incidence of coronary heart disease. Results of the system performance by using a variation of the threshold level, both the performance parameters of accuracy, sensitivity, specificity, AUC, PPV, NPV and accuracy, as shown in Fig. 6. Referring to Fig. 6, the threshold for determining that patients in the high category, the threshold on the value 7,00%. The changes in the value of sensitivity and specificity started to descend and ascend relatively significant.

Table 2: The example of rules used in this study

R_n	Fuzzy rule
R_1	IF (Age is Young) and (Gender is Male) THEN (Prediction is Very_Low)
R_2	IF (Age is Less_Mid) and (Gender is Male) THEN (Prediction is Very_Low)
R_3	IF (Blood_Pressure is Very_High) and (Gender is Male) THEN (Prediction is High)
R_4	IF (Cholesterol is Very_High) and (Gender is Female) THEN (Prediction is High)
R_5	IF (Cholesterol is High) and (Gender is Male) THEN (Prediction is High)
R_6	IF (Blood_Pressure is High) and (Gender is Female) THEN (Prediction is Medium)
R_7	If (Cholesterol is Medium) and (Gender is Female) then (Prediction is Low)

If using a threshold value at 12,00%, then on the threshold value of high specificity. The high specificity, have an understanding that, when the examination was declared healthy is more dominant than the pain. The screening will be a better-dictated pain, for a further follow-up test, of the otherwise healthy but apparently sick. The resulting system performance by using a threshold value of 7,00%, is shown in Fig. 7. The performance parameters of the system used are the sensitivity, specificity, PPV, NPV, accuracy, and AUC. Tests carried out using datasets cleveland, hongarian, switzerland and long beach.
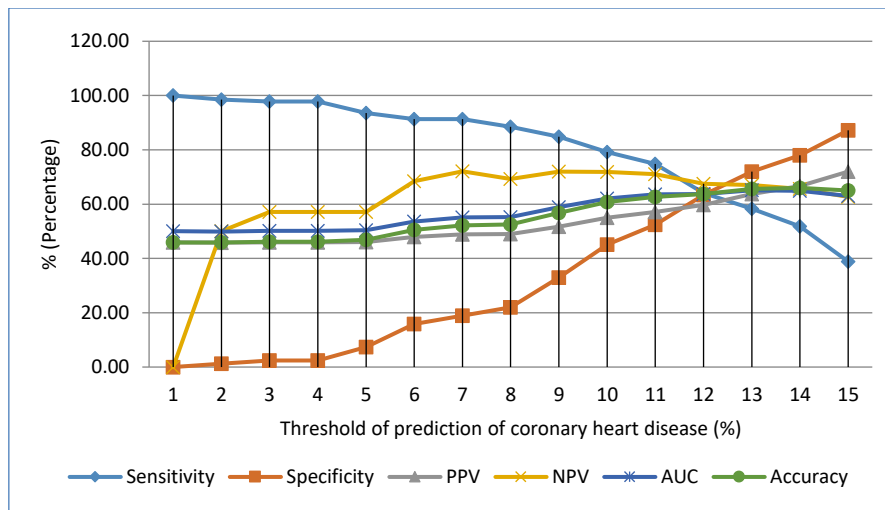

Fig. 6: Threshold of prediction of coronary heart disease

Early screening system model for the prediction of coronary heart disease events by reference to the risk factors capable of delivering performance with relatively good sensitivity value. Referring to Fig. 7, the value of sensitivity generated in this study for the dataset by cleveland is 91.37%, hongarian 86.79%, switzerland 84.35% and the long beach of 96.64%. The performance showed that when patients were positive for coronary heart disease, screening systems, the strongly positive value of coronary heart disease by 91.37% to cleveland dataset. Unfortunately, the high sensitivity is not balanced with high specificity. Based prediction of coronary heart disease risk factors, the initial screening stage in a process diagnosis. Initial screening is the preferred performance parameters sensitivity performance parameters (Wiharto et al., 2016). These considerations indicate better diagnose positive but then in a follow-up examination negative, rather than screening negative, but if it turns out advanced diagnosis is positive.

Early screening system using a panel of four attributes, namely age, sex, systolic blood pressure and total cholesterol. The fourth attributes, according to research conducted Feshki and Shijani (2016), divided into two groups. The first group namely the age and the type of gender The second group is systolic blood pressure and total cholesterol. Both groups are grouped based on the cost of inspection, the first group at an acceptable cost, while the second group is relatively cheap cost (Feshki and Shijani, 2016). Referring to this, the cost of the initial screening examination for coronary heart disease events are relatively cheap.

The previous studies showed that prior to conversion to a fuzzy system, first made the decision tree process to establish rule based. Rule-based is converted into a fuzzy system. This is different from the proposed screening models, namely, by performing modeling fuzzy rule-based and function-based membership FRS. The concept is also used in research Khatibi and Montazer (2010), only in the study that combines two standard Framingham risk score and PROCAM, and uses five risk factors attribute. In the proposed research, with reference to the research conducted Versteylen et al. (2011),

explains that standard FRS, able to provide better
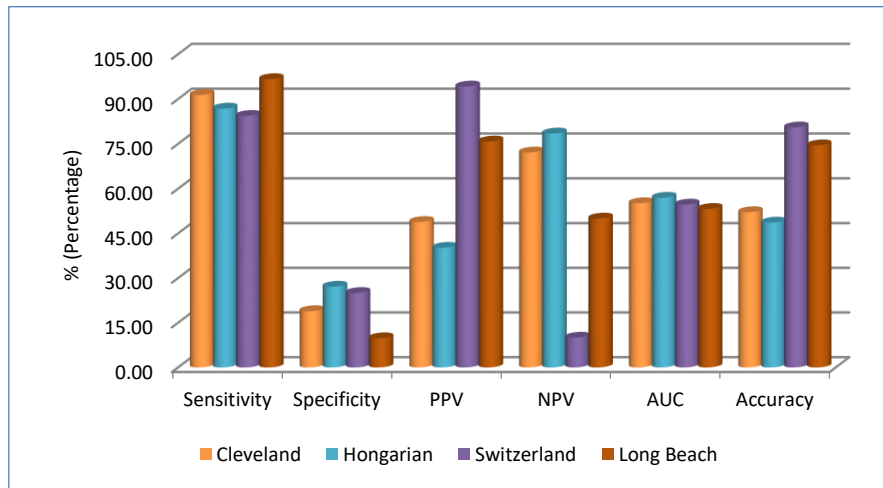
performance in comparison with another standard.

Fig. 7: Performance of system

The performance of the proposed system is relatively the same as the number of previous studies, such as that carried out the study Kim et al. (2014) able to provide accuracy reached 69.22%, while the sensitivity is not shown. The accuracy obtained by using seven attributes of risk factors. Research Kim et al. (2015), when seen from the result of sensitivity reached 93.1% and accuracy was 69.51%. The performance obtained with the 6 attributes risk factors. Also in the study Pal et al. (2011), the resulting system a provide 72.66% accuracy when using algorithm CART, whereas when using ID3 algorithm accuracy of 69.66%. The accuracy value obtained using nine attributes of risk factors. When compared to the proposed system is not much different sensitivity value, even if using a dataset long beach can be reached 96.64% with 74.5% accuracy. Besides the highest accuracy is obtained when testing using switzerland dataset, ie 80.89%.

The next comparison with research conducted Yang et al. (2015). The study uses a neural network, using seven attributes of risk factors. The resulting performance is 85.7% for sensitivity and 52.8% for specificity. When compared with the proposed system, the sensitivity of proposed system is increasing as much as 5.67% when used cleveland dataset. It can be increased also if used long beach dataset as much as 10.9%.

The proposed system, has a forte in terms of the number of attributes that are used, are relatively small, compared with Kim et al. (2015), Kim et al. (2014) and Yang et al. (2015). Attributes used risk factors are age, gender, total cholesterol and systolic blood pressure at rest. The fourth attributes these risk factors necessary inspection fees are relatively inexpensive (Feshki and Shijani, 2016).

4. Conclusion

Initial screening models for the prediction of coronary heart disease events by using fuzzy inference is able to provide a relatively good performance. System performance measured by observing the performance parameters of sensitivity. The consideration refers to the pre-diagnosis stage, which serves as a screening so that the sensitivity parameters as the parameter main performance. The model proposed system besides having good sensitivity performance, also requires only attribute risk factors of age, gender, cholesterol and blood pressure, so it is relatively cheap.

References

Anderson C, Henner T, and Burkey J (2013). Tablet computers in support of rural and frontier clinical practice. International Journal of Medical Informatics, 82(11): 1046–1058.

Assmann G, Cullen P, and Schulte H (2002). Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Münster (PROCAM) study. Circulation, 105(3): 310–315.

Detrano R, Jonasi A, Steinbrunn W, and Pfisterer M (1988). Heart disease dataset. University California Irvine, Irvine, USA. Available online at: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Feshki MG and Shijani OS (2016). Improving the heart disease diagnosis by evolutionary algorithm of PSO and feed forward neural network. In the Artificial Intelligence and Robotics (IRANOPEN), IEEE, Qazvin, Iran: 48–53. https://doi.org/10.1109/RIOS.2016.7529489

Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, Sam J, Haynes RB (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. Jama, 293(10): 1223–1238.

Graham I, Atar D, Borch-Johnsen K, Boysen G, Burell G, Cifkova R, and Herrmann-Lingen C (2007). European guidelines on cardiovascular disease prevention in clinical practice: executive summary: Fourth joint task force of the european society of cardiology and other societies on cardiovascular disease prevention in clinical practice (Constituted by

representatives of nine societies and by invited experts). European Heart Journal, 28(19): 2375–2414.

Khatibi V and Montazer GA (2010). A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment. Expert Systems with Applications, 37(12): 8536–8542.

Kim J, Lee J, and Lee Y (2015). Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree. Healthcare Informatics Research, 21(3): 167–174.

Kim JK, Lee JS, Park DK, Lim YS, Lee YH, and Jung EY (2014). Adaptive mining prediction model for content recommendation to coronary heart disease patients. Cluster Computer, 17(3): 881–891.

Lee HK and Kim HK (2016). Lifestyle and health-related quality of life by type-d personality in the patients with ischemic heart disease. Indian Journal of Science and Technology, 9(13): 1–10.

Pal D, Chakraborty C, and Mandana KM (2011). Data mining approach for coronary artery disease screening. In the International Conference on Image Information Processing, IEEE, Himachal Pradesh, India: 1–6. https://doi.org/10.1109/ICIIP.2011.6108972

Popovic B, Carillo S, Agrinier N, Christophe C, Selton-Suty C, Juillière Y, and Aliot E (2013). Ischemic stroke associated with left cardiac catheterization: The importance of modifiable and non-modifiable risk factors. American Heart Journal, 165(3): 421–426.

Ramentol E, Cabllero Y, Bello R, and Herrera F (2012). SMOTE-RSB: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowledge Infromation System, 33(2): 245–265.

Selvarajah S, Kaur G, Haniff J, Cheong KC, Hiong TG, van der Graaf Y, and Bots ML (2014). Comparison of the framingham risk score, SCORE and WHO/ISH cardiovascular risk prediction models in an Asian population. International Journal of Cardiology, 176(1): 211–218.

Versteylen MO, Joosen IA, Shaw LJ, Narula J, and Hofstra L (2011). Comparison of Framingham, PROCAM, SCORE, and Diamond Forrester to predict coronary atherosclerosis and cardiovascular events. Journal of Nuclear Cardiology, 18(5): 904–911.

Wiharto W, Kusnanto H, and Herianto H (2016). Tiered model based on fuzzy inference system for the diagnosis of coronary heart disease. Far East Journal of Electronics and Communications, 16(4): 985–1000.

Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, and Kannel WB (1998). Prediction of coronary heart disease using risk factor categories. Circulation, 97(18): 1837–1847.

Yang J, Lee Y, and Kang UG (2015). Comparison of prediction models for coronary heart diseases in depression patients. International Journal of Multimedia and Ubiquitous Engineering, 10(3): 257–268.