Sussex Research Online

hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm

Article (Accepted Version)

Tayefi, Maryam, Tajfard, Mohammad, Saffar, Sara, Hanachi, Parichehr, Amirabadizadeh, Ali Reza, Esmaeily, Habibollah, Taghipour, Ali, Ferns, Gordon A, Moohebati, Mohsen and Ghayour-Mobarhan, Majid (2017) hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm. Computer Methods and Programs in Biomedicine, 141. pp. 105-109. ISSN 0169-2607

This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/67239/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

http://sro.sussex.ac.uk

Accepted Manuscript

Computer Methods and Programs in Biomedicine
An international journal devoted to the development, implementation and exchange of computing methodology and software systems in biomedical research and medical practice

Hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm

Maryam Tayefi , Mohammad Tajfard , Sara Saffar ,
Parichehr Hanachi , Ali Reza Amirabadizadeh ,
Habibollah Esmaeily , Ali Taghipour , Gordon A. Ferns ,
Mohsen Moohebati , Majid Ghayour-Mobarhan

PII: S0169-2607(16)31190-7
DOI: 10.1016/j.cmpb.2017.02.001
Reference: COMM 4345

To appear in: *Computer Methods and Programs in Biomedicine*

Received date: 30 October 2016
Revised date: 25 January 2017
Accepted date: 2 February 2017

Please cite this article as: Maryam Tayefi , Mohammad Tajfard , Sara Saffar , Parichehr Hanachi , Ali Reza Amirabadizadeh , Habibollah Esmaeily , Ali Taghipour , Gordon A. Ferns , Mohsen Moohebati , Majid Ghayour-Mobarhan , Hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm , *Computer Methods and Programs in Biomedicine* (2017), doi: 10.1016/j.cmpb.2017.02.001

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Highlights

- Associated risk factors of CAD using decision tree,
- Sensitivity, specificity, accuracy of 96%, 87%, 94% and respectively.
- Serum hs-CRP levels as most important variable associated with CAD,
- Model is accurate, specific and sensitive for investigating risk factors of CAD.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm

Maryam Tayefi[1,4*], Mohammad Tajfard[2*] , Sara Saffar[5], Parichehr Hanachi[7], Ali Reza Amirabadizadeh[6], Habibollah Esmaeily[3], Ali Taghipour[3], Gordon A. Ferns[9], Mohsen Moohebati[8#] ,Majid Ghayour-Mobarhan[1,4,8#]

Affiliations:

1) *Metabolic Syndrome Research Center, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.*
2) *Department of Health Education and Health Promotion, School of Health, Management and Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Iran.*
3) *Department of Biostatistics and Epidemiology, School of Health, Management and Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Iran.*
4) *Department of New Sciences and Technologies, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.*
5) *Neurogenic Inflammation Research Center, Department of New Sciences and Technologies, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.*
6) *Medical Toxicology and Drug Abuse Research Center (MTDRC), Birjand University of Medical Sciences, Birjand, Iran.*
7) *Department of Biology, Biochemistry Unit, Alzahra University, Tehran, Iran.*
8) *Cardiovascular Research Center, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.*
9) *Brighton & Sussex Medical School, Division of Medical Education, Falmer, Brighton, Sussex BN1 9PH, UK*

*** Equally contributed Athours**

# Corresponding Authors:
Majid Ghayour-Mobarhan MD, PhD, Metabolic Syndrome Research Center, School of Medicine, Mashhad University of Medical Sciences, 99199-91766, Mashhad, Iran; Tel:+985138002288, Fax: +985138002287; Email: ghayourm@mums.ac.ir
Mohsen Moohebati MD, Cardiovascular Research Center, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran. Tel: Tel:+985138002288, Fax: +985138002287; Email: Mouhebatim@mums.ac.ir

Running title: Hs-CRP as a risk factor of CHD by decision tree

Conflict of interest: The authors have no conflict of interest to disclose

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Abstract

Background and aims: Coronary heart disease (CHD) is an important public health problem globally. Algorithms incorporating the assessment of clinical biomarkers together with several established traditional risk factors can help clinicians to predict CHD and support clinical decision making with respect to interventions. Decision tree (DT) is a data mining model for extracting hidden knowledge from large databases. We aimed to establish a predictive model for coronary heart disease using a decision tree algorithm.

Methods: Here we used a dataset of 2346 individuals including 1159 healthy participants and 1187 participant who had undergone coronary angiography (405 participants with negative angiography and 782 participants with positive angiography). We entered 10 variables of a total 12 variables into the DT algorithm (including age, sex, FBG, TG, hs-CRP, TC, HDL, LDL, SBP and DBP).

Results: Our model could identify the associated risk factors of CHD with sensitivity, specificity, accuracy of 96%, 87%, 94% and respectively. Serum hs-CRP levels was at top of the tree in our model, following by FBG, gender and age.

Conclusion: Our model appears to be an accurate, specific and sensitive model for identifying the presence of CHD, but will require validation in prospective studies.

Key word: Data mining, Decision tree, Coronary artery disease, hs-CRP

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Abbreviation list:

CHD: coronary heart disease; CA: coronary angiography; LR: logistic regression; CART: classification and regression tree, MLP: multi-layer perceptron; RBF: radial basis function; SOFM: self-organizing feature maps; DT: decision tree; SMO, Sequential minimal optimization, BMI: body mass index; SBP: systolic blood pressure; DBP: diastolic blood pressure; LDL: low density lipoprotein; HDL: high density lipoprotein; TC: total cholesterol; TG: triglyceride; FBG: fasting blood glucose; hs-CRP: highly sensitive C-reactive protein; ROC: receiver operating characteristics; AUC: area under the curve; LAD: left anterior descending; RCA: right coronary artery; LCX: left circumflex artery; ACCF/AHA: the American College of Cardiology Foundation and the American Heart Association.

Introduction

Coronary heart disease (CHD) is responsible for 13.3% deaths globally, and is also the most common cause of disability (1). The most important traditional risk factors of CHD are age, high body mass index, male gender, type 2 diabetes, hypertension, and dyslipidemia (2). Atherosclerosis is the most prevalent cause of CHD. It is a complex process, that is initiated by changes in endothelium and leads to atherosclerotic plaque formation. Atheroma narrows the coronary artery and reduces the blood supply to the myocardium (3). The gold standard method for determination of coronary arteries stenosis is coronary angiography (CA). It provides a detailed view of coronary anatomy, but is expensive and is associated with a significant morbidity and mortality (2). Hence the use of novel biomarkers with high sensitivity and specificity and their application to new algorithms to predict CHD remains an important approach to risk stratification.

Data mining is a retrospective computational method for extracting knowledge from large databases. Different data mining algorithms were applied recently to define new models for CHD. These include: decision tree, neural network, and association rule mining, (4-7). Decision tree is easy to implement and

ACCEPTED MANUSCRIPT

interpret. It provides a tree-based classification for developing a predictive model according to independent variables (8). Decision tree appears to be one the most accurate algorithms among data mining tools in CHD. Alizadehsani et al. used 4 different algorithms for classification of 303 CHD records with 54 attributes. They found that among the Bagging, SMO, Neural network and Naïve Bayes methods the highest accuracy was 89% for the Bagging or SMO methods (4). Negahbani et al. used data mining applying Fuzzy c-mean method on a dataset containing 303 patients and 74 features achieved maximum accuracy of 88% (9). Kurt et al. compared logistic regression (LR), classification and regression tree (CART), multi-layer perceptron (MLP), radial basis function (RBF), and self-organizing feature maps (SOFM) for data mining in 1245 CHD records. They showed that classification and regression tree (CART), tree-based rules were the closest to medical reasoning and could be useful in CHD predication (6). In a comparison between Neural network, Naïve Bayes and decision tree algorithm, Soni et al. found that decision tree was the most accurate of the three in CHD prediction (10).

Here we aim to establish a predictive model for coronary artery disease according to demographic, clinical and para-clinical attributes using decision tree algorithm. To the best of our knowledge, this is one of the largest data mining studies using coronary angiographic reports with more than 2300 records. To improving our predictive model we used the novel biomarker of CHD, hs-CRP. This the only predictive model for CHD including hs-CRP. hs-CRP is an acute phase response protein and a biomarker for CHD and related conditions (11, 12). Several studies have shown that serum hs-CRP is an independent risk factor in different aspects of CHD (13); serum hs-CRP has now been recommended as part of the ACCF/AHA Guidelines to be used in primary prevention (14).

Methods

Subjects and data set

This study was based on a case-control study design of patients referred to Ghaem Hospital, Mashhad-Iran for coronary angiography, between September 2011 and May 2013.

ACCEPTED MANUSCRIPT

A random sampling method was conducted in cardiology clinic of Ghaem Hospital for the proposed study as a case control design. The clinic has seven cardiologists, but only one of them recruited patients to the current study. Then individuals were only recruited to the study on three days per week. On average, 5 patients met inclusion and exclusion criteria daily (15 individuals per week). Finally, after 20 months of data collection, a total of 1187 patients were recruited. Following angiography, these individuals were divided into two groups: those with significant angiographically defined CHD [Angiography (+)] (the case group) who had ≥ 50% occlusion in at least one coronary artery and those with a normal angiogram (< 50% obstruction in coronary arteries) [Angiography (–)].

One thousand one hundred fifty-nine healthy controls were selected among people who attended clinics for routine medical assessment or pre-employment medical examinations. The healthy subjects had no signs or symptoms of CHD (such as cardiac chest pain, ECG changes, unstable angina, exertional angina, and no positive finding in cardiologist examination) and also did not have any of the traditional risk factors of CHD such as diabetes, metabolic syndrome, hypertension and dyslipidemia.

The inclusion criteria of the healthy group were age >18 years, who able to understand the study procedures and who provided written consent to participate in the study, who were in good health on the basis of examination, without symptoms of heart disease, who were not pregnant or breast-feeding, and without a history of hospitalization for any illness during the 5 years ago.

A study questionnaire was used to collect demographic information (age, gender, educational level, marital status, occupation), family history of CHD, lifestyle behaviors (smoking habits); physical measurements included BMI and blood pressure measurement, and biochemical measurements were made on blood samples including: hs-CRP, FBG level, and serum lipid profile as previously explained (15).

All the variables which were significantly different between participants with positive angiography, negative angiography and healthy participants were considered as input variables. The input variables were age, sex, SBP, DBP, LDL, HDL, TC, TG, FBG and hs-CRP shown in Table 1. The model evaluated

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

in this study had 10 input variables and one target variable. The target variable consisted of 3 classes as healthy, negative angiography and positive angiography.

Data mining tools are popular and applicable in medical studies to explore unknown patterns or prediction rules. One of the data mining tools is the decision-tree. In this study, a decision tree technique was used to generate the rules. Formal rules were extracted from the continuous dataset of observations by rule induction. CART is known to be a useful approach for pruning leaf nodes, which enhances the generalization capability of learned trees when the generated trees have an excessive number of steps and leaf nodes. CART can also perform analyses and interpretations to generate propositional knowledge, which is a set of rules used to generate ‘If-Then’ rules. Therefore, a CHD prediction model for Koreans was produced by applying the CART rule induction algorithm to KNHANES-VI. The decision-tree procedure is a non-parametric method which creates a tree-based classification model. It classifies cases into groups or predicts values of a target variable based on values of predictor variables. The main purpose of decision tree is to make a predictive model for the target variable according to predictors. The decision tree algorithms includes three types of nodes, the root node, internal node, end node or target

Decision-tree algorithms use splitting criteria to break a node to from a tree. The aim of these criteria is reducing the impurity of a node. Splitting criteria provide a rate for each predictor variable. Variables that have the best rate of splitting criterion, are selected as staying in the model. Information Gain, Gini index and Gain ratio are the most popular and important splitting criteria. Classification and regression tree (CART) is one of the applicable decision tree algorithms. CART is constructed by splitting subsets of data set using all predictor variables. By this procedure, repeatedly all root nodes are created.

The CART algorithm creates a binary division of the tree and pruning a tree on the cost-complexity (16). Also, CART algorithm uses the Gini impurity index for selecting the best variable. The Gini index measures impurity:

$$\mathrm{Gini(D)} = 1 - \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{P_i}^2$$

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

where p_i is the probability that a record in D belongs to class Ci and is estimated by |Ci,D|/|D| (16). The sum is computed over m classes. In the decision-tree, the first variable or root node is the most important factor and other variables can be classified in order of importance (17, 18). It can be stated also that the root node is the variable that can divide the whole population with the highest information gain.

It is common in data mining methods to divide the data set into two parts; a training data set, generally 70% of the subjects, and the testing dataset, 30% of the subjects. The model is constructed on training dataset and it is tested on testing dataset.

Statistical analysis

For statistical analyses, R version 3.0.2 was used. The Kolmogorov-Smirnov test was used to check the normality of variables. Values are reported as mean ±SD for normally distributed variables (or Median and IQR for non-normal distributed variables). Baseline demographics and clinical characteristics were compared among groups using, one-way ANOVA test (normal distributed variables), Kruskal-Wallis test (non-normal distributed variables) and chi-square test for qualitative variables. A P value < 0.05 was regarded as statistically significant.

There were 2346 participants who were considered in the model. As a common rule in decision tree, data were divided into training and testing groups, 70% of total participants (1640 cases) were randomly selected to make training group for constructing the Decision tree. The remaining 30% (706 cases) were considered as testing group to evaluate the performance of decision tree. In this study totally there were 1159 healthy, 405 participants with negative angiography and 782 participants with positive angiography.

A confusion matrix was used to evaluate the performance of the decision-tree for classification of participants. The accuracy, sensitivity, specificity and the receiver operating characteristics (ROC) curve were measured for comparison.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Results

The characteristics of the total 2346 subjects divided into the three groups (healthy, negative angiography and positive angiography) are shown in Table 1. Data were analysed using chi-square, one-way Anova and Kruskal-Wallis tests respectively.

A decision tree was built on the training group (1640 records). The testing group (706 records) was used to evaluate the model. The algorithm used the Gini index for selecting the variables, and the final tree was pruned. In this model, of total 10 input variables, hs-CRP, FBG, age, TC, SBP, sex and remained in the model. The final decision tree, with size 25, 14 leaves and 9 layers is shown in **Error! Reference source not found.**. The if-then rules created by tree is shown in Table 2. The evaluation of the tree was undertaken using confusion matrix on a testing dataset and shown in **Error! Reference source not found.**. The decision tree had an accuracy of 94%. Of the 375 healthy individuals in testing datasets, 321 were classified correctly using the decision-tree. 131 cases had negative angiography, of whom 53 were classified correctly and of the 240 cases with positive angiography, 193 cases were classified correctly. The specificity and sensitivity of this tree were 87% and 96% respectively. A ROC curve was obtained by applying decision-tree on testing the dataset which (**Error! Reference source not found.**).

The tree showed that in a subgroup with hs-CRP<3.2, FBG<100 and age>=66, the probability of having a positive angiogramme was 67%. In the subgroup with hs-CRP<3.2, FBG<100, age<66 and SBP<140, the probability of being healthy was 93%. In a same situation with SBP>=140, if age wasl <58, 37% was the probability of having negative angiography or positive angiography, while if age was58-66, there was an 86% probability of being healthy. In the male subgroup with hs-CRP<3.2, FBG>=100 and TC<152, 77% of individuals had positive angiography. In a female subgroup if hs-CRP<3.2, FBG>=100 and TC<152, HDL has an important role. If HDL>=28, 61% of individuals had negative angiography and if HDL<28, the probability of having positive angiography was 77%. In the female subgroup that hs-CRP is between 3.2 to 5.3, FBG <134, the individuals had negative angiography were 87%, while in a male subgroup, for FBG between 112 to 134, there was an 81% probability of a negative angiography. In a

ACCEPTED MANUSCRIPT

same situation, if FBG<112, 69% had a positive angiography. In a subgroup with hs-CRP>=5.3 and age <56, the probability of having negative angiography was 58%, while if age>=56 y, 93% of individuals had positive angiography (Table 2.)

Error! Reference source not found. shows the sensitivity, specificity, accuracy values for the tree as 96%, 87%, 94% respectively. The under the ROC curve (AUC) was 0.95 (Fig2).

Discussion

We conducted a retrospective study to create a tree to identify the associated risk factors for coronary artery disease. Amongst 11 traditional related factors of CHD including age, gender, PAL, BMI, SBP, DBP, FBG, TC, TG, LDL, HDL and hs-CRP, we entered 10 significant attributes of 2346 records in decision tree. This is the only study of data mining algorithms which consider serum hs-CRP as an attribute for risk factors of angiographic results. Interestingly hs-CRP was at the apex of the tree which divided the population with the highest information. We found that high levels of hs-CRP together with FBG, gender and age were more than 95% of the determinants for the presence of CHD.

Decision tree (DT) is a data mining algorithm for predicting diseases as well as coronary artery disease using different risk factors. Ture et al. in a data mining assay applying DT algorithm on 1381 patients, considering 8 major traditional risk factors of CHD. Among traditional risk factors of CHD sex, age, type II diabetes mellitus, smoking status and family history of CHD entered the algorithm but BMI, hypercholesterolemia and systemic hypertension did not consider because they were no significant different between healthy and CHD patients. They found sex and age at the top of the tree. Sensitivity and specificity was 95.9% and 31.9%, respectively (19). Likewise, our analysis revealed that these unmodifiable risk factors are important in CHD, although hs-CRP and FBG appeared to be more important in our group. Furthermore, the sensitivity and specificity for the tree of CHD, were better at respectively, 96% and 87%.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Data mining has been conducted on 2949 records with 40 features including 10 non-genetic and 30 candidate gene to predict CHD. Chen et al. used 6 different algorithms including DT. They reported sensitivity, specificity and accuracy rate of 88.4% including all variables, comparable to other used algorithms (20). As we have shown all three sensitivity, specificity and accuracy rate were higher in our decision tree. In a data mining survey of 573 records and 15 attributes including age, sex, chest pain type, resting blood pressure, resting electrographic results, serum cholesterol, FBG, results of exercise test, number of major vessels colored by fluoroscopy, defect type, smoking and obesity. They applied three data mining classification techniques namely Naive Bayes, DT and Neural Network. The accuracy of these models was respectively, 90.74%, 99.62% and 100% (21). Karaolis et al. carried out a data mining analysis using decision tree algorithm in 528 cases including myocardial infarction (MI), percutaneous coronary intervention (PCI) and coronary artery bypass graft (CABG). They entered 15 attributes to their model including demographic and biochemical characteristics and information about past medical history. They reported the highest accuracy for MI, PCI and CABG events as 66%, 75% and 75%, respectively. Age, smoking and positive family history and history of HTN and diabetes were the most important risk factors for these three events (22).

The association of C-reactive protein has been extensively studied in past and is recommended as an important biomarker in primary prevention and screening guidelines for coronary disease and are used in tailoring the treatment/therapy (23-29). Hs-CRP has also been demonstrated to be an important independent predictor for MI. Furthermore, previous cohort studies suggested that hs-CRP> 3 mg/dl indicating high risk (28). Interestingly, our model confirmed almost the same cut-off point for hs-CRP in CHD risk assessment. Moreover, our algorithm define FBG>100 mg/dl as a risk factor of CHD which is previously defined in cohort studies. To confirm the key role of hs-CRP in our model we ran the algorithm without hs-CRP which gave a sensitivity, specificity, accuracy as 83.7%, 88.7% 80.9% respectively. In the model without hs- CRP, FBG with cut-off point of 100 mg/dl was at the apex of the tree followed by age, diastolic blood pressure, total cholesterol and LDL.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Conclusion

Using serum hs-CRP and other traditional CHD risk factors through decision tree algorithm we have obtained an accuracy rate of 94%. To the best of our knowledge this is one of the most accurate tree-based model based on decision tree algorithm and can be used in clinic to differentiate healthy and CHD patients. We indicated that hs-CRP as a new biomarker is strongly associated with CHD even more than traditional biomarkers such as FBG and LDL. Further studies should be conducted to investigate novel biomarkers of CHD such as hs-CRP. Future novel biomarkers may improve the models of CHD risk assessments.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

References

1. Mann DL, Zipes DP, Libby P, Bonow RO, Braunwald E. Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine: Elsevier/Saunders; 2015.

2. Kasper D, Fauci A, Hauser S, Longo D, Jameson J, Loscalzo J. Harrison's Principles of Internal Medicine 19/E (Vol.1 & Vol.2): McGraw-Hill Education; 2015.

3. Libby P, Ridker PM, Hansson GK. Inflammation in atherosclerosis: from pathophysiology to practice. Journal of the American College of Cardiology. 2009;54(23):2129-38.

4. Alizadehsani R, Habibi J, Hosseini MJ, Mashayekhi H, Boghrati R, Ghandeharioun A, et al. A data mining approach for diagnosis of coronary artery disease. Computer methods and programs in biomedicine. 2013;111(1):52-61.

5. Nahar J, Imam T, Tickle KS, Chen Y-PP. Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications. 2013;40(4):1086-93.

6. Kurt I, Ture M, Kurum AT. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Systems with Applications. 2008;34(1):366-74.

7. Tsipouras MG, Exarchos TP, Fotiadis DI, Kotsia AP, Vakalis KV, Naka KK, et al. Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. Information Technology in Biomedicine, IEEE Transactions on. 2008;12(4):447-58.

8. Berry MJ, Linoff G. Data mining techniques: for marketing, sales, and customer support: John Wiley & Sons, Inc.; 1997.

9. Negahbani M, Joulazadeh S, Marateb H, Mansourian M. Coronary Artery Disease Diagnosis Using Supervised Fuzzy C-Means with Differential Search Algorithm-based Generalized Minkowski Metrics. Peertechz J Biomed Eng 1 (1): 006. 2015;14(006).

10. Soni J, Ansari U, Sharma D, Soni S. Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications. 2011;17(8):43-8.

11. Pradhan AD, Manson JE, Rifai N, Buring JE, Ridker PM. C-reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. Jama. 2001;286(3):327-34.

12. Dehghan A, Kardys I, de Maat MP, Uitterlinden AG, Sijbrands EJ, Bootsma AH, et al. Genetic variation, C-reactive protein levels, and incidence of diabetes. Diabetes. 2007;56(3):872-8.

13. Koenig W. High-sensitivity C-reactive protein and atherosclerotic disease: from improved risk prediction to risk-guided therapy. International journal of cardiology. 2013;168(6):5126-34.

14. Greenland P, Alpert JS, Beller GA, Benjamin EJ, Budoff MJ, Fayad ZA, et al. 2010 ACCF/AHA Guideline for Assessment of Cardiovascular Risk in Asymptomatic Adults: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines Developed in Collaboration With the American Society of Echocardiography, American Society of Nuclear Cardiology, Society of Atherosclerosis Imaging and Prevention, Society for Cardiovascular Angiography and Interventions, Society of Cardiovascular Computed Tomography, and Society for Cardiovascular Magnetic Resonance. Journal of the American College of Cardiology. 2010;56(25):e50-e103.

15. Mirhafez SR, Pasdar A, Avan A, Esmaily H, Moezzi A, Mohebati M, et al. Cytokine and growth factor profiling in patients with the metabolic syndrome. British Journal of Nutrition. 2015;113(12):1911-9.

16. Han J, Kamber M, Pei J. Data mining: concepts and techniques: concepts and techniques: Elsevier; 2011.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

17. Tayefi M, Esmaeili H, Karimian MS, Zadeh AA, Ebrahimi M, Safarian M, et al. The application of a decision tree to establish the parameters associated with hypertension. Computer Methods and Programs in Biomedicine. 2017;139:83-91.

18. Ramezankhani A, Pournik O, Shahrabi J, Khalili D, Azizi F, Hadaegh F. Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study. Diabetes research and clinical practice. 2014;105(3):391-8.

19. Ture M, Kurt I, Kurum T. Analysis of intervariable relationships between major risk factors in the development of coronary artery disease: a classification tree approach/Koroner arter hastaligi gelisiminde major risk faktorlerinin birbirleri arasindaki iliskilerin incelenmesi: Bir siniflandirma agaci yaklasimi. The Anatolian Journal of Cardiology (Anadolu Kardiyoloji Dergisi). 2007;7(2):140-6.

20. Chen Q, Li G, Leong T-Y, editors. Predicting coronary artery disease with medical profile and gene polymorphisms data. Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems; 2007: IOS Press.

21. Dangare CS, Apte SS. Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications. 2012;47(10):44-8.

22. Karaolis MA, Moutiris JA, Hadjipanayi D, Pattichis CS. Assessment of the risk factors of coronary heart events based on data mining with decision trees. IEEE Transactions on information technology in biomedicine. 2010;14(3):559-66.

23. Ridker PM. High-sensitivity C-reactive protein potential adjunct for global risk assessment in the primary prevention of cardiovascular disease. Circulation. 2001;103(13):1813-8.

24. Zacho J, Tybjærg-Hansen A, Jensen JS, Grande P, Sillesen H, Nordestgaard BG. Genetically elevated C-reactive protein and ischemic vascular disease. New England Journal of Medicine. 2008;359(18):1897-908.

25. Bisoendial RJ, Boekholdt SM, Vergeer M, Stroes ES, Kastelein JJ. C-reactive protein is a mediator of cardiovascular disease. European heart journal. 2010:ehq238.

26. Nissen SE, Tuzcu EM, Schoenhagen P, Crowe T, Sasiela WJ, Tsai J, et al. Statin therapy, LDL cholesterol, C-reactive protein, and coronary artery disease. New England Journal of Medicine. 2005;352(1):29-38.

27. Ridker PM, Danielson E, Fonseca F, Genest J, Gotto Jr AM, Kastelein J, et al. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. New England Journal of Medicine. 2008;359(21):2195.

28. Mirhafez S, Ebrahimi M, Karimian MS, Avan A, Tayefi M, Heidari-Bakavoli A, et al. Serum high-sensitivity C-reactive protein as a biomarker in patients with metabolic syndrome: evidence-based study with 7284 subjects. European Journal of Clinical Nutrition. 2016;70(11):1298-304.

29. Kazemi-Bajestani SMR, Ghayour-Mobarhan M, Ebrahimi M, Ebrahimi M, Moohebati M, Esmaeili H, et al. C-reactive protein associated with coronary artery disease in Iranian patients with angiographically defined coronary artery disease. Clinical laboratory. 2007;53(1/2):49.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table1. Comparison of baseline characteristics between healthy, negative angiography and positive angiography

	Healthy	Angiography-	Angiography+	p-value
Age(year)	52.97±9.35	53.69±11.42	58.89±10.59	<0.001
Sex				
Male	584(50.4%)	130(32%)	480(61.5%)	<0.001
Female	575(49.6%)	276(68%)	301(38.5%)	
PAL	1.43±0.26	1.42±0.22	1.41±0.22	0.24
BMI(kg/m^2)	26.77±4.18	26.88±5.15	27.12±5.20	0.27
SBP(mmHg)	120.61±15.49	129.75±23.73	134.51±25.00	<0.001
DBP(mmHg)	74.88±9.89	79.94±11.97	81.94±11.16	<0.001
LDL(mg/dl)	113.27±30.78	97.70±35.20	98.22±34.55	<0.001
HDL(mg/dl)	44.47±9.54	42.48±11.60	41.13±15.33	<0.001
TC(mg/dl)	182.83±34.33	168.15±43.76	169.04±43.04	<0.001
TG(mg/dl)	114.74±59.46	138.11±71.26	150.0±73.45	<0.001
FBG(mg/dl)	83.95±17.76	115.74±46.77	133.14±62.62	<0.001
hs-CRP(mg/l)	1.40±0.66	5.34±6.61	6.92±8.82	<0.001

Abbreviation: PAL, physical activity level; BMI, body mass index; LDL, low-density lipoprotein; HDL, high-density lipoprotein; TC, total cholesterol; TG, Triglyceride; FBG, Fasting blood glucose; hs-CRP, high-sensitivity C-reactive protein

ACCEPTED MANUSCRIPT

Table 2. The performance of the decision tree to identify associated risk factors of CVD(On 30%, testing dataset).

Variable	Decision tree model
Sensitivity(95%CI)	97.8(94.6-98.2)
Specificity(95%CI)	92.9(88.2-95.8)
Accuracy(95%CI)	95.3(91.3-97.6)
AUC(95%CI)	95.4(92.9-98.1)

Abbreviation: AUC: area under ROC

Table3. Confusion matrix of testing dataset

	Predicted outcome		
Actual outcome	Healthy	Angio-	Angio+
Healthy	328	4	3
Angiography-	11	56	64
Angiography+	15	32	193

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 4. The 15 rules extracted through decision tree.

R1: If Hs-CRP <3.2 and FBG<100 and age>=66, THEN class: person with Anglo$^+$(20/30 or 67%)

R2: If Hs-CRP <3.2, FBG<100, age<66 and SBP<140, THEN class: person with health (671/719 or 93%)

R3: If Hs-CRP <3.2, FBG<100, age<66 , SBP>=140 and age<58, THEN class: person with health (67/78 or 86%)

R4: If Hs-CRP <3.2, FBG<100, age<66 , SBP>=140 and age>=58,THEN class: person with Angio+,- (15/41 or 37%)

R5: If Hs-CRP <3.2, FBG>=100, TC<152 and sex=male, THEN class: person with Anio$^+$(27/35 or 77%)

R6: IF Hs-CRP <3.2, FBG>=100, TC<152, sex=female and HDL>=28, THEN class: person with Anglo- (19/31 or 61%)

R7: If Hs-CRP <3.2, FBG>=100, TC<152, sex=female and HDL<28, THEN class: person with Angio$^+$((10/13 or 77%)

R8: If Hs-CRP <3.2, FBG>=100, TC>=152 ,THEN class: person with health (66/139 or 47%)

R9: If Hs-CRP between3.2 to 5.3 and FBG<134 and sex=female, THEN class: person with Angio- (67/77 or 87%)

R10: IF Hs-CRP between3.2 to 5.3 and FBG<134 and sex=male and FBS<112, THEN class: person with Angio$^+$(25/36 or 69%)

R11: If Hs-CRP between3.2 to 5.3 and FBG<134 and sex=male and FBS>=112, THEN class: person with Angio- (25/31 or 81%)

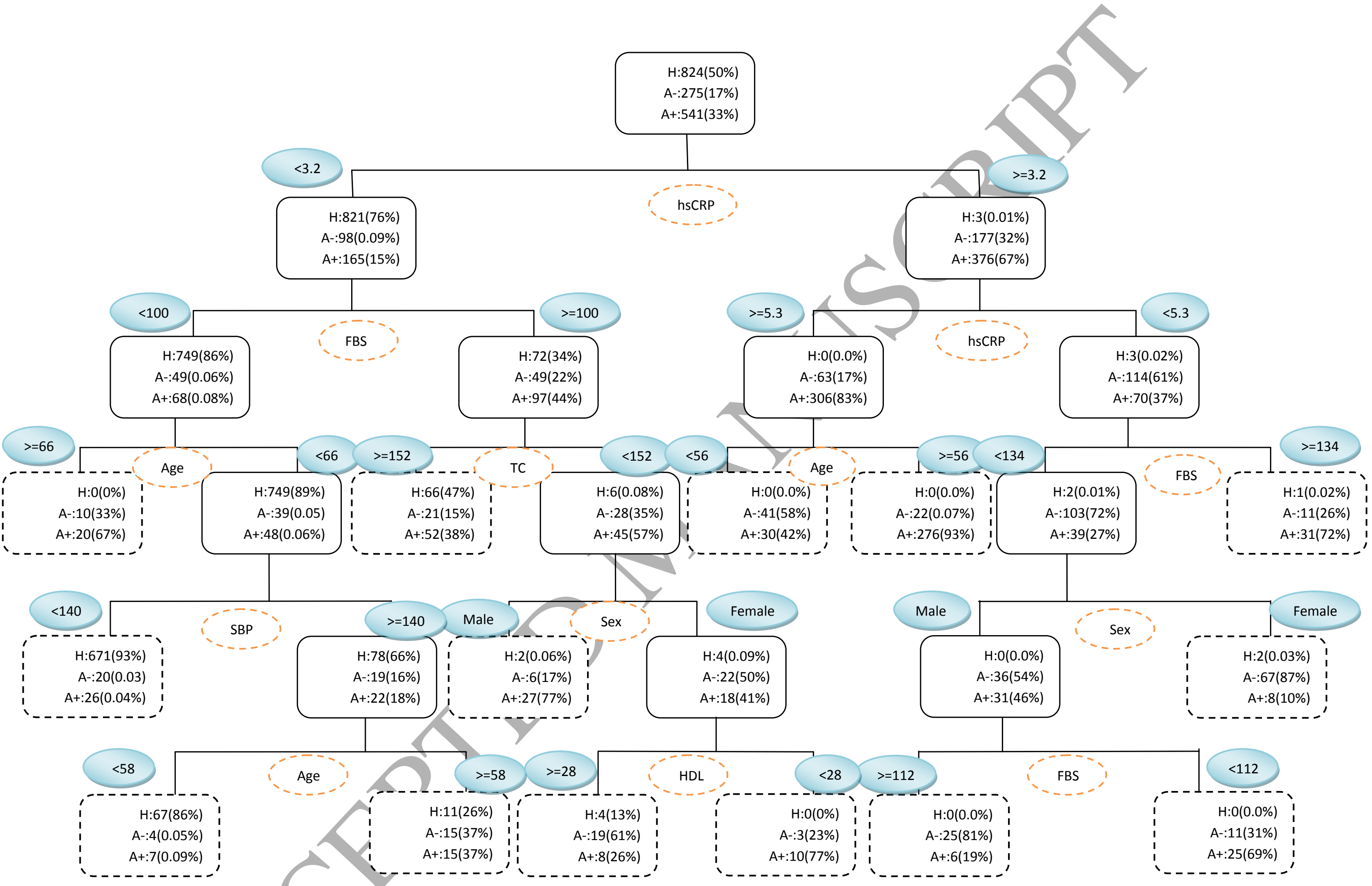
R12: If 3.2=<Hs-CRP<5.3 and age<56, THEN class: person with Anio$^-$ (41/71 or 58%)

R13: If 3.2=<Hs-CRP<5.3 and age>=56, THEN class: person with Anio$^+$ (276/298 or 93%)

R: abbreviation of rule.

ACCEPTED

ACCEPTED MANUSCRIPT



ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Fig 1. Decision tree with training dataset

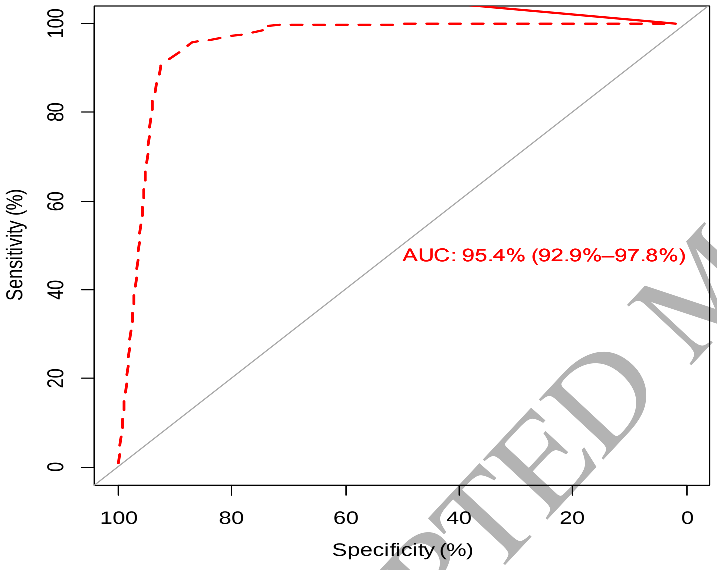


Fig2. Roc curve of the decision tree for testing dataset

ACCEPTED MANUSCRIPT