Int Cardiovasc Res J.2019;**13**(3):85-90.icrj.64781

International Cardiovascular Research Journal

www.ircrj.com

The Discovery of Major Heart Risk Factors among Young Patients with Ischemic Heart Disease Using K-Means Techniques

Samane Sistani [1], Somayeh Norouzi [1], Mohammad Reza Hassibian [1], Mahmoud Tara [1], Hamed Tabesh [1], Sepideh Hasibian [2], Mostafa Dastani [3, *]

[1] Department of Biomedical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, IR Iran
[2] Department of Biology, Mashhad Islamic Azad University, Mashhad, IR Iran
[3] Department of Cardiology, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, IR Iran

ARTICLE INFO

Article Type:
Research Article

Article History:
Received: 4 Dec 2018
Revised: 5 Feb 2019
Accepted: 1 May 2019

Keywords:
Risk Factors
Myocardial Ischemia
Coronary Disease

ABSTRACT

Background: Ischemic Heart Disease (IHD) is the leading cause of mortality in both developed and developing countries. It accounts for more than 15% of the total mortality worldwide. At the global scale, the massive occurrence of this disease can have tremendously negative effects on economy, especially among young people.
Objective: The present study aimed to investigate the relationship between IHD and several important risk factors involved. It also looked into the prevalence of IHD among patients between 20 and 40 years of age.
Methods: The present cross-sectional, retrospective survey was conducted in three referral heart hospitals affiliated to Mashhad University of Medical Sciences, Iran. The required data were extracted from integrated Hospital Information System (HIS) from 2010 to 2012. The data included clinical and demographic information, such as age, gender, marital status, occupation, diabetes, blood pressure, blood cholesterol, cigarette smoking, and family history. In the next phase, clustering technique and k-means algorithm were applied using the WEKA (3-6-9) software.
Results: Totally, 88623 patients suffered from heart diseases between 2010 and 2012. When the specific inclusion and exclusion criteria were considered, the number of records was restricted to 776, which included 548 males. The clustering technique was done in two phases. Firstly, there were four clusters extracted and secondly, cluster analysis was done in terms of age and gender. According to the findings, cigarette smoking in males aged between 20 and 40 years was the main risk factor.
Conclusion: The present research aimed to investigate the risk factors of heart diseases among patients between 20 and 40 years of age. Those below 40 years old were known as the main human resource in the community. The early prevalence of IHD in this population disabled them for the rest of their lives. This disability could also lead to irreparable physiological effects along with financial costs. It could also impose high costs on the society. Recognition of the risk factors of heart diseases at younger ages could contribute to healthcare policies.

1. Background

Cardiovascular Diseases (CVD), defined as the diseases of heart and circulatory system, are the main cause of mortality, morbidity, and hospitalization in males and females all over the world. Statistics have shown that CVD resulted in death of about 17.5 million people throughout the world in 2012, which made up 31% of all global deaths. Among these people, an estimated 7.4 million died due to Coronary Heart Disease (CHD) (1). According to a body of research conducted in Iran, occurrence of CVD is very high and the related mortality rate is on the rise (2, 3). Ischemic Heart Disease (IHD) is mostly revealed with angina and myocardial infarction and is a leading cause of mortality worldwide. It led to 7,249,000 mortalities in 2008, accounting for 12.78% of the total global mortality rate (4, 5). IHD is usually common among the aged. The mortality rate associated with IHD is very low for those below 40 years old. According to the World Health Organization's

*Corresponding author: Mostafa Dastani, Cardiology Department, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran. Cellphone: +98-9151511229,
E-mail: dastanim@mums.ac.ir.

(WHO) statistics (2001) on IHD mortality rate, the total number of male deaths between 5 and 40 years old in the Middle East and North Africa was 50, while it was 7,174 for those above 40 years old (the numbers were respectively 18 and 5,907 for females). In fact, IHD was traditionally considered a disease among males, but it can now be considered a leading cause of death among both males and females (4, 6).

The side effects of IHD among the youth threaten social health and economy. IHD places a major economic burden on economy and the public health system (7, 8). Fortunately, despite the harmful consequences of IHD on social health and economy, it is considered a preventable and non-epidemic disease.

IHD is associated with a wide range of risk factors. Some risk factors including hypertension and hyperlipidemia are controllable, while some others such as age and family history are not (9). In 2002 - 2003, a multi-center study was conducted to measure the prevalence of four conventional risk factors (cigarette smoking, diabetes, hyperlipidemia, and hypertension) among 122,458 patients with CHD. Among these patients, 84.6% of females and 80.6% of males showed to have at least one of the risk factors in question. Indeed, premature CHD was correlated to smoking cigarettes among males and diabetes among females (10).

There have been reports on the impact of geographical differences on the correlation between IHD and the relevant risk factors. As an example, hypertension played a critical role in the growth of IHD in Asia towards developed Western countries (11). Furthermore, hypertension might be more salient among Chinese people than those living in European countries (12, 13). Thus, it is essential to raise public awareness of the risk factors associated with heart diseases prevailing in the society.

IHD accounts for 50% of all mortalities per year in Iran (2, 14). Many risk factors are responsible for the development of IHD in Iran. The related body of academic research indicates that IHD usually appears during middle and old ages in Iran (15). The high prevalence of IHD among the youth could point to a serious health crisis in communities. Awareness of this crisis is the greatest opportunity for medical interventions to minimize the spread of the disease (16). However, no research has been conducted on the prevalence of IHD in the Iranian context at younger ages, especially between 20 and 40.

2. Objectives

This study aims to assess the prevalence of IHD and to identify the relevant risk factors for patients between 20 and 40 years old in Mashhad based on statistical analyses and the clustering technique. The results would express their benefit to public health programming.

3. Patients and Methods

The present applied research was retrospective, descriptive, and cross-sectional. It was conducted in Mashhad, the Spiritual Capital City of Iran and a main medical and therapeutic destination in the East of the country. The data were collected from 88,623 patients with IHD at Ghaem, Imam Reza, and Shahid Hasheminejad hospitals in Mashhad between 2010 and 2012. Purposive sampling was used to collect the required data based on two inclusion criteria; age range of 20 - 40 years and diagnosis of IHD. The IHD diagnosis codes were I20.00 - I25.99 in ICD.

The demographic information included age, gender, marital status, occupation, and the final IHD diagnosis codes extracted from the Hospital Information System (HIS). The clinical data included the main risk factors of heart diseases set by heart specialists and the related body of literature. These included diabetes, hypertension, hyperlipidemia, family history, smoking cigarettes, obesity, and physical activity according to physicians' descriptions included in the patients' paper records. Next, the data were prepared for pre-processing. As there were instances of low-quality data such as recurrent, incorrect, ambiguous, heterogeneous, and unrecorded data, data cleaning was done initially to have precise, valid, and flawless data for analysis.

3.1. Descriptive Statistics

To start data analysis and knowledge extraction, it is essential to know the features of the variables and their distribution. In practice, data analysis is weak and defective without descriptive statistics. At first, the variables that were continuous and quantitative were categorized. These variables included age divided into four groups and the length of stay divided into three groups. Then, the distribution of the variables was determined via the binominal test as well as chi-square test in SPSS 19 and Prism 6 software. The former is run when the target variable is bimodal and the aim is to estimate a particular ratio in population. This test was run to check the distribution of age, marital status, and the risk factors of heart disease. Chi-square test was used to check the distribution of occupation and length of stay. $P < 0.05$ was considered to be the significance level.

3.2. Data mining Model

Different data mining techniques can be divided in two general categories according to their application. The first category includes predictive techniques that aim to predict unknown cases and the second category includes descriptive factors that discover comprehensible patterns of data for human beings including descriptive methods, discovery of association rules, and clustering method adopted in the present study.

Association rules explore and find interrelationships within a dataset. The major algorithms used to discover the association rules are Eclat, FP-Growth, and APRIORI, the last of which was used in the present research. This algorithm is built upon databases that include interactions and aim to find the dependencies between different datasets (18, 19). Patients' data, preprocessed again, were prepared for the association rule algorithms. Then, WEKA (3-6-9) software was used to extract

2,212 rules with a confidence level of > 0.9.

Clustering algorithms aim to discover the natural internal grouping of data, which collect a set of data somehow similar to each other in the same cluster. K-means, two-step, and Kenhonen algorithms are the major clustering algorithms among which, K-means was used in the present research. This technique is used for separate clustering of

data when there is no information available on the quality of the clusters (17, 18).

K-means works by defining an initial set of cluster centers extracted from the data. Then, it allocates each record to the most similar cluster based on the input fields of the record. The WEKA software (3-6-9) was used for cluster extraction. In this algorithm, variables were assumed to have an equal weight and the Euclidean distance criterion was used for clustering.

4. Result

4.1. Statistical Analyses

In the present survey, 1.49% of the patients with heart diseases suffered from IHD. Additionally, the incidence of IHD was higher among males compared to females in patients between 20 and 40 years of age in the population.

Among the patients, 70.6% were male and 20.4% were female, indicating a significant difference between the number of males and females affected by IHD (P = 0.0001). Moreover, the incidence of IHD increased with age (Figure 1). The results also indicated a sharp increase in the number of patients affected by IHD who aged 30 years or above.

The study findings indicated that married individuals were more prone to IHD compared to the single ones, except for the patients between 20 and 25 years of age (P = 0.0001) (Figure 2).

In this study, the prevalence of IHD was quite lower among the employees of public or private organizations in comparison to self-employed individuals (50.9%).

The distribution of the risk factors that could be extracted from the patients' records has been presented in Table 1. Some important risk factors, including physical activity and obesity, and laboratory tests were not available. Thus, other risk factors were used as the research variables. It should be noted that no laboratory findings (hyperlipidemia, diabetes, hypertension) were acquired from the patients' history documented in their medical records. Therefore, the results obtained were simply in the form of a patient's "yes" or "no" answer to the susceptibility of the target risk factor. In this survey, most patients (79.5%) with IHD did not have any hyperlipidemia problems (P = 0.0001). In addition, most of the patients with IHD did not show to have low or high blood pressure (P = 0.0001). Besides, 87.4% of the patients affected by IHD were non-diabetic, while 12.6% were diabetic (P = 0.0001), and the difference was statistically significant. Furthermore, 33.67% of the patients affected by IHD were smokers (mostly between 35 and 40 years of age) (P = 0.0001). Smoking cigarettes was almost exclusive to males (only 2% of the females smoked). Finally, 30% of the patients had a family history of IHD (P = 0.0001) (Table 1).

4.2. Data Mining Association Rules

Extraction of the association rules was done on the records with no missing risk factors via the WEKA (3-6-9) software. Totally, 338 records were analyzed and 2,212 rules with a

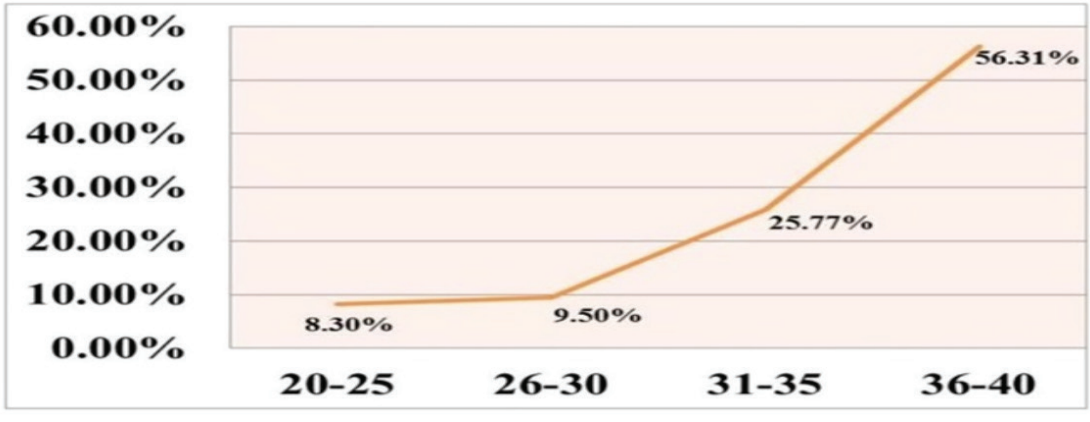

Figure 1. Age Distribution of the Patients Affected by IHD

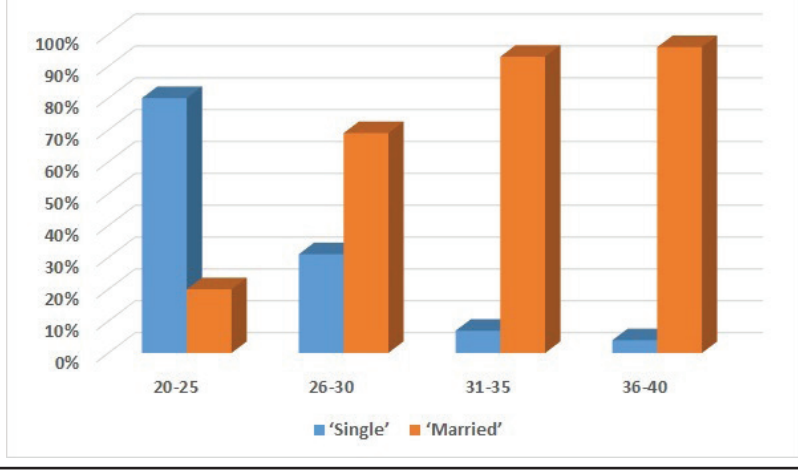

Figure 2. Distribution of Married IHD Patients

Table 1. Distribution of the Major Risk Factors of IHD

Risk Factor	Variables	Qty	Percentage
Diabetes	Positive	74	12.6
	Negative	512	87.4
	Missing	190	
Hypertension	Positive	122	20.5
	Negative	473	79.5
	Missing	181	
Hyperlipidemia	Positive	109	19.7
	Negative	445	80.3
	Missing	222	
Cigarette Smoking	Positive	180	33.7
	Negative	440	66.3
	Missing	192	
Family History	Positive	117	30.2
	Negative	270	69.8
	Missing	389	

Abbreviations: QTY, Quantity

confidence level of 0.9 were extracted. The confidence level was estimated through the following formula:

Conf (A→B) = SUP (A "≡"B)/SUP (A)

All meaningless rules were excluded. Totally, seven rules remained (Table 2). Data clustering

In the first clustering, four clusters were defined (Figure 3).

Cluster 0 included all males with the mean age of 35 years. The risk factors associated with this group were diabetes (11%), hypertension (17%), cigarettes smoking (0%), hyperlipidemia (13%), and family history (31%). Cluster 1 included all single males with the mean age of 27 years. The risk factors associated with this group were diabetes (0%), hypertension (3%), cigarette smoking (34%), hyperlipidemia (13%), and family history (24%). Cluster 2 consisted of all females with the mean age of 33 years. It should be mentioned that 80% of these females were married. The risk factors associated with this group were diabetes (16%), hypertension (24%), cigarette smoking (18%), hyperlipidemia (24%), and family history (22%). Cluster 3 included all married males with the mean age of 36 years. The risk factors associated with this group were diabetes (12%), hypertension (18%), cigarettes smoking (100%), hyperlipidemia (18%), and family history (42%).

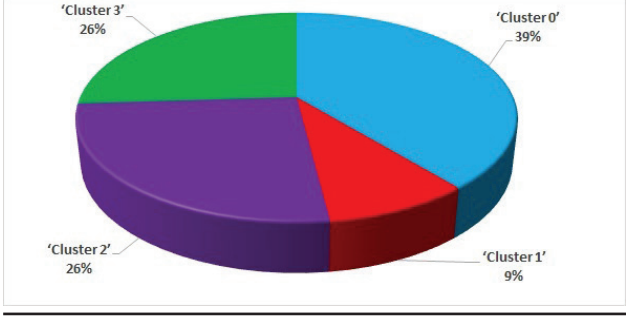

Figure 3. Phase 1 Clustering

The second clustering was done based on the age group and gender (Table 3).

5. Discussion

According to the previous investigations, the incidence of IHD is uncommon among individuals under the age of 40 years. According to a survey, IHD in patients below 40 years of age was found to represent only 3% of all patients afflicted with heart diseases and most of the IHD cases were found among males (19, 20). The results of the present study with regard to the effects of age and gender on the outbreak of IHD were consistent with the findings reported in other studies. The findings indicated that the incidence of IHD increased with age. Moreover, the prevalence of IHD was significantly higher among Iranian males in comparison to females.

The current study findings revealed that self-employed individuals; i.e., 50.9% of those affected by IHD, were more engaged with IHD awareness. No reason could be found for this association unless we accepted that self-employment was a stressful condition. Evidently, this conclusion might not be fully true as the study lacked any

Table 2. The Association Rules for IHD Patients between 20 and 40 Years Old Age

Confidence	Result	Rule
0.93	150 of them did not have diabetes.	162 patients did not have hypertension and hyperlipidemia and did not smoke.
0.96	88 of them did not have diabetes.	92 patients had hyperlipidemia and smoked cigarettes.
0.9	152 of them did not have hyperlipidemia.	168 male patients were married and did not have diabetes and hypertension.
0.94	74 of them did not have diabetes.	126 patients aged 35-40 years and did not have hypertension and hyperlipidemia.
0.91	74 of them did not have diabetes.	81 married patients were self-employed and did not have hyperlipidemia and any trace of heart diseases in their family history.
0.92	81 of them did not have hyperlipidemia.	91 non-diabetic patients had heart diseases in their family history.
0.91	77 of them did not have hyperlipidemia.	85 non-diabetic patients did not have hypertension, but smoked cigarettes.

Table 3. Patient Clustering based on Age and Gender Segregation

Age Range (Years)	Females					Males				
	DM	HTN	SM	LPH	FH	DM	HTN	SM	LPH	FH
20 - 25	10%	20%	30%	20%	10%	0%	9%	17%	13%	13%
26 - 30	18%	9%	9%	27%	36%	14%	10%	38%	7%	34%
31 - 35	14%	24%	24%	14%	19%	11%	21%	27%	18%	44%
36 - 40	18%	29%	16%	29%	24%	11%	17%	43%	16%	33%

Abbreviations: DM, diabetes mellitus; HTN, hypertension; SM, cigarette smoking; LPH, hyperlipidemia; FH, family history

data on hypertension among the population of IHD patients. Furthermore, no reasonable relationship could be found between marriage and IHD, except that it might result from the definite effects of age on IHD and other risk factors related to the population's culture.

It has been proven that smoking cigarettes is one of the preventable leading risk factors of IHD (21, 22). The present study results showed that smoking was a major risk factor among males. However, smoking cigarettes was only notable in females aged 20 to 25 years.

The current study findings indicated that family history was one of the main risk factors among both males and females. Diabetes mellitus is frequently present among young IHD patients. Although the relationship between diabetes and IHD is well-understood, the importance of this risk factor has not been fully characterized in details (23). There is a body of research indicating that diabetes, in the absence of other risk factors, prevails in 15 - 20% of all young IHD patients. The present study results revealed that diabetes mellitus and hypertension were not significantly correlated to IHD in young patients under the age of 40 years. Moreover, diabetes might play a more serious role in females than in males. Similar results were obtained in other studies (24).

5.1. Limitations

The present study had several limitations. Firstly, the risk factors thought to be correlated to the prevalence of IHD did not exist in the patients' electronic medical records. Thus, information regarding the risk factors was extracted from paper records. Secondly, there was no specific information about the risk factors in the patients' records. Thirdly, some risk factors including obesity and physical activities, which were initially explored in the study, were not included in the patients' records.

As for the increasing incidence of IHD in the Iranian society, especially among individuals aged below 40 years, it is essential to assess the main risk factors correlated to IHD prevalence nationally. This will enable the authorities to formulate policies to promote healthy lifestyles and reduce the prevalence of IHD. Moreover, informing and advising smokers, even those with myocardial infarction, to stop smoking will lower the mortality and morbidity rates and reduce the treatment costs for both patients and the society.

5.2. Conclusion

People under 40 years old are considered to be the most effective manpower in any society. Even though IHD mortality is declining, the incidence of IHD can cause disability for even healthy people in the prime of life. Disability will have significant psychological effects on patients and can cause financial stress forcing them to take medicines for the rest of their lives.

5.3. Ethical Approval

The current study was retrospective, descriptive, and cross-sectional. Therefore, written consent was not considered in this study.

Acknowledgements

The results reported in this paper were extracted from a thesis submitted by the first author for an MSc degree in Medical Informatics. The researchers are grateful for Ms. Zahra Eslami for typing and reviewing the article.

Authors' Contribution

SS, MH, SN, MD, and MT contributed to the conception of the work. SS, SN, and SH contributed to data collection. SS, SN, and HT contributed to statistical analysis and data interpretation. SS, MH, and SN contributed to drafting of the work. MH was the study supervisor.

Funding/Support

The study was supported by the Vice-chancellor for Research Affairs of Mashhad University of Medical Sciences (grant No. 930959).

Financial Disclosure

The authors have no financial interests related to the material in the manuscript.

References

1. Noor L, Shah SS, Adnan Y, Sawar S, ud Din S, Amina*, et al.* Pattern of coronary artery disease with no risk factors under age 35 years. *Journal of Ayub Medical College, Abbottabad : JAMC*. 2010;**22**(4):115-9.
2. Hatmi ZN, Tahvildari S, Gafarzadeh Motlag A, Sabouri Kashani A. Prevalence of coronary artery disease risk factors in Iran: a population based survey. *BMC cardiovascular disorders*. 2007;**7**:32.
3. Sarraf-Zadegan N, Sayed-Tabatabaei FA, Bashardoost N, Maleki A, Totonchi M, Habibi HR*, et al.* The prevalence of coronary artery disease in an urban population in Isfahan, Iran. *Acta cardiologica*. 1999;**54**(5):257-63.
4. Finegold JA, Asaria P, Francis DP. Mortality from ischaemic heart disease by country, region, and age: statistics from World Health Organisation and United Nations. *International journal of cardiology*. 2013;**168**(2):934-45.
5. Scarborough Pa, Bhatnagar P, Wickramasinghe K, Smolina K, Mitchell C, Rayner M. Coronary heart disease statistics 2010 edition. *British Health Foundation Health Promotion research group, Department of Public Health, University of Oxford*. 2010.
6. Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. Growing epidemic of coronary heart disease in low- and middle-income countries. *Current problems in cardiology*. 2010;**35**(2):72-115.
7. Capewell S, Ford ES, Croft JB, Critchley JA, Greenlund KJ, Labarthe DR. Cardiovascular risk factor trends and potential for reducing coronary heart disease mortality in the United States of America. *Bulletin of the World Health Organization*. 2010;**88**(2):120-30.

8. Roger VL, Go AS, Lloyd-Jones DM, Benjamin EJ, Berry JD, Borden WB*, et al*. Heart disease and stroke statistics--2012 update: a report from the American Heart Association. *Circulation*. 2012;**125**(1):e2-e220.
9. Hajar R. Risk Factors for Coronary Artery Disease: Historical Perspectives. *Heart views : the official journal of the Gulf Heart Association*. 2017;**18**(3):109-14.
10. Khot UN, Khot MB, Bajzer CT, Sapp SK, Ohman EM, Brener SJ*, et al*. Prevalence of conventional risk factors in patients with coronary heart disease. *Jama*. 2003;**290**(7):898-904.
11. Sasayama S. Heart disease in Asia. *Circulation*. 2008;**118**(25):2669-71.
12. Pais P, Pogue J, Gerstein H, Zachariah E, Savitha D, Jayprakash S*, et al*. Risk factors for acute myocardial infarction in Indians: a case-control study. *Lancet*. 1996;**348**(9024):358-63.
13. Yusuf S, Reddy S, Ounpuu S, Anand S. Global burden of cardiovascular diseases: Part II: variations in cardiovascular disease by specific ethnic groups and geographic regions and prevention strategies. *Circulation*. 2001;**104**(23):2855-64.
14. Azizi F, Rahmani M, Emami H, Mirmiran P, Hajipour R, Madjid M*, et al*. Cardiovascular risk factors in an Iranian urban population: Tehran lipid and glucose study (phase 1). *Sozial- und Praventivmedizin*. 2002;**47**(6):408-26.
15. Sayadi M, Zibaeenezhad M, Taghi Ayatollahi SM. Simple Prediction of Type 2 Diabetes Mellitus via Decision Tree Modeling. *International Cardiovascular Research Journal*. 2017;**11**(2).
16. Nadeem M, Ahmed SS, Mansoor S, Farooq S. Risk factors for coronary heart disease in patients below 45 years of age. *Pakistan journal of medical sciences*. 2013;**29**(1):91-6.
17. Han J, Pei J, Kamber M. *Data mining: concepts and techniques*. Elsevier; 2011.
18. Krishnaiah V, Narsimha G, Chandra NS. Heart disease prediction system using data mining techniques and intelligent fuzzy approach: A review. *International Journal of Computer Applications*. 2016;**136**(2):43-51.
19. Rubin JB, Borden WB. Coronary heart disease in young adults. *Current atherosclerosis reports*. 2012;**14**(2):140-9.
20. Klein LW, Nathan S. Coronary artery disease in young adults. Journal of the American College of Cardiology; 2003.
21. Ambrose JA, Barua RS. The pathophysiology of cigarette smoking and cardiovascular disease: an update. *Journal of the American College of Cardiology*. 2004;**43**(10):1731-7.
22. Kannel W, McGee D, Castelli W. Latest perspectives on cigarette smoking and cardiovascular disease. *Journal of Cardiac Rehabilitation*. 1984;**4**(7):267-77.
23. Krolewski AS, Kosinski EJ, Warram JH, Leland OS, Busick EJ, Asmal AC*, et al*. Magnitude and determinants of coronary artery disease in juvenile-onset, insulin-dependent diabetes mellitus. *The American journal of cardiology*. 1987;**59**(8):750-5.
24. Tomono S, Ohshima S, Murata K. The risk factors for ischemic heart disease in young adults. *Japanese circulation journal*. 1990;**54**(4):436-41.