CRIMSONpublishers
http://www.crimsonpublishers.com

Research Article

ISSN: 2577-2007

The Use of Artificial Neural Network and Logistic Regression to Predict the Influence of Lifestyle on Cardiovascular Risk Factors

Jahandideh S[1]*, Jahandideh M[2], Asefzadeh S[3] and Ziaee A[4]

[1]*Griffith University, Australia*

[2]*Zanjan University, Iran*

[3]*Qazvin University of Medical Sciences, Iran*

[4]*Qazvin University of Medical Sciences, Iran*

***Corresponding author:** Jahandideh S, School of Medicine & Menzies Health Institute Queensland, Gold Coast Campus, Griffith University, Queensland, Australia

Submission: September 19, 2017; **Published:** November 15, 2017

Abstract

Objective: The first cause of death worldwide is cardiovascular disease (CVD). CVD covers a wide array of disorders, including diseases of the cardiac muscle and of the vascular system supplying the heart, brain, and other vital organs. This research aims to the comprehensive impact that a series of lifestyle data from a population has on the main cardiovascular risk factors. Such predictions of the influence of lifestyle on cardiovascular risk factors could be useful.

Material and methods: A cross-sectional study was designed; the subjects who lived in Minodar, Iran were interviewed by trained nurses using a structural questionnaire. Data were processed from a sample of 393 subjects of both sexes aged 26-79 years old. The output data were: high/low cholesterolemia, HDL-C cholesterol, triglyceridemia. The input data were: sex, age, build, weight, marital status, Individual's status in the family, physical activity, hours of sleep per day, smoking, tobacco type, BMI. Two predictor models including artificial neural network and linear regression were applied.

Results: Logistic regression (LR) as a conventional model obtained poor prediction performance measure values. However, LR distinguished that relationships exist between inputs and dichotomous output variables (sex and BMI in TG and sex, weight and tobacco type in HDL-C and sex in total cholesterol as more significant parameters). On the other hand, artificial neural network as a more powerful model showed high response accuracy in predicting CVD risk factors. Such pleasing results could be attributed to the non-linear nature of ANN in problem solving which provides the opportunity to predict independent variables to dependent ones non-linearly.

Conclusion: The results displayed that our ANN-based model approach is very hopeful and may play a useful role in developing a better method for assessing the influence of lifestyle on cardiovascular risk factors.

Keywords: Cardiovascular risk factors; Lifestyle; Neural network; Logistic regression

Introduction

Recent years have witnessed a dramatic decline in cardiac mortality, but ischemic heart disease is still the leading killer in many parts of the world [1]. In Iran, similarly, Coronary Artery Disease (CAD) is the principle culprit for mortality, morbidity, and disability [2].

Direct (hospitalization and treatment) and indirect (absenteeism and unemployment) costs caused by CAD are estimated at 26.77 billion Rials in the Iranian Oil Industry [3]. The most important established risk factors for CAD include: high blood cholesterol (total cholesterol, LDL), hypertension, smoking, diabetes, and poor eating habits [4]. Although these risk factors have been diagnosed as the main causes of CAD, many studies have shown that more than 50 percent of patients with CAD have, in most cases, an absence of all risk factors except high cholesterol [5-8]. Smoking, hypertension, dyslipidemia, and inactivity are the main and controllable factors in CAD. Control and treatment of risk factors can significantly reduce mortality and the costs associated with CAD [9]. Primary and secondary prevention of coronary heart disease in residents of South Asia is a major health priority because the risk factors, including for CAD, in this population are very common [10]. Iran is also one of the countries in southwestern Asia that is not exempted from this matter. The death rates from CAD in Iran show it is necessary to find a solution to reduce the incidence of these illnesses and deaths.

Ischemic Heart Disease (IHD) is the single largest cause of death in developed countries and is one of the main contributors to the disease burden in developing countries. The two leading manifestations of IHD are angina and acute myocardial infarction. In 2001, IHD was responsible for 7.3 million deaths and 58 million Disability-Adjusted Life Years (DALYs) lost world Wide Organization [11].

The risk of developing Cardiovascular Disease (CVD) depends to a large extent on the presence of several risk factors. The major risk factors for CVD include tobacco use, high blood pressure, high blood glucose, lipid abnormalities, obesity, and physical inactivity. The global variations in CVD rates are related to temporal and regional variations in these known risk factors. Discussions of the strength of the associations of the various factors with CVD are widely discussed in the literature. Although some risk factors, such as age, ethnicity, and gender, obviously cannot be modified, most of the risk is attributable to lifestyle and behavioral patterns, which can be changed [12].

Artificial Neural Networks (ANNs) have attracted growing interest in recent years as a supplement or alternative to standard statistical techniques to predict complex phenomena in medicine and biological studies [13]. A neural network is a non-linear statistical data modeling tool that is able to capture and represent complex input/output relationships.

In this study, ANNs as a non-algorithmic model are used in predicting the influence of life style in cardiovascular risk factors. Prediction of risk factors will be helpful in assessing the comprehensive impact that a set of data demonstrating lifestyle from society has on the main cardiovascular risk factors. Most of the previous research has focused on predicting heart disease by considering risk factors [14-16].

Materials and Methods

Data set

This study used a cross-sectional design and a convenience sample of 393 subjects. Subjects' participation in the research was voluntary. The subjects were interviewed randomly by trained nurses and physicians using a structural questionnaire with each interview taking around 30 minutes. The questionnaire contained anthropometric, laboratory and physical activity questions. The subjects were used to compile the dataset. Both sexes were represented, were aged from between 26 to 75 years (average age 41.97 years old), and were all living in Minodar, Iran in 2015.

Model development

Logistic regression and artificial neural network as two algorithmic and non-algorithmic models were used.

Definition of the input parameters

The input data from the network for each subject were as follows: Sex (1: male, 2: female), age (years), height (cm), weight (kg), marital status (1: single, 2: married, 3: divorced, 4: wife dead), Individual's status in the family (1: head of family, 2: wife, 3: children), physical activity (1: high, 2: moderate, 3: low), hours of sleep per day (1: Less than 8 hours, 2: more than 8 hours), smoking (1: yes, 2: no), tobacco kind (Cigarette, pipe, hubble-bubble, none of them), BMI (underweight: <18.5, normal: 18.5-24.9, overweight: 25-29.9, obesity grade1: 30-34.9, obesity grade 2:>35). More details about HDL-Cholesterol, Cholesterol, Triglycerides and values of selected parameters are presented in (Table 1).

Table 1: Characterization of the study population.

	Cholesterol		HDL-C		Triglycerides	
	1	**2**	**1**	**2**	**1**	**2**
sex						
Male (1)	29	142	37	135	75	97
Female (2)	7	198	9	196	13	192
Age (Year)						
26-29	1	13	1	13	0	14
30-39	3	110	9	104	12	101
40-49	20	164	25	160	53	132
50-59	7	31	8	30	14	24
60-69	2	9	1	10	4	7
70-79	1	2	1	2	2	1
Height (Cm)						
140-149	2	11	1	12	2	11
150-159	5	111	13	103	11	105
160-169	13	126	12	127	36	103
170-179	13	73	15	71	32	54
180-189	3	16	4	16	6	14
190-199	0	1	0	1	1	0
Weight (Kg)						
30-39	0	0	0	0	0	0
40-49	1	5	1	5	1	5
50-59	2	33	5	30	2	33
60-69	9	122	19	113	21	111
70-79	11	107	13	105	35	83
80-89	11	57	6	62	24	44
90-99	2	12	1	13	4	10
100-109	0	2	0	2	1	1
110-120	0	1	0	1	0	1
Marital Status						
Single	1	5	0	6	0	6
Married	35	330	46	320	87	276

divorced	0	2	0	2	0	2
Wife dead	0	3	0	3	1	2
smoking						
Yes	5	44	6	43	18	31
No	31	296	40	288	70	258
Hours of sleep per day						
Less than 8 hours	29	251	36	43	69	211
More than 8 hours	5	81	9	288	16	71
Physical activity						
High	15	125	25	115	40	100
Moderate	7	68	8	67	16	59
Low	23	155	21	157	40	138

We used the WHO Global physical Activity Questionnaire (GPAQ) to calculate the physical activity. It comprises of 16 questions to collect information on physical activity participation in three settings as well as sedentary behavior. The domains are activity at work; travel to and from places, and recreational activities. To calculate a categorical indicator, the total time spent in physical activity during a typical week, the number of days as well as the intensity of the physical activity is taken into account. The three levels of physical activity suggested for classifying populations are low, moderate, and high (Organization, 2012). According to the Analysis Guide, physical activities of subjects were calculated and clustered using MATLAB programming language. Each subject was put into one of 3 categories of physical activity. The cut-off values for the data were: triglyceride: 160mg/dl; cholesterol: 220mg/dl; HDL-cholesterol: 45mg/dl for males, 50mg/dl for females, with the values coming from the ranges of normality indicated by the literature [17]. Of the 393 subjects in the group study, 80 subjects were used for the test phase and the remaining 313 for the training phase.

Logistic regression analysis

Regression is the study of dependence. It is used to answer questions such as: do changes in cholesterol depend on age, sex, physical activity? The goal of regression is to summarize observed data as simply, and usefully as possible.

Artificial neural network models

The human brain has been used to design and develop ANNS. Accordingly, they are a cellular information processing system. The neural network consisted of an interconnected set of artificial neurons. The neurons perform collectively and simultaneously as summing and non-linear mapping junctions for all data and inputs. Changes take place in structure on the basis of internal and external information that flows through the network during the learning phase, so ANN is called an adaptive system. To model complex relationships between inputs and outputs or finding patterns in data, modern neural networks are usually applied (Figure 1).

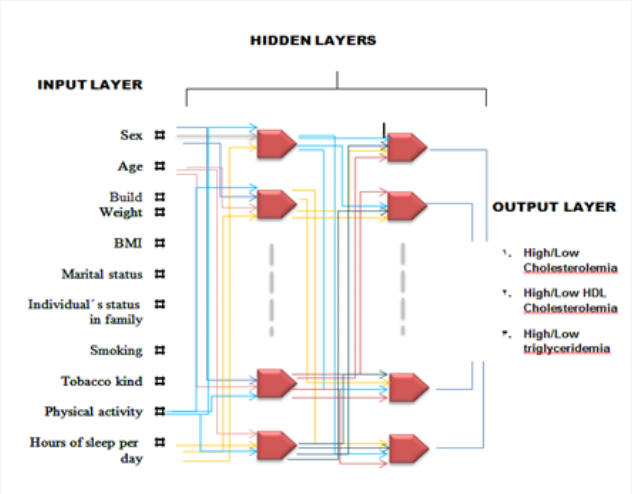

Figure 1: Diagram illustrating the structure of the developed neural network.

In this study, we trained an independent network for each output (Cholesterol, HDL-C and triglycerides). Our networks were trained perfectly over two layers of neurons. The factors of the optimized neural network are shown in Table 2. One or two hidden layers, different learning constants and hidden nodes were tested to train with inputs. An in-house program written in the MATLAB programming language was used to build the neural network.

Table 2: Optimized neural network parameters.

Triglycerides	HDL -C	Cholesterol
Learning rate	0.1	0.1	0.1
Error goal	0	0	0
Number of input nodes	12	12	12
Number of output nodes	1	1	1
Number of hidden layers	2	2	2
Number of layer neurons1	19	10	20
Number of layer neurons2	20	12	10
Training function	Trainrp	Trainrp	Trainrp
Response accuracy	83.75 %	91.25 %	93.75%

Result

Results of ANNs

The ANN-based models were fed with the twelve mentioned factors. The ANN was used as the predictor models on the data base by using the "trainrp" method. For each network, the optimized structure of four layer neural networks included one output neurons, two hidden layers and an input layer. The variables (Cholesterol, HDL-C and triglycerides) were used as the output

in a dichotomous form in the neural network. 313 subjects were entered into the network as a training set and the remainders (80 subjects) were considered as a testing set.

In the first phase, all of the inputs were entered into the network and the output Triglycerides, HDL-C, total cholesterol were entered separately. Different learning constants of .08, 0.1, and 0.2 were tested and the learning constant of 0.1 was selected. The results showed that prediction accuracy in each networks respectively was 85.75%, 91.25% and 93.75%. The results are shown in Table 2.

Results of Logistic regression

The results indicate a strong influence of the sex variable on Cholesterol (P-Value: 0.000), HDL-C (P-Value: 0.000) and triglycerides (P-Value: 0.000) of BMI on triglycerides (P-Value: 0.012), and weight (0.005) and tobacco kind (P-Value: 0.02) on HDL-C. Changes in other factors were not associated with the cardiovascular risk factors. The statistical characteristics of the developed models are shown in Table 3-5.

Table 3: Statistical characteristics of the developed Logistic regression model.

Co-efficients[a]					
Model	**Unstandardized Coefficients**		**Standardized Coefficients**	**test**	**P-value**
	B	**Std. Error**	**Beta**		
(Constant)	1.351	0.085		15.96	0
sex	0.38	0.042	0.44	9.144	0
BMI	-0.065	0.026	-0.122	-2.526	0.012
a. Dependent Variable: Triglycerides					

Table 4: Statistical characteristics of the developed Logistic regression model.

Coefficients[a]					
Model	**Unstandardized Coefficients**		**Standardized Coefficients**	**test**	**P-value**
	B	**Std. Error**	**Beta**		
(Constant)	1.341	0.102		13.187	0
sex	0.195	0.036	0.281	5.439	0
Weight	0.044	0.015	0.141	2.832	0.005
Tobacco kind	0.095	0.041	0.118	2.328	0.02
a. Dependent Variable: HDL-C					

Table 5: Statistical characteristics of the developed Logistic regression model.

Co-efficients[a]					
Model	**Unstandardized Coefficients**		**Standardized Coefficients**		**P-value**
	B	**Std. Error**	**Beta**	**test**	
(Constant)	1.697	0.052		32.689	0
sex	0.122	0.032	0.19	3.817	0
a. Dependent Variable: Total Cholesterol					

Discussion and Conclusion

Several studies have indicated the presence of correlations among variables for example body weight, smoking, age and classic cardiovascular disease risk factors [18-20]; that is, one single factor actually is not the determining factor but rather the comprehensive interaction of all of them together are determining.

In this study, logistic regression as a conventional model obtained poor prediction performance measure values. However, LR distinguished that relationships exist between inputs and dichotomous output variables (sex and BMI in TG and sex, weight and tobacco kind in HDL-C and sex in total cholesterol as more significant parameters). On the other hand, ANNs as a more powerful model showed high response accuracy in predicting. To examine the complex relationship between input variables and output variables, ANNs are broadly used (Nelson & Illingworth) and there are reasons for supporting the ANN-based model. The ANN-based model can handle factors without specifying their complex

non-linear relationships. In positions where determining factors are numerous and/or not completely understood or controlled, this is particularly beneficial. However, when wanting to determine the relative influence of individual factors on performance of prediction, this black box nature of ANN-based models may be disadvantageous. As prediction of the influence of lifestyle on cardiovascular risk factors was the objective of this research, the predictive accuracy is the greatest advantage. In such research, the black box nature is not a weakness. Gueli [21] in their research, which had a similar aim, presented that a neural network offered data for a single individual with a high probability (up to 93.33%) [21].

In conclusion, our results are promising and confirm a beneficial role of neural networks in predicting the influence of lifestyle on cardiovascular risk factors. To have the most advantage of data mining techniques, it is suggested to apply two successful data mining tools, neural networks and genetic algorithms to cover the weakness of ANNs.

References

1. Shaw LJ, Merz CNB, Pepine CJ, Reis SE, Bittner V, et al. (2006) Insights from the NHLBI-Sponsored Women's Ischemia Syndrome Evaluation (WISE) Study: Part I: gender differences in traditional and novel risk factors, symptom evaluation, and gender-optimized diagnostic strategies. J Am Coll Cardiol 47(3 suppl): S4-S20.
2. Hatmi Z, Tahvildari S, Motlag AG, Kashani AS (2007) Prevalence of coronary artery disease risk factors in Iran: a population based survey. BMC Cardiovasc Disord 7(1): 32.
3. Larijani B, Fakhrzadeh H, Mohaghegh M, Pourebrahim R, Akhlaghi M, et al. (2003) Burden of coronary heart disease on the Iranian oil industry [1999-2000]. East Mediterr Health J 9(5-6): 904-10.
4. Grundy SM, Pasternak R, Greenland P, Smith S, Fuster V, et al. (1999) Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations: a statement for healthcare professionals from the American Heart Association and the American College of Cardiology. Circulation 100(13): 1481-92.
5. Mosca L (2002) C-reactive protein-to screen or not to screen. N Engl J Med 347(20): 1615-1617.
6. Lefkowitz RJ, Willerson JT (2001) Prospects for cardiovascular research. JAMA 285(5): 581-587.
7. Magnus P, Beaglehole R (2001) The real contribution of the major risk factors to the coronary epidemics: time to end the only-50% myth. Arch intern med 161(22): 2657-2660.
8. Ridker PM, Rifai N, Rose L, Buring JE, Cook NR, et al. (2002) Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. N Eng J Med 347(20): 1557-1565.
9. Stone NJ (1996) The clinical and economic significance of atherosclerosis. Am J Med 101(4A): 4A6S-9S.
10. Nelson M, Illingworth BA (1991) Practical guide to neural nets. Addison-Wesley Publishing Co., Reading, Boston, Massachusetts, USA.
11. World Health Organization (2002) The world health report 2002: reducing risks, promoting healthy life. World Health Organization, Geneva, Switzerland, pp. 7-14.
12. Jamison DT, Breman JG, Measham AR, Alleyne G, Claeson M, et al. (2006) Disease control priorities in developing countries. Oxford University Press, New York, USA.
13. Patel JL, Goyal RK (2007) Applications of artificial neural networks in medical science. Curr clin pharmacol 2(3): 217-226.
14. Amin SU, Agarwal K, Beg R (2013) Genetic neural network based data mining in prediction of heart disease using risk factors. Paper presented at the Information & Communication Technologies (ICT), IEEE Conference on, Thuckalay, Tamil Nadu, India.
15. Anbarasi M, Anupriya E, Iyengar N (2010) Enhanced prediction of heart disease with feature subset selection using genetic algorithm. International Journal of Engineering Science and Technology 2(10): 5370-5376.
16. Voss R, Cullen P, Schulte H, Assmann G (2002) Prediction of risk of coronary events in middle-aged men in the Prospective Cardiovascular Münster Study (PROCAM) using neural networks. Int J Epidemiol 31(6): 1253-1262.
17. World Health Organization (2012) Global physical activity questionnaire (GPAQ) analysis guide. World Health Organization. Geneva, Switzerland.
18. Otterstad JE (2003) Influence on lifestyle measures and five-year coronary risk by a comprehensive lifestyle intervention programme in patients with coronary heart disease. Eur J Cardiovasc Prev Rehabil 10(6): 429-437.
19. Wallis EJ, Ramsay LE, Haq IU, Ghahramani P, Jackson PR, et al. (2000) Coronary and cardiovascular risk estimation for primary prevention: validation of a new Sheffield table in the 1995 Scottish health survey population. Bmj 320(7236): 671-676.
20. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, et al. (1998) Prediction of coronary heart disease using risk factor categories. Circulation 97(18): 1837-1847.
21. Gueli N, Piccirillo G, Troisi G, Cicconetti P, Meloni F, et al. (2005) The influence of lifestyle on cardiovascular risk factors: analysis using a neural network. Arch Gerontol Geriatr 40(2): 157-172.