Original Article
http://mjiri.iums.ac.ir

Medical Journal of the Islamic Republic of Iran (MJIRI)
Iran University of Medical Sciences

Artificial neural networks versus bivariate logistic regression in prediction diagnosis of patients with hypertension and diabetes

Mehdi Adavi[1], Masoud Salehi[2], Masoud Roudbari*[3]

Received: 12 January 2015 Accepted: 12 May 2015 Published: 3 January 2016

Abstract

Background: Diabetes and hypertension are important non-communicable diseases and their prevalence is important for health authorities. The aim of this study was to determine the predictive precision of the bivariate Logistic Regression (LR) and Artificial Neutral Network (ANN) in concurrent diagnosis of diabetes and hypertension.

Methods: This cross-sectional study was performed with 12000 Iranian people in 2013 using stratified-cluster sampling. The research questionnaire included information on hypertension and diabetes and their risk factors. A perceptron ANN with two hidden layers was applied to data. To build a joint LR model and ANN, SAS 9.2 and Matlab software were used. The AUC was used to find the higher accurate model for predicting diabetes and hypertension.

Results: The variables of gender, type of cooking oil, physical activity, family history, age, passive smokers and obesity entered to the LR model and ANN. The odds ratios of affliction to both diabetes and hypertension is high in females, users of solid oil, with no physical activity, with positive family history, age of equal or higher than 55, passive smokers and those with obesity. The AUC for LR model and ANN were 0.78 (p=0.039) and 0.86 (p=0.046), respectively.

Conclusion: The best model for concurrent affliction to hypertension and diabetes is ANN which has higher accuracy than the bivariate LR model.

Keywords: Artificial neutral Network, Joint logistic regression, Prediction, Diabetes, Hypertension.

Cite this article as: Adavi M, Salehi M, Roudbari M. Artificial neural networks versus bivariate logistic regression in prediction diagnosis of patients with hypertension and diabetes. *Med J Islam Repub Iran* 2016 (3 January). Vol. 30:312.

Introduction

One of the problems facing the medical research is the prediction of concurrent affliction to two diseases according to some common risk factors. These diseases are considered as joint distribution of two or several response variables and the prediction of concurrent affliction to both diseases are made via predictor variables. These models are called bivariate or multivariate models. If the both response variables are quantitative or qualitative, then the standard statistical methods such as bivariate regression or logistic regression (LR) are used for the modeling between response and predictor variables (1-2). The most important application of these statistical methods is finding the relationship between the variables to build a model and predict according to available information (3-6).

There are some assumptions for modeling of relationship between variables in classical methods. Some of these assumptions are normal distribution for response variables, linear relation among response and explanatory variables, and the homogeneity of variances of the error

[1]. MSc in Biostatistics, Department of Biostatistics, School of Public Health, Iran University of Medical Sciences, Tehran, Iran. mehdi.adavi@yahoo.com
[2]. Assistant Professor, Department of Biostatistics, School of Public Health, Iran University of Medical Sciences, Tehran, Iran. salehi74@yahoo.com
[3]. (**Corresponding author**) Professor, Antimicrobial Resistance Research Center, Rasoul-e-Akram Hospital, Department of Biostatistics, School of Public Health, Iran University of Medical Sciences, Tehran, Iran. mroudbari@yahoo.co.uk

terms.

If some of the above mentioned assumptions are not held in actual data, the model cannot be used or has considerable errors. Also, the more sensitivity of the models to outliers and missing data are from the limitations of these models. Therefore, finding alternative methods with less limitation are interested.

The artificial neutral Network (ANN) is one of most suitable method to sort out this problem without any assumptions. Also, ANN has no limitation for the form of relationship between response and predictor variables. In this method ANN finds the form of relationship which is not necessarily linear. Furthermore, the data are implicitly analyzed in ANN, so it has a high probability of finding the correct solution even if a part of the network layers is deleted or works incorrectly (7-8). Also, ANNs have some important disadvantages such as black box nature, complicated computation, and proneness to overfitting (9).

The hypertension and diabetes are non-communicable diseases with the prevalence of 85% in developed countries, 70% in countries with moderate income, and 50% in developing countries (10). Due to the considerable loss of untimely mortality and weakness of these diseases, the governments, especially in developed countries, spend more money to cure these diseases (10). Therefore, it is too important for the countries to prevent non-communicable disease for their high prevalence and cost.

Because of the high prevalence of hypertension and diabetes, especially diabetes type II, almost half of diabetic patients have hypertension. Furthermore, concurrent affliction to both diseases increases the risk of cardiovascular and renal diseases, and cerebro vascular accidents (11-12).

Therefore, the best solution for controlling the diseases' mortality is to predict their concurrent affliction with high precision. The objection of this study is comparing the precision of ANN and joint LR in concurrent diagnosis of patients with hypertension and diabetes.

Methods

Data

In this cross-sectional study which was performed in 2013, the target population is Iranian adults with the age range of 25-64 years. The data are a part of a national study of non-communicable disease risk factors in 2007 by the Ministry of Health and Medical Education (MHME) in Iran. The data are collected using a standard questionnaire via an interview with a sample of 12000 Iranian people.

The sampling method was stratified-cluster sampling where the provinces were the strata and in each province fifty clusters were chosen randomly. The sampling was performed by non-communicable center of MHME.

Statistical analysis

For predicting the concurrent affliction to both hypertension and diabetes, bivariate LR model and ANN methods were used.

To build an ANN, a 8400 sample (70%) for the learning and a 3600 sample (30%) for the test and prediction of the topology of the final network were used. In learning stage, the weight of entrance, middle and output layers were determined randomly. Then, the model processes the data for each unit and sends it to the next unit to compute the values of the response variables. In this stage, the values of calculated response variables were compared with the actual values to find the model error. If the least square error is less than the model error, then the system returns back to change the weights and repeats the same procedures to find new values for the response variables to find the suitable error (5-6).

In this study a 3-layer perceptron ANN with two middle layers (ANN with one middle layer was not converged) and transition function of hyperbolic was built. There were seven entrance layers for seven independent variables. In the next stages with

adding neurons to the middle layer, a perceptron ANN with similar variables were developed. With repeat of these stages, if the mean square error (MSE) was less than 0.01, the process terminates and the best model obtains. For predicting the concurrent affliction to both hypertension and diabetes, a joint LR was used and all independent variables entered to the model. The variables were type of cooking oil, family history, age (less than 55 years, more or equal to 55 years), gender, obesity, passive smokers and physical activity. Since the objective of the study was concurrent affliction to both diseases and the values of independent variables repeated for both hypertension and diabetes as response variables, therefore generalized estimating equation models was used for model building in SAS 9.2 software, and Matlab software was utilized to construct ANN.

Finally, predicting of concurrent affliction to both diseases was performed using the ANN and LR model. Since the response variables have two levels, the area under curve (AUC) was used as a criterion to find the best model for predicting (0.05 was chosen for significant level).

Results

The used variables and their levels together with the frequencies are shown in Table 1. According to the result, the most frequency belongs to non-diabetic people. The sample prevalence of diabetes and hypertension was almost 12% and 13%, respectively.

The results of the joint LR for concurrent affliction to both diseases are summarized in Table 2.

After calculating the mean square errors for the different ANN models, the best ANN with two middle layers with 11 neurons in the first middle and 10 neurons in the second middle layer was chosen and this ANN was used for modeling of concurrent affliction to both hypertension and diabetes.

According to the result, age (OR=1.89, $p<0.001$), family history (OR=1.17, $p<0.001$), and physical activity (OR=1.13, $p<0.001$) were the variables which had significant relationship with concurrent affliction to both diseases.

To compare the precision of the joint LR model and ANN to the concurrent affliction of the both diseases, the AUC for joint LR model and ANN were 0.78 ($p=0.039$) and 0.86 ($p=0.046$), respectively. Therefore,

Table 1. The frequency distribution of participants' variables

Variables	Levels	N	%
Gender	Male	6005	50.0
	Female	5995	50.0
Type of cooking oil	Solid	4282	35.7
	Others	7703	64.3
Age (Year)	Equal or more than 55	4740	39.5
	Less than 55	7160	60.5
Family history	Yes	2337	19.5
	No	9663	80.5
Passive smokers	Yes	1489	12.4
	No	10511	87.6
Obesity	Yes	9449	79.0
	No	2506	21.0
Physical activity	Low	4404	36.7
	Moderate	2913	24.3
	High	4683	39.0
History of hypertension	Yes	1572	13.1
	No	10428	86.9
History of Diabetes	Yes	1483	12.4
	No	10517	87.6

Table 2. The result of logistic regression model

Variables	Category	Estimation	SE	OR	95% CI for OR		p
					Lower	Upper	
Family history	Yes	0.159	0.024	1.17	1.12	1.23	<0.001
	No	Reference Category					
Obesity	No	-0.013	0.020	0.99	0.95	1.03	0.516
	Yes	Reference Category					
Physical activity	No	0.121	0.019	1.13	1.09	1.17	<0.001
	Yes	Reference Category					
Passive smokers	No	-0.012	0.027	0. 99	0.94	1.04	0.660
	Yes	Reference Category					
Type of cooking oil	Other	-0.015	0.009	0.98	0.97	1.00	0.094
	Solid	Reference Category					
Age (Year)	More than 55	0.63	0.011	1.89	1.84	1.92	<0.001
	Less than 55	Reference Category					
Gender	Female	0.001	0.018	1.00	0.97	1.04	0.965
	Male	Reference Category					

ANN method has more precision in concurrent affliction to hypertension and diabetes than the joint LR.

Discussion

Hypertension and diabetes are from the most frequent health problems in the world which allocate the considerable proportion of health resources and facilities in developing countries. These diseases are from common diseases in Iran as well, which are categorized as non-communicable diseases (14).

Some researches on ANN were performed in different countries. In a research by Jefferson and colleagues (15), ANN and LR model were compared for predicting post-surgery complication in cancer patients. It was proved that ANN has better result than LR for complication prediction. This result is the same as current result. Green et al (13) showed in their research in 2006 on prediction of acute coronary syndrome in the emergency room, that ANN has better prediction than LR using ROC curve, which is the same as this study.

Jaimes et al (16) in their study in 2005, using ANN and LR model for predicting the patients' death rate with suspected sepsis in emergency room, showed that area under the ROC curve for LR and ANN were 0.7517 and 0.8782, respectively. Therefore, ANN has better performance than LR model to predict the death rate and this result is the same as current study.

It is necessary to say that in the above studies, the models are univariate but in this study a joint LR model was applied.

There are also some similar local studies as well. Sedehi (17) and colleagues performed a study to find healthy people or patients with metabolic syndrome. They compared the results of LR, discriminate analysis, (1:8:15) ANN, and (1:10:15) ANN to predict metabolic syndrome. According to the result, the correct prediction for LR, discriminate analysis, (1:8:15) ANN model, and (1:10:15) ANN were 72.4%, 66.7%, 73.6% and 87.4%, respectively. The results suggest that ANN has better performance than LR which is the same as in the current study.

In another study by Biglarian et al (18) in 2010 on survival prediction of gastric cancer patients after surgery, ANN and Cox Regression model were used. It was shown that the area under the ROC curve for ANN and Cox Regression model were 0.826 and 0.754, respectively. Therefore ANN has better performance for survival prediction. The result of this study in choosing ANN model as better model is the same as this study.

Conclusion

We conclude that ANN has more precision than other models. Also, it is possible to use the best topology of ANN for concurrent affliction to hypertension and diabetes to control these non-communicable diseases using prevention methods.

It is suggested to use some new variables such as stress and smoking to find better predictions. Also, precision comparison of the ANN with different topology with Bayesian models is suggested. Besides, using ANN in survival analysis and comparison of ANN with time series models (ARIMA, ARMA) is recommended for future study.

Acknowledgement

This study was founded and supported by Iran University of Medical Sciences; grant number: 19303. The study was approved in the Ethical Committee of Iran University of Medical Sciences. Also, the authors have no competing interest.

References

1. Regan MM, Catalano PJ. Likelihood models for clustered binary and continuous outcomes Application to Developmental toxicology, Biometrics 1999; 55: 760-768.

2. Sammel MD, Ryan LM, Legler JM. Latent variable models for mixed discrete and continuous outcomes, Journal of the Royal Statistical Society. Series B (Methodological) 1997; 59: 667-678.

3. Kutner MH, Nachtsheim CJ, Neter J, Li W. Applied linear regression models, 4th ed., New York: McGraw-Hill/Irwinn 2004.

4. Jobson JD. Applied Multivariate Data Analysis: Categorical and Multivariate Methods, New York: Springer, 1992.

5. Lindsey JK, Jones B. Choosing among generalized linear models applied to medical data, Stat. Med 1998; 17: 59-68.

6. Kay JW, Titterington TW. Statistics and neural networks: Advanced at the interface. Oxford: Oxford University Press1999.

7. Anderson JA. An Introduction to Neural Networks. Cambridge: MIT Press, 1995.

8. Sadeghian S. The knowledge of hospitalized patients about major risk factors of IHD in University hospitals of Tehran, Daneshvar Medicine 2001; 35:55-60.[Persian]

9. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin. Epidemiol 1996 .; 49(11):1225-31.

10. Fausett L. Fundamentals of Neural Networks Architectures, Algorithms and Applications. Prentice Hall: 1994.

11. Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. Lancet 1995; 346: 1075–1079.

12. Mitchell TM. Machine Learning. Boston: McGraw Hill, 1997.

13. Preventing Chronic Diseases. A Vital Investment, World Health Organization, 2005.

14. Jefferson MF, Pendleton N, Lucas SB, Horan MA. Comparison of a genetic algorithm neural network with logistic regression for predicting outcome after surgery for patients with non-small cell lung carcinoma. Cancer 1997;79:1338-1342.

15. Green M, Bjork J, Forberg J, Ekelund U, Edenbrandt L, Ohlsson M. Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room, Artif intel med 2006, 38:305-318.

16. Jaimes F, Farbiarz J, Alvarez D, Martínez C. Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room, Critical Care 2005, 9: R150-R156.

17. Sedehi M, Mehrabi Y, Kazemnejad A, Hadaegh, F. Comparison of Artificial Neural Network, Logistic Regression and Discriminant Analysis Methods in Prediction of Metabolic Syndrome, Ir J Endocrino & Metabol 2010, 11:639-646.

18. Biglarian A, Hajizadeh E, Kazemnejad A. Comparison of artificial neural network and Cox regression models in survival prediction of gastric cancer patients, Koomesh 2010, 11: 215-220. [Persian].