



## **Comparison of Logistic Regression and Artificial Neural Network in Low Back Pain Prediction: Second National Health Survey**

*M Parsaeian<sup>1</sup>, K Mohammad<sup>1</sup>, M Mahmoudi<sup>1</sup>, \*H Zeraati<sup>1</sup>*

*Dept. of Epidemiology and Biostatistics, Tehran University of Medical Sciences, Tehran, Iran*

**(Received 19 Feb 2012; accepted 12 Apr 2012)**

### **Abstract**

**Background:** The purpose of this investigation was to compare empirically predictive ability of an artificial neural network with a logistic regression in prediction of low back pain.

**Methods:** Data from the second national health survey were considered in this investigation. This data includes the information of low back pain and its associated risk factors among Iranian people aged 15 years and older. Artificial neural network and logistic regression models were developed using a set of 17294 data and they were validated in a test set of 17295 data. Hosmer and Lemeshow recommendation for model selection was used in fitting the logistic regression. A three-layer perceptron with 9 inputs, 3 hidden and 1 output neurons was employed. The efficiency of two models was compared by receiver operating characteristic analysis, root mean square and -2 Loglikelihood criteria.

**Results:** The area under the ROC curve (SE), root mean square and -2Loglikelihood of the logistic regression was 0.752 (0.004), 0.3832 and 14769.2, respectively. The area under the ROC curve (SE), root mean square and -2Loglikelihood of the artificial neural network was 0.754 (0.004), 0.3770 and 14757.6, respectively.

**Conclusions:** Based on these three criteria, artificial neural network would give better performance than logistic regression. Although, the difference is statistically significant, it does not seem to be clinically significant.

**Keywords:** Artificial neural network, Logistic regression, Low back pain, Second national health survey

### **Introduction**

Nearly everyone at some point suffers from low back pain (LBP) that interferes with his work and recreational activities (1, 2). In a few cases, there is a serious cause but generally, it is not possible to identify a specific cause of the pain. Symptoms, pathology, and radiological appearances are poorly correlated (3). It is often reported that there is no evident objective findings in 80 to 90% of back pain cases and therefore difficult to establish a pathological basis of the pain (4).

Since LBP is a multifactor disease, each factor influencing it gets a small part of the risk, so has little effect on preventing the disease. Contrary to the application of modern and advanced techniques, experts have failed to diagnose and fully comprehend the real cause of LBP (3, 5). All these flaws are indicative of the fact that we need to conduct more extensive and profound researches together with using more precise statistical procedures to specify the risk factors of this disease.

---

**\*Corresponding Author:** E-mail: zeraatih@tums.ac.ir Fax: 00982188989127

Available at: <http://ijph.tums.ac.ir>

In recent years, successful applications of artificial neural network (ANN) to predict medical outcomes have been demonstrated in many articles (6-9). ANNs are flexible mathematical data modeling tools that are extension of traditional model based methods. They have been used in situations where analyses based on discriminant analysis or logistic regression would have been standard statistical techniques. In this paper, we have predicted LBP, through its associated risk factors using two models: logistic regression, a widely used model for prediction of binary data, and ANN, a new and more sophisticated model. Hence, the purpose of this article is to compare empirically the predictive ability of logistic model with ANN in prediction of LBP.

## Materials and Methods

### Data collection

This study was based on the information obtained from the second national health survey in the year 2000, in Iran. Sampling design was described in a previously published article (10). Totally, 34589 people aged 15 years and older were interviewed. The information was obtained by means of a questionnaire, which included demographic, personal habits and personal conditions (11). We focused our study on the information of low back pain and its associated risk factors among Iranian adult people. LBP was defined as a non-traumatic pain in back and lumber region with no stiffness or with stiffness that relived by rest. Stiffness in this study was a kind that lasted less than an hour. Nine factors, including age (year), gender, education (literate, illiterate), residential area (urban, rural), smoking habits (smoker, non-smoker), hard working conditions (yes, no), body mass index (BMI), mental disorder (yes, no), marital status (married, unmarried) were used to compare the performance of these two models. Agriculture, animal husbandry, and laboring were considered as hard working condition. Mental disorders were assessed by using the General Health Questionnaire (GHQ) including 28 items. Scores equal to or greater than six were classified as suspected mental disorder. These variables were selected based on Hosmer and Lemeshow recommendation for model selection in

the logistic regression (12). All binary variables were coded as 0 for absence and 1 for presence of that characteristic. The marital status was considers 0 for unmarried subjects and 1 for married ones. Males were coded as 0 and females were coded as 1. In addition, urban areas were coded as 0 and rural areas were coded as 1.

### Statistical analysis

The whole dataset is first split into two parts, ANN and logistic model were developed using the first part (n=17294) and they were validated in the second part (n=17295) which we have called the latter the test set. All above variables were used with two prediction tools, logistic regression, and ANN, to determine the prevalence of LBP.

### Artificial Neural Network

A supervised multilayer perceptron, the most popular artificial neural network, was performed by Statistica neural network software (release 3.0 D). This network is comprised of an input layer, a hidden layer, and an output layer. Unlike the number of nodes in the input and output layers which is determined due to data structure, finding the optimum number of hidden node is a crucial step in the architecture of the neural network. The number of hidden nodes greatly increases the learning capability of the network, which may lead to over fitting of the training data. Indeed, excess nodes in the hidden layer may make ANNs learn the training examples too correctly while they cannot be generalized to new cases. This situation arises in over complex networks when the ability of the network remarkably exceeds the needed free parameters. The most popular method of finding the optimal number of hidden nodes is cross-validation technique. This procedure requires two distinct set. A training set, which is used to learn patterns presented in the data and verification, set which is used to evaluate over-fitting. Accordingly, the first part of the data is randomly divided into two sets, a training set (n=13835) and a verification set (n=3459). We trained different network by changing the number of nodes in the hidden

layer and compare the performance of these networks by root mean square (RMS) in the verification set. Since the network with three nodes had the minimum RMS, we use three hidden nodes in our architecture. The network was trained by back-propagation algorithm with the learning rate (weight adjustment at each iteration) in the range of (0.1-0.5) and momentum value (how past weight changes affect current weight changes) in the range of (0.1-0.3) on the training set. The activation function for both hidden layer and output layer was the sigmoid function. The training was stopped when there is no decrease in RMS after 100 epochs in the training set. By altering the initial weights values, learning rate and momentum value, we repeat the training algorithm. Then the best network, which has the lowest RMS in the verification set, was considered as a properly trained network. Finally, this network was used as a predictive tool in the test set to assess the accuracy of ANN for unseen data.

### **Logistic regression**

A well-known statistical method for modeling a binary response variable is logistic regression. Like in ANN, Logistic regression was developed based on the nine aforesaid variables in the first part of the data (Stata software program, version 10.0). For both of ANN and logistic regression models, the area under the receiver operating characteristic (ROC), RMS and -2Loglikelihood criteria was calculated in the test set. In all data analysis, p-values of 0.05 or less were considered significant.

### **Result**

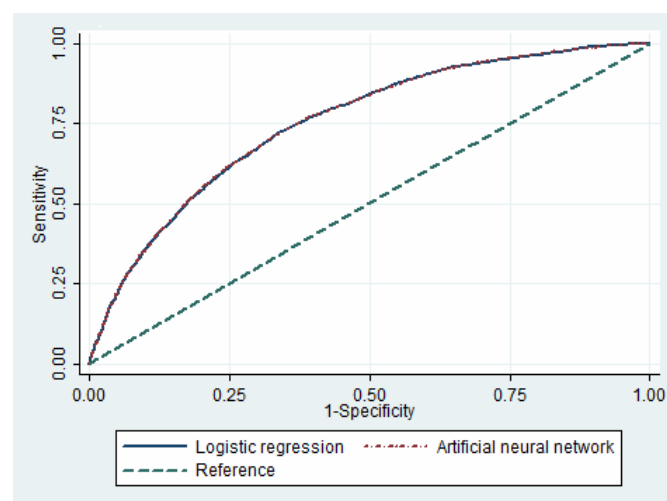
The study population was composed of 34589 people of whom 7286 people (21.4%) had LBP. Descriptive statistics of this population is shown in Table 1.

An ANN with nine input nodes, three hidden nodes, one output node, and sigmoid activation function in both hidden and output layer was trained. With the same variables, a main effect logistic regression was fitted to the first part of the data. The result can be summarized in the

following way: LBP was significantly associated with age, OR (CI):1.02(1.01, 1.02), gender 2.76 (2.47, 3.08), education 0.74(0.66, 0.82), residential area 1.24 (1.13, 1.36), smoking 1.45(1.29, 1.63), hard working condition 1.24 (1.08, 1.42), BMI 1.03(1.02, 1.04), mental disorder 1.93 (1.76, 2.11), marital status 2.50 (2.18, 2.85).

In order to test the generalization ability, we evaluate ANN and logistic regression in the test set using the area under ROC curves. As shown in Table 2 the areas under the ROC curves were 0.752 and 0.754 for the logistic regression and the ANN, respectively ( $p=0.0035$ ). The ROC curve were displayed in Fig. 1.

In addition, RMS and -2Loglikelihood of these two models was calculated. The RMS and -2Loglikelihood are single summary measures, which compare the observed with the estimated probability of LBP. The RMS and -2Loglikelihood of the logistic regression was 0.3832 and 14769.2 respectively. The RMS and -2Loglikelihood of ANN was 0.377 and 14757.6 respectively. Based on these three indices, neural network provide a better fit on the test set by comparison with logistic regression. Although, the difference is statistically significant, it does not seem to be clinically significant.



**Fig. 1:** ROC curves for artificial neural network and logistic regression in prediction of low back pain in the test set



**Table 1:** Prevalence of low back pain according to gender, age, place of residence, smoking and other socioeconomic factors in a random sample of 34589 Iranian people, in year 2000

| Characteristics        | LBP           |               | P      |
|------------------------|---------------|---------------|--------|
|                        | Yes           | No            |        |
| Gender                 |               |               | < 0.01 |
| Male                   | 2131† (13.3)  | 13771 (86.7)  |        |
| Female                 | 5173 (28.4)   | 13069 (71.6)  |        |
| Residential Area       |               |               | < 0.01 |
| Urban                  | 4414 (19.9)   | 17812 (80.1)  |        |
| Rural                  | 2872 (24.1)   | 9028 (75.9)   |        |
| Education              |               |               | < 0.01 |
| Literate               | 4347 (16.5)   | 21977 (83.5)  |        |
| Illiterate             | 2935 (37.9)   | 4814 (62.1)   |        |
| Hard Working Condition |               |               | < 0.01 |
| Yes                    | 989 (17.6)    | 4633 (82.4)   |        |
| No                     | 6297 (22.1)   | 22203 (77.9)  |        |
| Smoking                |               |               | < 0.01 |
| Yes                    | 1262 (25.4)   | 3715 (74.6)   |        |
| No                     | 6022 (20.7)   | 23100 (79.3)  |        |
| mental disorder        |               |               | < 0.01 |
| Yes                    | 2327 (35.8)   | 4166 (64.2)   |        |
| No                     | 4793 (18.3)   | 21378 (81.7)  |        |
| Marital Status         |               |               | < 0.01 |
| Unmarried              | 742 (7.2)     | 9589 (92.8)   |        |
| Married                | 6544 (27.5)   | 17243 (72.5)  |        |
| Age (Mean ± SD)        | 42.65 (16.96) | 32.82 (15.98) | < 0.01 |
| BMI (Mean ± SD)        | 24.98 (4.91)  | 23.52 (4.43)  | < 0.01 |

\*Differences between the patients and the controls were analyzed by the chi-square test for categorical variables and t-test for continuous variables.

† different number of total observation for each variable is the result of different response rate.

**Table 2:** Comparison of logistic regression and artificial neural network by the area under the ROC curve

| Models                    | Asymptotic 95% Confidence Interval |            |                  |             |             |
|---------------------------|------------------------------------|------------|------------------|-------------|-------------|
|                           | Area                               | Std. Error | Asymptotic Sig.* | Lower Bound | Upper Bound |
| Logistic regression       | 0.752                              | 0.004      | 0.001            | 0.744       | 0.761       |
| Artificial neural network | 0.754                              | 0.004      | 0.001            | 0.745       | 0.762       |

\* Null hypothesis: true area=0.5

## Discussion

In this article, an attempt was made to evaluate the accuracy level of two pattern classifiers, conventional logistic regression, and ANN. The data from the second national health survey were considered for model comparison. The results show

that ANN performs well comparing with its classical alternative.

ANNs are semi parametric classifiers, which are more flexible than parametric models but require fewer parameters than nonparametric methods, which are totally flexible (13). A very appealing

characteristic of these automated networks is their learning method, i.e. learning by examples. When there is little knowledge of the actual relationship, this characteristic makes them more powerful in pattern recognition. Moreover, since ANN has no limitation regarding its formulated function, it is more flexible and has more strength in mimicking complicated patterns than logistic regression (14). Another desirable feature of ANNs is their ability to find patterns despite of missing data. They are robust networks, which are tolerant in probing incomplete noisy patterns (15).

Although powerful in concept, both the logistic regression and the ANN have almost similar performance in predicting LBP. Similar results has been reported elsewhere. In about 90% of studies with large sample size, both ANN and logistic regression have identical performance (16).

There are some drawbacks in practical application of ANNs: firstly, designing a network is not so easy and good understanding of the fundamental theory is necessary. Secondly, there are no formal techniques to test the relative relevance of the independent variables and to carry out the variable selection process in non-linear methods (17). Thirdly, there is no etiologic interpretation for the calculated weights in the network as compared with classical models (18). We cannot determine a mathematical relationship between target and input variables (17). Finally, learning an artificial neural network is computationally time consuming and require sophisticated software.

Notwithstanding these drawbacks, there are some situations where the use of traditional models is impossible and we need some other alternatives. Using the classical models require many assumptions that may not be true in some real applications. Violation of these assumptions may produce error in prediction and hypothesis testing (19). In addition, logistic models use linear combination of variables and, therefore, are not suitable for modeling multifaceted relationships (15, 20). The inability to capture pattern complexity and inability to capture process dynamic are two major pitfalls of traditional methods.

As a general conclusion, when complex dependencies and interactions exist in the dataset, ANN may be the best choice. Conversely, when the main purpose of modeling is causal inference among variables and we want to identify the effect of each variable on the response variable, logistic regression is particularly useful (21).

One the important advantage of our study is external validation of modeling results. The main purpose of modeling is to use this learned patterns for new cases. Therefore, in order to avoid over fitting, external validation is essentially required. Unfortunately, in many articles, there is no external validation and building and assessment of the model is done on the same data set. Also we employ a large data set, so our result may be considered as an evidence for comparison of ANN and logistic regression with large sample sizes.

This study also has some limitation: Firstly, the definition of LBP in this study was subjective rather than diagnostic and the study population was a part of general national health survey of which relatively few direct questions could be allocated to LBP. Including questions about other risk factors may improve the diagnostic accuracy of both logistic regression and ANN as predictive models and would permit their reliable use in a clinical setting. Secondly, Multilayer feed-forward which is a basic form of neural network was used in this study. There have been well-designed modifications to the neural network model to extend its range of utilization. Further studies are needed to use other neural network architecture and compare these networks to find whether there are practical or clinical advantages of one approach over the others.

### **Ethical considerations**

Ethical issues (Including plagiarism, Informed Consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc) have been completely observed by the authors.

## Acknowledgements

The authors wish to acknowledge participants in the Second National Health for providing database, assistance, and technical support. The authors declare that there is no conflict of interests.

## References

1. Foster NE (2011). Barriers and progress in the treatment of low back pain. *BMC Medicine*, 9:108.
2. Hahne AJ et al. (2011). Specific treatment of problems of the spine (STOPS): design of a randomised controlled trial comparing specific physiotherapy versus advice for people with subacute low back disorders. *BMC Musculoskelet Disord*, 12: 104.
3. Chou R (2011). LowBackPain(Chronic). *American Family Physician*, 84(4): 437-8.
4. Hagen EM. Light mobilization treatment of subacute, non-specific low back pain [PhD thesis]. Department of Physical Medicine and Rehabilitation, University of Bergen, Norway; 2006.
5. Avins AL (2010). Glucosamine and the ongoing enigma of chronic low back pain. *JAMA*, 304(1): 93-4.
6. El-shafie A, Mukhlisin M, Najah AA, Taha MR (2011). Performance of artificial neural network and regression techniques for rainfall-runoff prediction. *International Journal of the Physical Sciences*, 6(8): 1997-2003.
7. Ludwig SA, Roos S, Frize M, Yu N (2010). Medical Outcome Prediction for Intensive Care Unit Patients. *IJCMAM*, 1(4): 19-30.
8. Hearty AP, Gibney MJ (2008). Analysis of meal patterns with the use of supervised data mining techniques-artificial neural networks and decision trees. *Am J Clin Nutr*, 88(6):1632-42.
9. Suzuki K (2011). *Artificial Neural networks Methodological Advances and Biomedical Application*. 1<sup>st</sup> ed. InTech, India, p.:71.
10. Bakhshi E, Eshraghian MR, Mohammad K et al. (2008). Sociodemographic and smoking associated with obesity in adult women in Iran: Results from the national health survey. *J Public Health*, 30 (4): 429-435.
11. Saiepour N, Mohammad K, Abhari R, Zeraati H, Noorbala AA (2008). Mental disorder assessed by general health questionnaire and back pain among postmenopausal iranian women. *Pak J Biol Sci*, 11 (5): 809-12.
12. Hosmer DW, Lemeshow S (2000). *Applied logistic regression*. 2<sup>nd</sup>. John Wiley & Sons, New York, pp.:92-99.
13. Dybowski R, Gant V (2001). *Clinical applications of artificial neural networks*. 1<sup>st</sup> ed. Cambridge University Press, London, p.:5.
14. Song JH, Venkatesh SS, Conant EA, Arger PH, Sehgal CM (2005). Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses. *Acad Radiol*, 12: 487-495.
15. Eftekhari B, Mohammad K, Ardebili HE, Ghodsi M, Ketabchi E (2005). Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. *BMC Med Inform Devis Mak*, 5 (3).
16. Bartfay E, Mackillop WJ, Pater JL (2006). Comparing the predictive value of neural network models to logistic regression models on the risk of death for small-cell lung cancer patients. *Eur J Cancer Care (Engl)*, 15(2):115-24.
17. Gan C, Limsombunchai V, Clemes M, Weng A (2005). Consumer choice prediction: Artificial neural networks versus logistic models. *Journal of Social Sciences*, 1 (4): 211-219.
18. Jaimes F, Farbiarz J, Alvarez D, Martinez C (2005). Comparison between logistic regression and neural networks to predict

- death in patients with suspected sepsis in the emergency room. *Crit Care*, 9:150-156.
19. Sadat-Hashemi SM, Kazemnejad A, Lucas C, Badie K (2005). Predicting the type of pregnancy using artificial neural networks and multinomial logistic regression: A comparison study. *Neural Comput & Applic*, 14 (3): 198–202.
  20. Eken C, Bilge U, Kartal M, Eray O (2009). Artificial neural network, genetic algorithm, and logistic regression applications for predicting renal colic in emergency settings. *Int J Emerg Med*, 2: 99-105.
  21. Venkatesan P, Anitha S (2006). Application of a radial basis function neural network for diagnosis of diabetes mellitus. *Current Science*, 91 (9): 1195-1199.