Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2018;7(2):350-4

Estimation of total prostate specific antigen values through artificial intelligence modelling

Seyma Yasar[1], Ali Beytur[2], Cemil Colak[1]

[1] *Inonu University Faculty of Medicine Department of Biostatistics and Medical Informatics, Malatya, Turkey*
[2] *Inonu University Faculty of Medicine Department of Urology, Malatya, Turkey*

Received 05 December 2017; Accepted 29 December 2017
Available online 20.03.2018 with doi: 10.5455/medscience.2017.06.8750

Abstract

It has been indicated that total prostate specific antigen (PSA) screening, one of the serum markers used for the diagnosis of prostate cancer, has been clinically beneficial. In this research, it was aimed to estimate the total PSA values by Multilayer Perceptron (MLP) artificial neural network (ANN) model. Data on total PSA values in this study (n = 1422) were randomly selected using the structured query language (SQL) from the database of patients records of Urology Department of Medical School at Inonu University. Total PSA values as a target/dependent variable, and age (year), blood group (A/B/0/AB), Exitus (EX) status (alive/death), Lymphocyte (LY) (%), Hemoglobin (HGB) (g / dL), Neutrophil (NE) (%), Albumin (g / dL), Calcium (mg / dL), Mean Corpuscular Hemoglobin (MCH), Leukocyte count (WBC) (103 / ml), Platelet (PLT) (103/ ml) as predictor variables were evaluated in the analyses. Outlier/extreme observations were analysed, and quantitative variables were rescaled by the transformation of Z-score or Box-Cox, and the MLP ANN model was constructed to estimate the total PSA values after variable selection method was used. Estimation performance of the model was examined by the values of mean absolute error, standard deviation and correlation coefficient. The MLP ANN model was created using a total of 1422 data sets as 993 of which were in training and 429 in the testing. Values of the mean absolute error, standard deviation, and correlation coefficient were calculated for training data set as; 0.744, 0.895 and 0.452; for the test data set as; 0.773, 0.935 and 0.355. The estimated accuracy of the generated model is predicted as 20.3%. In the MLP ANN model, the importance levels of the variables were obtained as 0.33 for HGB, 0.22 for NE, 0.16 for Calcium, 0.13 for PLT, 0.10 for age and 0.06 for EX. The MLP ANN model was established for the estimation of the total PSA values based on the selected variables, and calculated the importance levels of the related variables. Better prediction results in the estimation of total PSA values can be provided by using different additional variables, various resampling methods and alternative models.

Keywords: Artificial intelligence modelling, risk factors, total prostate specific antigen

Introduction

Prostate cancer is the most prevalent disease, especially in the elderly male population. It is reported that prostate cancer is the second most common cause of cancer deaths in males. Prostate cancer, a major health problem worldwide, is becoming increasingly common in society. The diagnosis of prostate cancer is made by various techniques. These are methods diagnosed by guided biopsy such as prostate specific antigen (PSA), digital rectal examination (PRM), free/total PSA ratio and transrectal ultrasonography (TRUS). The precise diagnosis of prostate cancer is made according to the pathological evaluation. The presence of prostate cancer in patients with prostate biopsy is the result of an unexpected state of PRM or elevated serum PSA level. The most prominent function of transrectal ultrasonography (TRUS) in the early diagnosis of prostate cancer is that it has pioneered the systematic prostate biopsy technique, which is the gold standard method of diagnosing the disease [1-5].

***Coresponding Author:** Seyma Yasar, Inonu University Faculty of Medicine Department of Biostatistics and Medical Informatics, Malatya, Turkey
E-mail: seyma.yasar@inonu.edu.tr

Artificial neural networks (ANNs) are an information processing system that is inspired by biological neural networks and include some performance characteristics similar to biological neural networks. Because of its ability to model both parametric and non-parametric relations, ANNs are one of the more frequently used methods in the literature in recent years. The ANNs, which mimic the way the human brain works in a simple way, have many important features such as being able to learn from the data, generalize, and work with many variables. Today, ANNs models are suitable for specific purposes and use in different areas. It is the multi-layer feedforward ANNs that have the most common usage area and are also used in our work [6,7].

Neurons in the MLP network are organized in layers. In MLP, the first layer is the input layer. The input layer provides the ANNs to retrieve information about the problem to be solved. The other layer is the output layer where information processed in the network is transmitted to the outside. MLP ANNs can have multiple hidden layers [8,9].

In this study, it was aimed to estimate the total PSA values in the patients with prostate cancer by MLP ANNs model.

Material and Method

The total PSA values of the 1422 patients used in this study were randomly selected from the database of patients records of Urology Department of Medical School at Inonu University by using structured query language (SQL). The descriptive information of the variables used in this context was given in Table 1.

Table1. Definitions of the variables used in the study

Variables	Variable Type	Variable Unit/ Categories	Variable Role
Total PSA	Numeric	ng/ml	Dependent/Output Variable
Age	Numeric	year	Independent/Input Variable
Blood Group	Categorical	A,B,0,AB	Independent/Input Variable
EX Status	Categorical	alive / dead	Independent/Input Variable
Lymphocytes	Numeric	%	Independent/Input Variable
Hemoglobin	Numeric	g/dL	Independent/Input Variable
Neutrophil	Numeric	%	Independent/Input Variable
Albumin	Numeric	g/dL	Independent Input Variable
Calcium	Numeric	mg/dL	Independent/Input Variable
Mean Erythro-cyte Hemoglobin	Numeric	pg	Independent/Input Variable
Leucocyte count	Numeric	103/ml	Independent/Input Variable
Platelet	Numeric	103/ml	Independent/Input Variable

Designing MLP ANNs Model

Outlier/extreme observations were analyzed, quantitative variables Z-score or Box-Cox transformation rescaled and variable selection method was used in the preliminary stage. The MLP is a model in which outputs are entered during the training phase and outputs expected to be generated for these inputs are shown. The generated network finds output according to the given input. The network structure consists of three different layers as shown in Figure 1. The layer that the incoming information is transmitted to the intermediate layer is the input layer. The middle layer can be one or more. In this layer, the information from the input layer is processed. In the output layer, output values are determined for each input in response to information received from the intermediate layer. Process elements are interconnected between all layers.

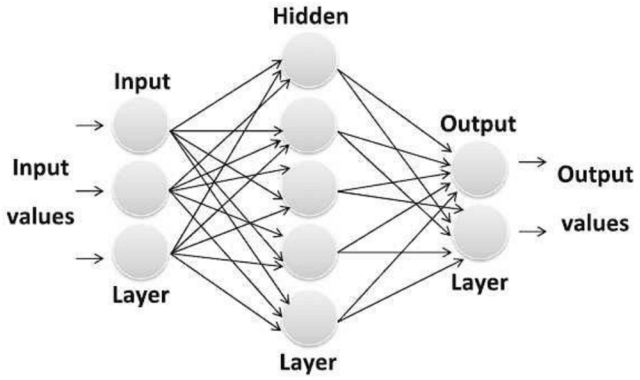


Figure 1. Topological Structure of Multilayer Perceptron Network

In the MLP model, there are two stages: a forward calculation where the output of the network is calculated and a backward calculation where the weights are updated. Let's assume that there are k process elements in the input layer during the phase of forward calculation. Accordingly, the output of the process elements in the input layer is determined by equation 1.

$$C_k^i = G_i \qquad 1$$

NET input values for process elements in the intermediate layer, A_{kj} where k: the input element, j: is the value of the weight attached to the intermediate layer element, are calculated by the equation 2.

$$NET_j^a = \sum_{k=1}^{n} A_{kj} C_k^i \qquad 2$$

Using NET input values in the intermediate layer and the activation function j., the output value of the intermediate layer element is found. The sigmoid function is usually used as the activation function. In case of the preference of sigmoid function, the output value of *j.* intermediate layer element is given by equation 3.

$$C_j^a = \frac{1}{1+e^{-(NET_j^a + \beta_j^a)}} \qquad 3$$

The value given as β^a in Equation 3 j. indicates the weight of the threshold value element connected to the element. The output of this threshold value element is equal to 1 and this value is fixed. The computation phase of the backward calculation starts with the error value. Let the (E_1,E_2,…) values are the expected value and (O_1,O_2,…) values are output values. Thus, the obtained E error values are calculated by Equation 4.

$$e_m = E_m - 0_m \qquad 4$$

The total error (TE) value occurring in the network is calculated given by equation 5 with the use of e_m values.

$$TE = 1/2 \sum_m e_m^2 \qquad 5$$

While MLP network is trained, it is aimed to reduce the worst of the error. The error that occurs after this calculation is distributed to weight values. Both the weights between the intermediate layer and the output layer are changed, and the weights between intermediate layers or the weights between the intermediate layer and the input layer are changed. Thus, the entire weight of the network will changed. The first iteration includes forward and backward calculations. A new example for the second iteration is shown to the network and the same operations are repeated for the second example [10].

In the MLP ANN model designed in this study, 7 input variables were used in the input layer. These input variables are HGB, NE, CAE, PLT, AGE and EX. In the hidden layer, 5 artificial neurons were used. The value in the output layer is Y_{TPSA}, which is the total prostate specific antigen value. The topological structure of the MLP ANN model constructed with 7 input variable input layers, 5 artificial neuron intermediate (hidden) layers and 1 output layer are shown in Figure 2.

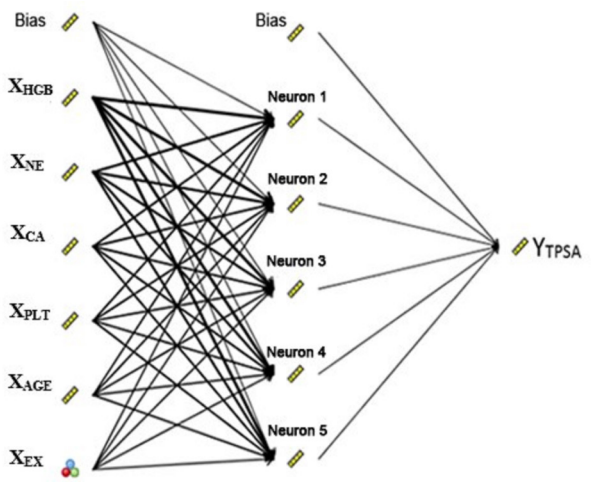


Figure 2. Architectural structure of the designed MLP ANN model

Biostatistical Analysis

IBM SPSS Statistics version 24.0 Windows package program was used in the analysis. Quantitative data were summarized as median (min-max) and qualitative data as number (%).The suitability of the data for normal distribution was assessed by the Kolmogorov Smirnov test. A value of p <0.05 was considered statistically significant.The predictive performance of the MLP ANN model constructed to estimate the PSA values was investigated with mean absolute error, standard deviation and correlation coefficient values.

Results

Descriptive statistics of the input and output variables are shown in Table 2.

Table 2. Descriptive statistics of the input and output variables

Input	Variables		Descriptive Statistics
X_{AGE*}	(Age)		73 (50-96)
X_{LY*}	(Lymphocytes)		1.8 (0.1-395)
X_{HGB*}	(Hemoglobin)		12.9 (1.8-161)
X_{NE*}	(Neutrophils)		4.8 (0.1-1157)
X_{ALB*}	(Albumin)		3.5 (0.4-5.1)
X_{CA*}	(Calcium)		9.1 (4.7-101)
X_{MCH*}	(Mean Corpuscular Hemoglobin)		29 (15.3-322)
X_{WBC*}	(Leucocyte count)		7.6 (0.8-1376)
X_{PLT*}	(Platelet)		244 (37-673)
X_{EX*}	(Ex status)		
		Alive	1242(87.3%)
		Death	180(12,7%)
XBLOOD*	(Blood Group)		
		A(+)	193(13.6%)
		A(-)	34(2.4%)
		B(+)	116(22.5%)
		B(-)	11(3.7%)
		0(+)	320(8.2%)
		0(-)	53(0.8%)
		AB(+)	19(1.3%)
Output	**Variable**		**Output Result**
Y_{TPSA*}	(Total PSA)		29.12 (2.5-7484)

*: Transformed Variables, The data are shown in median (min-max) and number (%).

The MLP ANNs model was created using a total of 1422 data sets, 993 of which were used in training and 429 were used in the testing. The absolute error, standard deviation and correlation coefficient values calculated from the model are given in Table 3.

Table 3. Information and Results of the MLP ANN Model

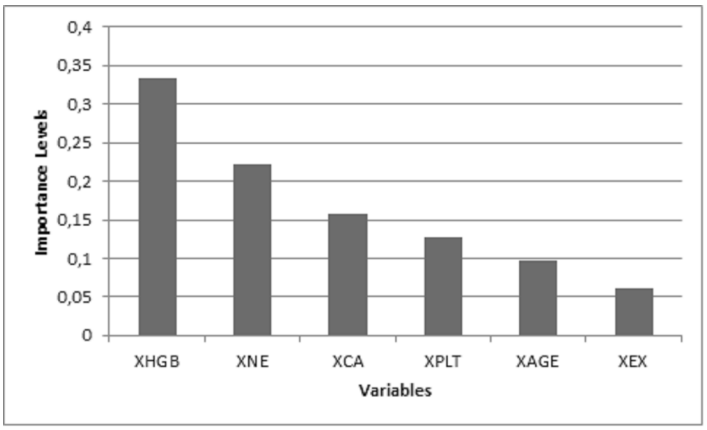
Data Set	n	Mean Absolute Error	Standard Deviation	Correlation
Training	993	0.744	0.895	0.452
Testing	429	0.773	0.935	0.355

The estimated accuracy of the generated model was estimated as 20.3%. The importance values of the variables in the MLP ANN model were given in Table 4.

Table 4. The importance values of the variables in the MLP ANN model

Variables	Importance Levels
XHGB	0.334
XNE	0.222
XCA	0.157
XPLT	0.128
XAGE	0.098
XEX	0.062

When table 4 was examined, while the HGB variable is the most important variable affecting the TPSA values, the EX variable has the lowest importance values. In addition, the importance values of the variables obtained from the model designed to estimate TPSA values were shown in Graph 1.



Graph 1. The importance values of the variables obtained from the model

Discussion

The present study intended to develop an ANN model to estimate total PSA values in patients with prostate cancer. Some studies reported that a significant incidence of prostate cancer happens in men with total serum PSA between 2.5 and 4.0 ng/mL [11]. Also, a study determined that nearly 20% of males in the screening procedures with total PSA levels in the range were diagnosed with prostate cancer [12]. From this clinical point of view, total PSA values were examined in the present study, and were estimated by the designed MLP ANNs model. In the ANNs architecture, MLP is used for constructing the predictive model based on the medical records of the studied patients. Principally, the data set was reorganized and the output and inputs variables were

determined to carry out the prediction task related to total PSA values in patients with prostate cancer. In the preprocessing stage, the data were inspected in terms of outlier/extreme observations, and quantitative variables were transformed to Z-score or Box-Cox transformation where appropriate. In the next stage, feature/variable selection technique was carried out for reducing the number of potential predictor variables. Then, MLP ANNs model was trained in the training data set, and was tested in the training data set. Afterwards, the prediction performance of the MLP ANNs model was assessed based on the defined metrics (i.e., mean absolute error, standard deviation and correlation coefficient).

Thereafter, significant factors that may be associated with TPSA values were determined from the obtained MLP ANNs model. Based on the results of the MLP ANNs model, the most significant variable associated with the TPSA values was HGB. According to the reported studies on prostate cancer based on risk factors for PSA decline (PSAD), HGB was significantly different between men who did not achieve a 30% PSAD and men who achieved a 30% PSAD [13]. Another study developed a pretreatment prognostic model for survival of patients with progressive metastatic prostate cancer. In the designed model, age, HGB and albumin were examined, and HGB and albumin were found to be significantly related to survival of patients with progressive metastatic prostate cancer [14]. Based on the findings of MLP ANNs model, neutrophil was the second important predictor for the total PSA values in prostate cancer. A number of researches reported that neutrophil-lymphocyte ratio (NLR) was a biomarker for metastatic prostate cancer and can be used for the prediction of prostate cancer[15,16]. Thus, this important finding inferred from this study is consistent with the reported studies[15,16]. Some clinical researches demonstrated the relation of calcium intake with the risk of prostate cancer, and concluded that high calcium intake can reduce the risk of advanced prostate cancer [17,18]. In this direction, the important results of the mentioned researches support the result of the current study. Platelet was another factor affecting total PSA values. As reported, tumor cells conduce activation of platelet [19]. When this information is evaluated, platelet may be associated with total PSA values that can be used for diagnosing prostate cancer. Many clinical studies report that advanced age is a significant risk factor for prostate cancer [20,21]. When this information is evaluated together, advanced age is an important factor related to the total PSA values. When the factors that may be associated with prostate cancer were examined, ex status had the least significance level to estimate total PSA values in patients with prostate cancer. As it is known, prostate cancer is one of the diseases that can cause death, and mortality rate is high [22]. For this reason, ex status may be related to the total PSA values that can be used to establish the pre-diagnosis of prostate cancer. However, more extensive clinical trials are needed to establish a precise relationships between the risk factors mentioned and total PSA values in patients with prostate cancer.

Conclusion

As a conclusion, the MLP ANN model was designed for the estimation of the total PSA values based on the selected variables, and calculated the significance levels of the related factors. Better prediction results in the estimation of total PSA values can be provided by using different additional variables, various resampling methods and alternative models.

Competing interests
The authors declare that they have no competing interest.

Financial Disclosure
The financial support for this study was provided by the investigators themselves.

References

1. Altunoluk B, Bahar MR, Eren M, et al. Prostat kanseri tanısında prostat spesifik antijen ve parmakla rektal muayene bulgularının rolünün değerlendirilmesi. JCEI. 2012;3(1):66-70.

2. Erol C. Prostat kanserinde görüntüleme teknikleri. Turkiye Klinikleri J Urology- Special Topics. 2017;10(2):121-32.

3. Çolak MC, Çolak C, Kocaturk H, et al. Predicting coronary artery disease using different artificial neural network models/Koroner arter hastaliginin degisik yapay sinir agi modelleri lie tahmini. The Anatol J Cardiol. 2008;8(4):249-55.

4. Gulkesen K, Koksal I, Bilge U, et al. Comparison of methods for prediction of prostate cancer in Turkish men with PSA levels of 0-10 ng/mL. J BUON. 2010;15(3):537-42.

5. Lawrentschuk N, Lockwood G, Davies P, et al. Predicting prostate biopsy outcome: artificial neural networks and polychotomous regression are equivalent models. Int Urol Nephrol. 2011;43(1):23-30.

6. Kaynar O, Taştan S. Zaman serisi analizinde Mlp yapay sinir ağları ve arıma modelinin karşılaştırılması. Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi. 2009(33):161-72.

7. Lee HJ, Hwang SI, Han S-m, et al. Image-based clinical decision support for transrectal ultrasound in the diagnosis of prostate cancer: comparison of multiple logistic regression, artificial neural network, and support vector machine. Eur Radiol. 2010;20(6):1476-84.

8. Kaya MO, Çolak C, Özdemir E. Prostat spesifik antijeni yardımı ile prostat kanserinin değişik yapay sinir ağı modelleri ile tahmini. İnönü Üniversitesi Sağlık Bilimleri Dergisi. 2013(1):19-22.

9. Psaltis D, Sideris A, Yamamura AA. A multilayered neural network controller. IEEE control systems magazine. 1988;8(2):17-21.

10. Gör İ,Çok katmanlı algılayıcı yapay sinir ağı ile lineer diferansiyel denklem sisteminin çözümü. 18 Akademik Bilişim Konferansı, 30 Ocak-5 Şubat 2016. Aydın, Türkiye, 236-240.

11. Stephan C, Jung K, Lein M. Editorial Comment. J Urol. 2017;180(2):543.

12. Catalona WJ, Smith DS, Ornstein DK. Prostate cancer detection in men with serum PSA concentrations of 2.6 to 4.0 ng/mL and benign prostate examination: enhancement of specificity with free PSA measurements. Jama. 1997;277(18):1452-5.

13. Armstrong AJ, Tannock IF, Wit Rd, et al. The development of risk groups in men with metastatic castration-resistant prostate cancer based on risk factors for PSA decline and survival. Eur J Cancer. 2010;46(3):517-25.

14. Smaletz O, Scher HI, Small EJ, et al. Nomogram for overall survival of patients with progressive metastatic prostate cancer after castration. J Clin Oncol. 2002;20(19):3972-82.

15. Sonpavde G, Pond GR, Armstrong AJ, et al. Prognostic impact of the neutrophil-to-lymphocyte ratio in men with metastatic castration-resistant prostate cancer. Clin Genitourin Cancer. 2014;12(5):317-24.

16. Templeton AJ, Pezaro C, Omlin A, et al. Simple prognostic score for metastatic castration-resistant prostate cancer with incorporation of neutrophil-to-lymphocyte ratio. Cancer. 2014;120(21):3346-52.

17. Giovannucci E, Rimm EB, Wolk A, et al. Calcium and fructose intake in

relation to risk of prostate cancer. Cancer. 1998;58(3):442-7.

18. Koh K, Sesso H, Paffenbarger R, et al. Dairy products, calcium and prostate cancer risk. Br J Cancer. 2006;95(11):1582-5.

19. Prandoni P, Falanga A, Piccioli A. Cancer and venous thromboembolism. Lancet Oncol. 2005;6(6):401-10.

20. Efstathiou JA, Bae K, Shipley WU, et al. Cardiovascular mortality after androgen deprivation therapy for locally advanced prostate cancer: RTOG 85-31. J Clin Oncol. 2009;27(1):92-9.

21. Middleton RG, Thompson IM, Austenfeld MS, et al. Prostate cancer clinical guidelines panel summary report on the management of clinically localized prostate cancer. J urol. 1995;154(6):2144-8.

22. Andriole GL, Crawford ED, Grubb RL 3rd, et al. Mortality results from a randomized prostate-cancer screening trial. N Engl J Med. 2009;360(13):1310-9.