Mathematical Biosciences and Engineering

MBE, 18(4): 3690–3698.
DOI: 10.3934/mbe.2021185
Received: 03 March 2021
Accepted: 09 April 2021
Published: 28 April 2021

http://www.aimspress.com/journal/MBE

Research article

Back propagation neural network model for medical expenses in patients with breast cancer

Feiyan Ruan[1,2], Xiaotong Ding[3], Huiping Li[1,*], Yixuan Wang[1], Kemin Ye[2] and Houming Kan[4]

[1] School of Nursing, Anhui Medical University, Hefei 230032, China
[2] Breast surgery, The First Affiliated Hospital of Anhui Medical University, Hefei 230022, China
[3] School of Nursing, Nanjing Medical University, Nanjing 211166, China
[4] Pain department, SIR RUN RUN Hospital of Nanjing Medical University, Nanjing 211166, China

*** Correspondence:** Email: 2184333239@qq.com.

Abstract: *Objective:* Breast cancer seriously endangers women's life and health, and brings huge economic burden to the family and society. The aim of this study was to analyze the medical expenses and influencing factors of breast cancer patients, and provide theoretical basis for reasonable control of medical expenses of breast cancer patients. *Methods:* The medical expenses and related information of all female breast cancer patients diagnosed in our hospitals from 2017 to 2019 were collected. Through SSPS Clementine 12.0 software, the back propagation (BP) neural network model and multiple linear regression model were constructed respectively, and the influencing factors of medical expenses of breast cancer patients in the two models were compared. *Results:* In the study of medical expenses of breast cancer patients, the prediction error of BP neural network model is less than that of multiple linear regression model. At the same time, the results of the two models showed that the length of stay and region were the top two factors affecting the medical expenses of breast cancer patients. *Conclusion:* Compared with multiple linear regression model, BP neural network model is more suitable for the analysis of medical expenses in patients with breast cancer.

Keywords: breast cancer; medical expenses; back propagation neural network; multiple linear regression; influencing factor

1. Introduction

Breast cancer is one of the most common malignant tumors in women. In 2013, there were about 180 million new cases of female breast cancer in the world [1,2]. In China, the incidence rate of breast cancer in 2013 was 42.02/10 million, ranking the first place in the incidence of cancer in women, and the mortality rate was 9.74/10 million, ranking fifth in the female cancer death cause [3–5]. Breast cancer seriously endangers women's life and health, and brings huge economic burden to the family and society [6].

At present, the traditional statistical methods to analyze the medical expenses and influencing factors of breast cancer include multiple linear regression, logistic regression analysis and so on [7–10]. Back propagation (BP) neural network is an artificial neural network model based on error back propagation [11]. Compared with traditional statistical methods, it has no special requirements for the type and distribution of data, and has some fault tolerance, so it has more advantages [12–15]. In recent years, BP neural network model and other data mining methods have also been applied to analyze the medical expenses of cancer patients such as gastric cancer, lung cancer, liver cancer, gynecological cancer and so on [16–18].

In this study, the medical expenses of female breast cancer patients in rural areas of Anhui Province were selected as the research object. After the BP neural network model was constructed, the results were compared with the multiple linear model. The aim was to verify the effectiveness and predictive power of the two models for the influencing factors of medical expenses of breast cancer patients, so as to provide scientific theoretical basis for reasonable control of medical expenses of breast cancer.

2. Materials and methods

2.1. General data

Table 1. Study variable and its assignment table.

Variable	Code	Value
Region	X1	1 = Feidong; 2 = Feixi; 3 = Changfeng; 4 = Lujiang
Diagnosis year	X2	1 = 2017; 2 = 2018; 3 = 2019
Medical payment	X3	1 = Rural insurance; 2 = Urban workers; 3 = Urban residents; 4 = Others
Operation	X4	1 = Yes; 2 = No
Chemotherapy	X5	1 = Yes; 2 = No
Radiotherapy	X6	1 = Yes; 2 = No
Clinical stages	X7	1 = I; 2 = II; 3 = III; 4 = IV; 5 = Unknown
Age (years)	X8	1 = < 45; 2 = 45～59; 3 = 60～4; 4 = ≥ 75
Length of stay (days)	X9	Logarithm of the actual value
Medical expenses	Y1	Logarithm of the actual value

According to the data of cancer registration report in Anhui Province, the medical records of all female breast cancer patients diagnosed from 2017 to 2019 and coded as C50 according to ICD-10

disease code were collected from four county people's hospitals (Feidong, Feixi, Changfeng, and Lujiang) in Anhui Province, including hospitalization number, medical insurance number, medical payment method, operation, chemotherapy and radiotherapy; Medical insurance number, name, date of birth, disease diagnosis, total cost of hospitalization, hospital and other information were collected from social security department. A total of 846 samples were collected. All research variables were assigned values (Table 1). After the samples with missing values and illogicality were excluded, 795 cases were included in the study, and the effective rate was 93.97%.

2.2. Back propagation neural network

BP neural network is the abbreviation of error back propagation neural network. It is a multilayer feedforward network trained by error back propagation algorithm. It is one of the most widely used neural network models at present. It is composed of an input layer, one or more hidden layers and an output layer. Each layer is composed of a certain number of neurons. Its structure is shown in Figure 1. First, the number of neurons in each layer is determined. The number of neurons in the input layer and the output layer is determined by the independent and dependent variables of the actual research. There is no unified calculation method for the number of neurons in the hidden layer. In this study, the formula (M represents the number of neurons in the hidden layer, n represents the number of neurons in the input layer, m represents the number of neurons in the output layer, and represents any value from 1 to 10) is used to calculate the number of neurons in the hidden layer. Secondly, network training. Through SSPS Clementine 12.0 software to simulate the training set, different BP neural network models are obtained. Finally, the network test is carried out. The test set is substituted into each model, and the model with the highest accuracy is the optimal model.

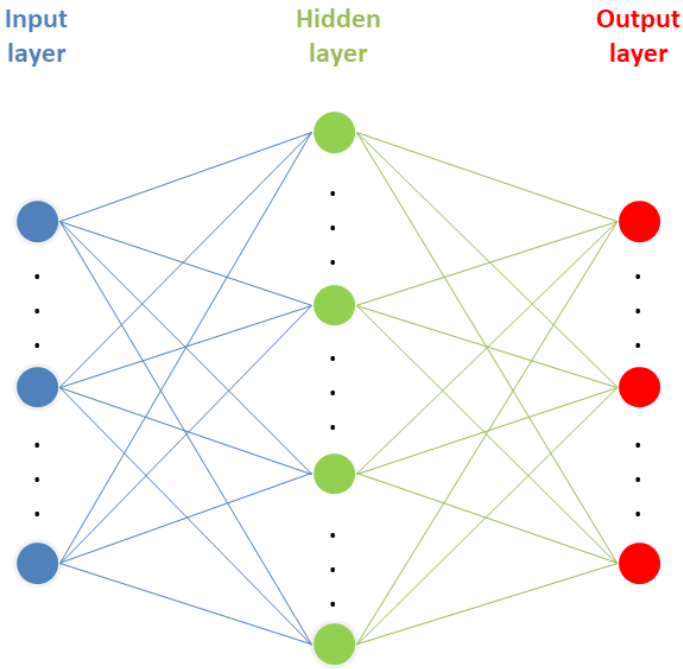

Figure 1. Structure of back propagation neural network model.

2.2. Multiple linear regression

Multiple linear regression is a regression analysis based on the given values of multiple explanatory variables. It is a method to study the linear relationship between a dependent variable and multiple independent variables [19]. The general form of multiple linear regression model is:

$$\mathbf{Y}=\boldsymbol{\beta_0}+\boldsymbol{\beta_1 x_1}+\boldsymbol{\beta_2 x_2}+\boldsymbol{\beta_3 x_3}+\boldsymbol{I}+\boldsymbol{\beta_i x_i}+\boldsymbol{e} \quad (1)$$

Where $\boldsymbol{\beta_0}$ is a constant term, $\boldsymbol{I}$ is the number of independent variables, $\boldsymbol{\beta_i}$ ($\boldsymbol{i}$ = 1, 2$\cdots$ $\boldsymbol{i}$) is partial regression coefficient, $\boldsymbol{e}$ is random error. The meaning of partial regression coefficient is the average change of dependent variable **Y** when the independent variable changes one unit while other independent variables remain unchanged. After the multiple linear regression parameters are obtained, it is necessary to carry out statistical tests to determine the reliability of the model performance, including the fitting test (coefficient of determination), the significance test of the overall linear equation (F test), the significance test of variables (t test), etc. [20].

2.3. Model evaluation method

The performance of the two models is evaluated by the following indicators, where $\boldsymbol{R^2}$ is the coefficient of determination, which can explain the percentage of the independent variable explaining the change of the dependent variable. The value range is between 0 and 1. The closer the value is to 1, the better the fitting degree of the model to the sample is [21]. The absolute mean error (MAE) reflects the actual error of the model, and the root mean square error (RMSE) is the arithmetic square root of the mean square error (MSE). The calculation formula of each index is as follows.

$$\boldsymbol{R^2}=\mathbf{1}-\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_l\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}_l\right)^2} \quad (2)$$

$$\mathbf{MAE}=\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{i=1}^{n}\left|\hat{\boldsymbol{y}}_{\boldsymbol{l}}-\boldsymbol{y_i}\right| \quad (3)$$

$$\mathbf{MSE}=\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{i=1}^{n}\boldsymbol{w_i}\left(\boldsymbol{y_i}-\hat{\boldsymbol{y}}_{\boldsymbol{i}}\right)^2 \quad (4)$$

$$\mathbf{RMSE}=\sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{i=1}^{n}\boldsymbol{w_i}\left(\boldsymbol{y_i}-\hat{\boldsymbol{y}}_{\boldsymbol{i}}\right)^2} \quad (5)$$

The predicted data of all patients were test by the normality test. The data both in back propagation neural network model and multiple linear regression model were consistent with normal distribution. After passing the homogeneity test, the difference between the two models were compared by *t* test.

3. Result

3.1. Basic information of patients

The median medical expenses of breast cancer patients were (24576 ± 4792) RMB, and the hospitalization days was (31.4 ± 6.7) d. The composition ratio of other variables is shown in Table 2.

Table 2. Basic information of hospitalized patients with breast cancer.

Variable	Case	Composition ratio (%)	Variable	Case	Composition ratio (%)
Region			Age		
Feidong	196	24.7	< 45	159	20.0
Feixi	258	32.5	45–59	378	47.5
Changfeng	164	20.6	60–74	214	26.9
Lujiang	177	22.3	≥ 75	44	5.5
Diagnosis year			Clinical stages		
2017	262	33.0	I	76	9.6
2018	254	31.9	II	348	47.8
2019	279	35.1	III	94	11.8
Medical payment			IV	26	3.3
Rural insurance	570	71.7	Unknown	251	31.6
Urban workers	64	8.1	Chemotherapy		
Urban residents	135	17.0	Yes	679	85.4
Others	26	3.3	No	116	14.6
Operation			Radiotherapy		
Yes	712	89.6	Yes	42	5.3
No	83	10.4	No	753	94.7

3.2. Modeling results

The SPSS Clementine 12.0 data mining platform was used to construct BP neural network model and multiple linear regression model for medical expenses of female breast cancer patients. BP neural network model takes nine indicators such as region, year of diagnosis and medical payment mode as input and logarithm of medical expenses as output. It adopts random sampling method and takes 70% samples as training set and 30% samples as test set. After repeated verification, a three-layer BP neural network model is finally constructed, with 9 neurons in the input layer, 10 neurons in the hidden layer and 1 neuron in the output layer, and the accuracy is 95.97%. The model results are shown in Table 3.

Table 3. Summary of Back propagation neural network model results.

Data set	Number of samples	Minimum error	Maximum error	Average error	Absolute mean error	Standard deviation	Linear correlation
Training set	557	−0.669	0.668	−0.003	0.127	0.167	0.889
Test set	238	−1.151	0.852	−0.004	0.149	0.221	0.863

In the multiple linear regression model, 9 indicators such as region, year of diagnosis and medical payment mode were taken as independent variables, and the logarithm of medical expenses was taken as dependent variable. The multiple correlation coefficient was 0.841, which indicated that the model fitted well. The probability of F statistic of the model population linear test ($P < 0.001$)

indicates that there is a significant linear relationship between independent variables and dependent variables. The model results are shown in Table 4.

Table 4. Summary of multiple linear regression models.

R	R^2	Adjust R^2	Error of standard estimation	*F*	*P*
0.838	0.696	0.689	0.202	126.647	< 0.001

3.3. Comparative analysis of influencing factors of medical expenses of breast cancer patients

BP neural network model gives the sensitivity of each variable, that is, the influence of each variable change on medical expenses. The analysis and comparison results of the two models are shown in Table 5. It can be seen that the top two influencing factors for medical expenses of breast cancer patients are length of stay and region; whether radiotherapy, surgery, age and chemotherapy also have a greater impact on medical expenses; medical payment method, diagnosis year and clinical stage have a smaller impact on medical expenses.

Table 5. Comparative analysis of influencing factors of medical expenses in patients with breast cancer.

Variable	Code	BP neural network model		Multiple linear regression model	
		Sensibility	Sort	Regression coefficient	sort
Length of stay	X9	0.552	1	0.731	1
Region	X1	0.161	2	0.259	2
Radiotherapy	X6	0.064	3	0.008	9
Operation	X4	0.057	4	0.082	3
Age	X8	0.045	5	0.038	6
Chemotherapy	X5	0.041	6	0.074	4
Medical payment	X3	0.035	7	0.051	5
Diagnosis year	X2	0.033	8	0.029	7
Clinical stages	X7	0.022	9	0.021	8

3.4. Comparison between back propagation neural network model and multiple linear regression model

All samples were substituted into the established BP neural network model and multiple linear regression model to evaluate the performance of the two models. The results are shown in Table 6. The coefficient of determination ($\boldsymbol{R^2}$) of BP neural network model was larger than that of multiple linear regression model, so the fitting degree of BP neural network model was better than that of multiple linear regression model. The MAE, MSE and RMSE values of BP neural network model were less than those of multiple linear regression model, so the prediction ability of BP neural network model was better than that of multiple linear regression model.

Table 6. Comparison between back propagation neural network model and multiple linear regression model.

Evaluation criteria	R^2	MAE	MSE	RMSE
BP neural network	0.771 ± 0.152	0.131 ± 0.028	0.034 ± 0.004	0.186 ± 0.034
Multiple linear regression	0.662 ± 0.117	0.168 ± 0.031	0.045 ± 0.005	0.221 ± 0.025
t/P value	8.767/ < 0.001	13.664/ < 0.001	26.503/ < 0.001	12.795/ < 0.001

4. Discussion

In this study, the median medical expenses and length of stay of breast cancer patients were (24576 ± 4792) RBM and (31.4 ± 6.7) d. The high cost of hospitalization not only affects the delay of treatment, but also hinders the choice of treatment [22,23]. At the same time, breast cancer brings huge disease burden to patients. Therefore, it is suggested to strengthen cancer screening for women, so as to achieve early detection, early diagnosis and early treatment, so as to reduce the economic burden and disease burden for individuals, families and society.

The results show that BP neural network model and multiple linear regression model can fit the data well. Through the comparative analysis of the influencing factors of medical expenses of breast cancer patients, it can be seen that the hospitalization days and regions are consistent in the two models, whether surgery, age of diagnosis, chemotherapy, medical payment method, year of diagnosis and clinical stage are basically consistent in the two models, but whether radiotherapy is inconsistent in the two models. The length of stay has the greatest impact on the cost of hospitalization, which is consistent with the existing research. Therefore, on the premise of ensuring the level of medical services, shortening the length of stay is an effective measure to reduce the medical expenses of breast cancer patients. There are also great differences in the medical expenses of breast cancer among regions, mainly due to the different economic development and medical technology levels. It is suggested that medical service providers can standardize the clinical pathway and provide efficient and affordable treatment for patients. Medical insurance is the main means to reduce the economic burden of rural patients. In this study, it has little impact on medical expenses. It is suggested that medical insurance policy makers should strengthen the compensation for rural cancer patients and reduce their economic risks.

In recent years, BP neural network model has been widely used in the field of medicine, and achieved good results [24,25]. Although BP neural network model has some disadvantages, such as over training, slow convergence speed, easy to fall into local optimum, it has no requirements for data type and distribution, has certain fault tolerance, and can correct errors repeatedly in the process of self-learning [26,27]. These advantages have great advantages in dealing with medical data with the characteristics of complexity and diversity. The results show that the determination coefficient of BP neural network model is greater than that of multiple linear regression model, and the values of MAE, RMSE and MSE are less than the corresponding values of multiple linear regression model, so its prediction ability is higher than that of multiple linear regression model.

5. Conclusions

Compared with multiple linear regression model, BP neural network model is more suitable for the analysis of medical expenses in patients with breast cancer. However, the model itself has no

advantages or disadvantages, only the applicable conditions of each model are different.

Acknowledgements

The authors would like to acknowledge the all the breast cancer patients for their participation and support.

Conflict of interest

The authors declared that there was no conflict of interests.

References

1. M. Akram, M. Iqbal, M. Daniyal, A. U. Khan, Awareness and current knowledge of breast cancer, *Biol. Res.*, **50** (2017), 33.
2. S. Winters, C. Martin, D. Murphy, N. K. Shokar, Breast cancer epidemiology, prevention, and screening, *Prog. Mol. Biol. Transl. Sci.*, **151** (2017), 1–32.
3. Z. Anastasiadi, G. D. Lianos, E. Ignatiadou, H. V. Harissis, M. Mitsis, Breast cancer in young women: an overview, *Updates Surg.*, **69** (2017), 313–317.
4. S. S. Coughlin, Epidemiology of breast cancer in women, *Adv. Exp. Med. Biol.*, **1152** (2019), 9–29.
5. M. A. Thorat, R. Balasubramanian, Breast cancer prevention in high-risk women, *Best Pract. Res. Clin. Obstet. Gynaecol.*, **65** (2020), 18–31.
6. F. Varghese, J. Wong, Breast cancer in the elderly, *Surg. Clin. North. Am.*, **98** (2018), 819–833.
7. S. I. Bangdiwala, Regression: multiple linear, *Int. J. Inj. Contr. Saf. Promot.*, **25** (2018), 232–236.
8. Y. H. Hu, S. C. Yu, X. Qi, et al., An overview of multiple linear regression model and its application, *Chi. J. Prev. Med.*, **53** (2019), 653–656.
9. R. Zemouri, N. Omri, C. Devalland, L. Arnould, B. Morello, N. Zerhouni, et al., *Breast cancer diagnosis based on joint variable selection and constructive deep neural network*, 2018 IEEE 4th Middle East Conference on Biomedical Engineering (MECBME), 2018.
10. R. Zemouri, N. Omri, B. Morello, C .Devalland, L. Arnould, N. Zerhouni, et al., Constructive deep neural network for breast cancer diagnosis, *IFAC PapersOnLine*, **51** (2018), 98–103.
11. S. Belciug, *Artificial Intelligence in Cancer: Diagnostic to Tailored Treatment*, Elsevier, New York, 2020.
12. Y. Deng, H. Xiao, J. Xu, H. Wang, Prediction model of PSO-BP neural network on coliform amount in special food, *Saudi. J. Biol. Sci.*, **26** (2019), 1154–1160.
13. Z. Li, Y. Li, A comparative study on the prediction of the BP artificial neural network model and the ARIMA model in the incidence of AIDS, *BMC Med. Inf. Decis. Mak.*, **20** (2020), 143.
14. X. Liu, Z. Liu, Z. Liang, S. P. Zhu, J. A. F. O. Correia, A. M. P. De Jesus, PSO-BP neural network-based strain prediction of wind turbine blades, *Materials*, **12** (2019), 1889.
15. R. Zemouri, N. Omri, F. Fnaiech, N. Zerhouni, N. Fnaiech, A new growing pruning deep learning neural network algorithm (GP-DLNN), *Neural Comput. Appl.*, **32** (2019), 18143–18159.

16. J. Li, W. Luo, Hospitalization expenses of acute ischemic stroke patients with atrial fibrillation relative to those with normal sinus rhythm, *J. Med. Econ.*, **20** (2017), 114–120.
17. M. E. Png, J. Yoong, C. S. Tan, K. S. Chia, Excess hospitalization expenses attributable to type 2 diabetes mellitus in Singapore, *Value Health Reg. Issues*, **15** (2018), 106–111.
18. J. Wang, P. Li, J. Wen, Impacts of the zero mark-up drug policy on hospitalization expenses of COPD inpatients in Sichuan province, western China: an interrupted time series analysis, *BMC Health Serv. Res.*, **20** (2020), 519.
19. B. Aline, A. M. Zeina, Z. Ryad, S. Valmary-Degano, Prediction of Oncotype DX recurrence score using deep multi-layer perceptrons in estrogen receptor-positive, HER2-negative breast cancer, *Breast Cancer*, (2020), 1007–1016.
20. N. Pandis, Multiple linear regression analysis, *Am. J. Orthod. Dentofacial Orthop.*, **149** (2016), 581.
21. D. G. Streeter, *Practical statistics for medical research*, New York, Chapman and Hall, 1991.
22. G. L. Yuan, L. Z. Liang, Z. F. Zhang, Q. L. Liang, Z. Y. Huang, H. J. Zhang, et al., Hospitalization costs of treating colorectal cancer in China: A retrospective analysis, *Medicine*, **98** (2019), e16718.
23. X. Zhuang, Y. Chen, Z. Wu, S. R. Scott, M. Zou, Analysis of hospitalization expenses of 610 HIV/AIDS patients in Nantong, China, *BMC Health Serv. Res.*, **20** (2020), 813.
24. J. Lyu, J. Zhang, BP neural network prediction model for suicide attempt among Chinese rural residents, *J. Affect. Disord*, **246** (2019), 465–473.
25. C. Zhang, R. Zhang, Z. Dai, B. Y. He, Y. Yao, Prediction model for the water jet falling point in fire extinguishing based on a GA-BP neural network, *PLoS One*, **14** (2019), e0221729.
26. R. Zemouri, N. Zerhouni, D. Racoceanu, Deep learning in the biomedical applications: recent and future status, *Appl. Sci.*, **9** (2019), 1526.
27. R. Zemouri, C. Devalland, S. Valmary-Degano, N. Zerhounid, Intelligence artificielle: quel avenir en anatomie pathologique?, *Ann. de Pathol.*, **39** (2019), 119–129.