Analysis of influencing factors of hospitalization cost based on BP neural network model*

Hongwei Xia*, Chaoqun Chenb
School of science, Changzhou institute of technology Changzhou, China
*a2227295639@qq.com, bchenchq@czu.cn

Abstract. In view of the present social status quo, "cost of medical care", according to a national disease surveillance center of 2700 medical record sample data, using BP neural network method, through SPSS software operation, get in 26 factors affecting hospitalization cost, age, level of hospital, hospital source code, disease types, cost the importance of the five factors relative to the highest, and the accuracy rate of more than 85%, so as to provide basis for government decision-making departments to formulate corresponding health policy.

1. Introduction
Recently, the movie "I am not the god of medicine" is popular all over the world and has become a hot topic of discussion after dinner and tea. The score of douban is as high as 9.2. The film's theme strikes at the heart of people's pain -- the difficulty and expense of getting medical treatment. "After decades of hard work, a serious illness." We have to admit that the medical reform for many years, the difficult and expensive phenomenon has not disappeared. What are the reasons for the high medical costs? In the abolition of drug addition, lower drug prices today, why are people still feeling sick? Now see a doctor expensive, basically be drug cost and examination cost, and medical consumable go up, wait like heart bracket, steel frame. However, hospitals can only get benefits from medical service charges and instrument inspection fees under the circumstance of eliminating drug addition and insufficient financial subsidy funds. In addition, the continuous improvement of medical care is also a factor leading to the increase of medical expenses. According to domestic and foreign studies, about 50-75% of the increase in medical expenses is due to technological progress. With the continuous progress of science and technology, the massive use of new technologies, new drugs, new materials and new equipment, as well as the continuous improvement of people's health requirements, also leads to the rise of medical expenses. The medical and health undertakings, which are related to social harmony, national stability and people's well-being, have a very prominent and important position.

This paper aims to explore the influencing factors of hospitalization cost by collecting relevant data and adopting appropriate mathematical models [1],

2. Data source and preprocessing
The monitoring center of a national hospital is responsible for the comprehensive data monitoring of all kinds in the country, and has accumulated millions of medical record data. We take the data from this database. In sampling, the principles of random sampling by region, classification and full coverage are followed. In addition to the first page of inpatient medical records, it also supplemented
the information of hospital conditions, treatment plans and other information. A total of 2,700 case data were sampled [2].

2.1. Hospitalization cost data distribution
First, the distribution of total hospital expenses and each itemized expense was investigated, and the descriptive statistical analysis function of SPSS software was used. The analysis results are shown in table 1:

| Table 1. Describes the statistical table |
|-----------------------------------------|
|                                        |
| The case number | The maximum | The average | The standard deviation |
|-----------------------------------------|
| Total hospital expenses                  | 2700 | 96965 | 4082.72 | 6995.05 |
| Daily room                               | 2700 | 4400 | 234.33 | 346.383 |
| Up for medicine                          | 2700 | 18614 | 335.04 | 1002.231 |
| Western medicine fee                     | 2700 | 66891 | 1771.42 | 3690.751 |
| Inspection fee                           | 2700 | 2981 | 120.08 | 223.033 |
| Blood fee                                | 2700 | 1320 | 5.79 | 69.834 |
| Oxygen fee                               | 2700 | 2608 | 28.4 | 142.334 |
| Operation fee                            | 2700 | 28977 | 992.36 | 2565.113 |
| Delivery fee                             | 2700 | 2758 | 62.57 | 231.823 |
| Other fee                                | 2700 | 114 | 0.04 | 2.194 |
| Emission fee                             | 2700 | 28978 | 160.13 | 723.545 |
| Emission assay                           | 2700 | 1810 | 35.84 | 93.823 |
| Chinese herbal medicine fee              | 2700 | 7103 | 91.25 | 311.918 |
| A nurse                                  | 2700 | 5895 | 106.07 | 272.233 |
| Treatment                                | 2700 | 6223 | 19.73 | 135.242 |
| Number of cases in force (listed)        | 2700 |      |      |      |

As can be seen from table 1:
1) The average of all the cost indicators is far less than the maximum, and the standard deviation is far greater than the mean, which is a typical positive skewness data distribution.
2) The average cost of blood, oxygen, surgery and delivery is very low, indicating that such patients account for a relatively low proportion.

2.2. Variable Combination
For the associated cost itemized indicators, after consulting the clinical experts, the original cost itemized indicators were merged as follows according to the professional knowledge:

1) The combination of Chinese herbal medicine and Chinese patent fee is collectively referred to as Chinese herbal medicine/Chinese patent fee.
2) Blood cost, oxygen cost, operation cost amalagmative, be called blood/oxygen/operation cost collectively.
3) Examination cost and cure cost are amalagmative, it is examination cure cost collectively.
4) Birth expenses are excluded.
5) Hospitalization costs were 0 yuan for 27 records and more than 50 thousand yuan for 13 records
6) One day was recorded in 157 cases, and more than 200 days in 12 cases.

In the management of real hospital inpatients, it is completely possible that the hospitalization cost is 0 yuan or the length of stay is 1 day, but such patients should not be the target group for analysis, and the corresponding records should not be included in the statistical analysis.

2.3. Investigation of disease distribution

The hospitalization cost is closely related to the type of disease. Frequency statistics are conducted for each disease type, and frequency functions in SPSS software are used. The results are shown in table 2:

| Disease Code | Frequency | The Effective Percentage | Cumulative Percentage |
|--------------|-----------|--------------------------|-----------------------|
| 1639         | 1629      | 43.0%                    | 1.6%                  |
| 8738         | 4018      | 275.0%                   | 10.2%                 |
| 4140         | 4140      | 596.0%                   | 22.1%                 |
| 4349         | 4349      | 318.0%                   | 45.6%                 |
| 4556         | 4556      | 127.0%                   | 50.3%                 |
| 4659         | 4659      | 391.0%                   | 64.8%                 |
| 4919         | 4919      | 361.0%                   | 78.2%                 |
| 5409         | 5409      | 322.0%                   | 90.1%                 |
| 7221         | 7221      | 175.0%                   | 96.6%                 |
| 8738         | 8738      | 92.0%                    | 100.0%                |
| Total        | 2700      |                          |                       |

As can be seen from table 2, the number of cases of diseases with codes of 1639 and 8738 was small, and the number of samples of the other 8 diseases 8 diseases were all sufficient.

2.4. Variable transformation

In the analysis, the total hospital expenses, the total number of days in hospital and so on are relatively important indicators, but these variables are obviously positively skewed distribution. In order to reduce the influence of extreme values, we conduct logarithmic transformation on them.

3. Introduction of BP neural network model

According to the results of previous data preprocessing, we used BP neural network method [3][4][5] to analyze the data and further evaluate the importance of each variable in the model, so as to determine the main influencing factors in the hospitalization cost.

3.1. BP neural network model

Artificial neural network (Ann) is an abstract mathematical model designed to reflect the structure and function of human brain. Since 1943, when American psychologist W. McCulloch and mathematician W. Pitts put forward the abstract mathematical model of formal neuron -- MP model, the theoretical technology of artificial neural network has undergone a tortuous development for more than 50 years. Especially in the 1980s, the research of artificial neural networks has made great progress, and relevant theories and methods have developed into an interdisciplinary subject between physics, mathematics, computer science and neurobiology. It in pattern recognition, image processing, intelligent control, combinatorial optimization, financial forecasting and management, communication, robots and expert system, and other fields has been widely used, more than 40 neural network model is put forward, which is famous with machine, the Hopfield net, Boltzman machine, the adaptive resonance theory and back propagation network (BP), etc[6].
BP (Back Propagation) network was proposed by a team of scientists led by Rumelhart and McClelland in 1986. It is a multi-layer feedforward network trained according to the error inverse Propagation algorithm and is one of the most widely used neural network models at present. BP network can learn and store a large number of input-output mode mapping relations without revealing the mathematical equation describing the mapping relations in advance. Its learning rule is to use the steepest descent method to constantly adjust the weight and threshold of the network through back propagation, so as to minimize the sum of squared errors of the network. BP neural network model topology includes input layer, hidden layer and output layer.

In the perceptron neural network model and linear neural network model learning algorithm, the difference between the ideal output and the actual output is used to estimate the neuron connection weight error. When the problem of linear indivisibility is solved and multilevel networks are introduced, it becomes a difficult problem to estimate the error of hidden layer neurons. In practice, it is impossible to know the ideal output value of any neuron in the hidden layer. In 1985, Rumelhart and McClelland put forward the BP network error back-propagation (BP) learning algorithm, which realized the multi-layer neural network model envisioned by Minsky [7].

The BP algorithm uses the error after output to estimate the error of the direct leading layer of the output layer, and then uses this error to estimate the error of the previous layer, so that the back propagation layer by layer can obtain the error estimation of all other layers. The multi-stage pulmonary circulation network which uses BP algorithm for learning is called BP network and belongs to the type of forward neural network. Although the accuracy of the error estimation itself will decrease with the "backward propagation" of the error itself, it still provides a more effective method for the training of multi-layer network. In addition, the multi-layer forward neural network can approximate any nonlinear function [8].

The neurons that make up the BP network are still neurons. According to the requirements of BP algorithm, the activation function used by these neurons must be differentiable everywhere. You always use an s-shaped function. For a neuron, its network input can be expressed as:

\[ net = x \omega \]

Where \( x \) represents the input received by the neuron, and \( \omega \) represents the connection weight corresponding to the neuron \( x \).

The output of this neuron is:

\[ y = f(net) = \frac{1}{1 + e^{-net}} \]

Further, we can find the derivative of \( y \) with respect to \( net \):

\[ f'(net) = \frac{e^{-net}}{(1 + e^{-net})^2} = \frac{1 + e^{-net} - 1}{(1 + e^{-net})^2} = (1 + e^{-net}) \left( 1 - \frac{1}{1 + e^{-net}} \right) = y(1 - y) \]

And: \( \lim_{net \to +\infty} \frac{1}{1 + e^{-net}} = 1; \lim_{net \to -\infty} \frac{1}{1 + e^{-net}} = 0 \)

According to the s-type activation function, the range of \( y \) is \((0, 1)\), so, the range of \( y \) is \((0, 0.25)\), and when \( y = 0.5 \), there is a maximum value. \( f'(net) f'(net) \)
Standard BP algorithm is a learning algorithm based on gradient descent method, the learning process is by cumulative value and threshold value, to delete the expectations and the actual output of the RBF network of minimum mean square error area and implementation, but it is only used to mean square error function of weights and threshold value of a derivative (gradient) of information, makes the algorithm exists defects such as slow convergence speed, easy to fall into local minimum.

Definition:
The input vector $\mathbf{x}$
Hidden layer input vector $\mathbf{h}_i$
Hidden layer output vector $\mathbf{h}_o$
Output layer input vector $\mathbf{y}_i$
The output vector $\mathbf{y}_o$
Expected output vector $\mathbf{d}$
The connection weight between the input layer and the hidden layer $\mathbf{w}_{ih}$
The connection weight between the hidden layer and the output layer $\mathbf{w}_{ho}$
The threshold of neurons in the hidden layer $\mathbf{b}_h$
The threshold value of each neuron in the output layer $\mathbf{b}_o$
Number of samples $k$
The activation function $f(\cdot)$

3.2. The steps of BP standard algorithm
The specific implementation steps of BP standard algorithm are as follows:
1) initialize the network, assign random Numbers within an interval (-1, 1), and set the error function as $e = \frac{1}{2} \sum_{a=1}^{q} (d_o(k) - y_o(k))^2$

Given the calculation accuracy value and the maximum learning times $M$, $\varepsilon$ 2) randomly select the KTH input sample and the corresponding expected output $\mathbf{x}(k)$ $\mathbf{d}(k)$ 3) calculate the input of neurons in the hidden layer, and then use the input and activation function to calculate the output of neurons in the hidden layer $\mathbf{h}_i(k)$ $\mathbf{h}_o(k)$

$$h_i(k) = \sum_i w_{ih} x_i(k) - b_h$$

$$h_o(k) = f(h_i(k))$$

$$y_i(k) = \sum_i w_{ho} h_i(k) - b_o$$

$$y_o(k) = f(y_i(k))$$

4) Use the actual output of the network expected output vector network to calculate the partial derivatives of the error function with respect to each neuron in the output layer $\delta_o(k)$
\[ \delta_h(k) = (d_o(k) - y_o(k))y_o(k)(1 - y_o(k)) \]

5) Calculate the partial derivative of the error function on each neuron in the hidden layer by using the connection weight from the hidden layer to the output layer, the output layer and the output of the hidden layer: \( \omega_{ho}(k) \) \( \delta_o(k) \) \( ho_o(k) \) \( \delta_h(k) \)

\[ \delta_h(k) = \left[ \sum_{o=1}^{O} \delta_o(k)\omega_{ho} \right] ho_o(k)(1 - ho_o(k)) \]

6) Use the output of neurons in the output layer and the output of neurons in the hidden layer to modify the connection weight and threshold: \( \delta_o(k) \) \( ho_o(k) \) \( \omega_{ho}(k) \) \( b_o(k) \)

\[ \omega^{N+1}_{ho}(k) = \omega^{N}_{ho}(k) + \eta \delta_o(k)ho_o(k) \]
\[ b^{N+1}_{o}(k) = b^{N}_{o}(k) + \eta \delta_o(k) \]

\( N \) is before adjustment, \( N+1 \) is after adjustment, vertical is learning rate, and value is between (0, 1).

7) Correct the connection weight and threshold by using the input of neurons in the hidden layer and the input of neurons in the input layer: \( \delta_i(k) \) \( x_i(k) \)

\[ \omega^{N+1}_{h}(k) = \omega^{N}_{h}(k) + \eta \delta_i(k)x_i(k) \]
\[ b^{N+1}_{h}(k) = b^{N}_{h}(k) + \eta \delta_i(k) \]

8) Calculate the global error \( E \)

\[ E = \frac{1}{2m} \sum_{k=1}^{m} \sum_{o=1}^{O} (d_o(k) - y'_o(k))^2 \]

9) Judge whether the network error meets the requirements. When the number of or learning is greater than the set maximum number \( M \), the algorithm ends. \( E < \varepsilon \) Otherwise, randomly select the next learning sample and the expected output of the corresponding year, return to the third step and enter the next round of learning process.

### 3.3. Considerations for using BP neural network

1) Interpretability

Neural networks are hard to explain, and there is no obvious way to explain them.

2) Overtraining problem

The neural network can fully extract sample information under the condition of sufficient nodes, but it cannot distinguish effective information from noise. Once the training sample is over-learned, more and more noise will be included, thus the network’s promotion ability will be reduced.

3) Sample size

Too low sample size will lead to unstable prediction effect and poor accuracy of extrapolation. Since net network can be regarded as the fitting of a nonlinear equation, the sample size is sufficient as long as the parameter estimation in the equation can be accurately estimated.

4) Analysis efficiency
The analysis efficiency of BP neural network depends on the relationship between training and learning. Training a neural network takes a considerable amount of time, of course, once the neural network is built, prediction and subsequent learning are very fast.

4. BP neural network software operation

At present, there are many software for BP neural network method to run and process, and the typical ones are Matlab and SPSS. Among them, the operation of SPSS is simple and the interpretation of results is convenient. Therefore, this paper adopts SPSS for calculation.

In order to prevent overfitting, it is necessary to split the sample data in the neural network method to form the analysis set, training set and verification set according to a certain proportion, such as 7:3 or 5:5 or other proportions. In this paper, the default ratio of 7:3 in SPSS software is adopted.

In the analysis, the network with single or double hidden layers can be established, and the number of hidden units in each layer can also be set by ourselves. In this paper, the default single-layer model of SPSS software is established first, and then the network is complicated to observe the follow-up analysis effect and whether the analysis result is improved.

The outcome at discharge (curative effect) was taken as the control factor, the length of stay was not controlled, the (total hospitalization cost) was taken as the dependent variable, the hospitalization code and other 26 factors were taken as the factors, the age was taken as the covariate, and the importance of independent variables was analyzed. The output results are shown in table 3, table 4, and table 5.

### Table 3. Summary of case processing

| The case number | The percentage |
|-----------------|----------------|
| sample training | 1859           |
|                 | 70.10%         |
| inspection      | 793            |
|                 | 29.90%         |
| effective       | 2652           |
|                 | 100.00%        |
| To rule out     | 48             |
| A total of      | 2700           |

As can be seen from table 3, among the 2,700 samples, 48 were excluded, including 1,859 training sets and 793 verification sets, accounting for 70.1% and 29.9%, respectively.

### Table 4. Model summary table

| training        | Sum of squares error | 267.333 |
|-----------------|----------------------|---------|
| The relative error |                     | 0.188   |
| Use the abort rule |                    |         |
| Training time   | 9 00:01 on Saturday  |         |
| (UK time). |
| inspection      | Sum of squares error | 129.719 |
| The relative error |                   | 0.144   |
| Dependent variable: |               |         |
| lg (total cost of hospitalization) | |       |
| A. Error calculation is based on test samples. | |         |
Table 4 shows that the relative error rate of both the training set and the verification set is about 14.4%, equivalent to the prediction accuracy of 85.6%, which is an excellent result.

As can be seen from table 5, the relative importance of age is the largest, indicating that with the increase of the age of patients, the hospitalization cost is also increasing. For China's aging country, this is a noteworthy information, which requires the country and government to issue corresponding measures and policies to deal with the aging trend. Secondly, the relative importance of the hospital level is as high as 75%.

Table 5. Importance table of independent variables

| Variable                                | The importance of | The importance of normalizing |
|-----------------------------------------|-------------------|-------------------------------|
| Hospital code                           | 0.087             | 62.20%                        |
| region                                  | 0.039             | 27.80%                        |
| level                                   | 0.106             | 75.50%                        |
| Discharge western medicine main diagnosis| 0.092             | 65.70%                        |
| year                                    | 0.021             | 15.30%                        |
| in                                      | 0.047             | 33.90%                        |
| gender                                  | 0.015             | 11.00%                        |
| Marital status                          | 0.04              | 28.30%                        |
| Whether the han nationality             | 0.013             | 9.50%                         |
| The hospital way                        | 0.041             | 29.60%                        |
| Results (western medicine)              | 0.019             | 13.40%                        |
| Exit diagnostic compliance sign         | 0.011             | 8.10%                         |
| In-out diagnostic conformance markers   | 0.017             | 12.40%                        |
| Presence of rescue                      | 0.016             | 11.10%                        |
| Treatment of categories                 | 0.02              | 14.40%                        |
| Drug allergy marker                     | 0.013             | 9.10%                         |
| Surgery sign                            | 0.038             | 26.80%                        |
| Nosocomial infection                    | 0.02              | 14.20%                        |
| Blood transfusion is                    | 0.015             | 10.60%                        |
| Infusion situation                      | 0.043             | 30.70%                        |
| Admission via                           | 0.025             | 17.70%                        |
| Before admission, he was treated by other hospitals | 0.011 | 7.90% |
| Illness during hospitalization: critical | 0.022 | 15.70% |
| Illness during hospitalization: acute      | 0.012 | 8.60%  |
| Illness during hospitalization: difficult | 0.018 | 13.20% |
| Sources of cost                         | 0.058             | 41.60%                        |
| age                                     | 0.125             | 93.10%                        |

and the relative importance of the hospital code is as high as 66.2%, which means that the same disease is diagnosed and treated in different levels and hospitals, and the costs vary greatly. The relative importance of the type of disease (discharge from the hospital mainly diagnosed by western doctors) is 65.7%, indicating that the hospitalization costs vary greatly among different diseases. For
diseases with high costs, the government should consider increasing the payment proportion of medical pooling. The important cost source is 41.6%, which tells us that for different payment sources, such as self-expense and public expense, the hospitalization cost will vary greatly even if the disease type, disease severity, treatment effect and other factors are completely the same.

Acknowledgments
This project is supported by the natural science foundation of Changzhou Institute of technology project "keep orthogonal mapping" (Project No.: E3-6107-17-039) and the teaching construction project of Changzhou Institute of technology project "Research on the cultivation of innovation ability of college students based on mathematical modeling" (Project No.: A3-4403-18-016).

References
[1] Deng Weibing, Zhou Yumin et al. SPSS23 statistical analysis using a tutorial electronic industry press 2017
[2] Zhang Wentong, Zhong Yunfei, IBM SPSS data analysis and mining of actual combat case essence, tsinghua university press, 2016
[3] Li Song, Liu Lijun, engineering theory practice of BP neural network optimization based on particle swarm optimization for short-term traffic flow prediction system, 2012 32 (9): 2045-2049.
[4] Zhao Cheng, Bai Xiu, MAO chunmei. ResourceS and Environment in the Yangtze BaSin, 2012 21(6)I
[5] application and implementation technology of BP artificial neural network by dong-chao long ma weiming et al. Journal of naval engineering university, 2000 93 (4)
[6] Yu Mingyan, Shi Yunbo et al. Quantitative identification of single gas sensor based on BP algorithm;Instrumentation technology and sensors; 08, 2011
[7] H. Kim and H. Melhem Damage detection of structures by complex analysis Engineering structures, 2004 26 (3) : 347-362.
[8] Bo Yan, Yao Cui Beam Structure Damage Identification Based on BP Neural Network and Support Vector Machine Mathematical Problems in Engineering 2014 (2) :1-8