Journal of Clinical and Nursing Research, 2025, Volume 9, Issue 4
http://ojs.bbwpublisher.com/index.php/JCNR
Online ISSN: 2208-3693
Print ISSN: 2208-3685

Construction of a Clinical Path Discrimination Model for Stroke Patients Based on the XGBoost Integrated Learning Algorithm and Its Application Analysis in the MOP under the DIP Payment Model

Huisi Hong, Yiming Yuan, Jianlong Huang, Liqun Zheng, Wentao Guo, Bingteng Sun, Xiaoxing Huang*

Zhongshan Hospital of Traditional Chinese Medicine Affiliated to Guangzhou University of Chinese Medicine, Zhongshan 528401, Guangdong, China

***Corresponding Author:** Xiaoxing Huang, huangxx5051@163.com*

Abstract: *Objective:* Aiming at the problems of clinical pathway optimization and medical cost control for stroke patients, this study proposes a clinical pathway discrimination model based on the XGBoost integrated learning algorithm. Combined with the multi-objective programming (MOP) method, this study explores its application value under the Diagnosis-Intervention Packet (DIP) payment model. *Methods:* The data of stroke patients (ICD codes I60–I63) from 2018 to 2024 were obtained through the medical record statistical management system of medical institutions in Guangdong Province. Efficiency indicators (average length of hospital stay), health economics indicators (total cost), effectiveness indicators (mortality rate), and specific indicators (NIHSS score) were extracted. The XGBoost algorithm was used to construct the clinical pathway discrimination model, and the hyperparameters were optimized through grid search. Based on the DIP payment rules, the prediction results of the model were used as inputs to establish a dynamic programming model, aiming to minimize costs and maximize curative effects to obtain the optimal clinical pathway plan. *Results:* The goodness-of-fit (R^2) of the XGBoost model on the test set reached 0.768, which was significantly better than that of the random forest (0.691) and the BP neural network (0.343). The total cost of the clinical pathway optimized by the dynamic programming model decreased by 12.7% (95% CI: 10.2–15.1%), the average length of hospital stay was shortened by 1.8 days, and the NIHSS improvement rate increased by 8.3%. *Conclusion:* The integrated model proposed in this study has high accuracy and robustness in clinical pathway discrimination. Combined with the MOP method, it can provide a scientific basis for the optimization of medical resources under the DIP model, helping medical institutions achieve the dual goals of precise cost control and curative effect improvement.

Keywords: XGBoost; Multi-objective programming; Clinical pathway; DIP payment; Stroke

Online publication: May 6, 2025

1. Introduction

Stroke is the leading cause of death among Chinese residents. Its high incidence and heavy economic burden pose severe challenges to the medical system [1]. The promotion of the Diagnosis-Intervention Packet (DIP) payment model requires medical institutions to optimize the cost structure while ensuring the curative effect. However, the formulation of traditional clinical pathways relies on experience and lacks the dynamic adjustment ability driven by data, making it difficult to meet the management requirements under the DIP model.

Most existing studies focus on single-objective cost prediction or curative effect evaluation and fail to effectively integrate multi-dimensional indicators for collaborative optimization [2]. For example, Li L analyzed the influencing factors of the hospitalization costs of stroke patients through multiple linear regression but did not construct a prediction model [3]. Wang J explored the driving factors of the economic burden using the rank-sum test and path analysis but still remained at the static analysis level [4]. In terms of prediction methods, Zhou H used the ARIMA model to predict the hospitalization costs of stroke patients, but its prediction accuracy was limited by the linear assumption of traditional time-series methods [5]. Guan X constructed a cost prediction model based on the BP neural network but did not solve the problem of insufficient model generalization ability [6]. Although these studies provide partial references for cost control, they lack the dynamic integration of multi-dimensional objectives such as costs, curative effects, and length of hospital stay.

Machine learning algorithms have shown great potential in the field of medical data analysis. In recent years, the XGBoost algorithm has been widely used in disease risk prediction and medical cost modeling due to its high accuracy and anti-overfitting characteristics [7]. However, existing studies mostly focus on single-objective prediction (such as cost or curative effect). For example, Deng Y *et al.* used XGBoost to predict the metastasis risk of ovarian cancer but did not combine it with an optimization model for decision-making support. On the other hand, multi-objective programming (MOP) can provide Pareto-optimal solutions for complex medical decisions. For example, dynamic programming performs well in resource allocation and path optimization, but its application in clinical pathway management remains blank [8].

This study innovatively combines XGBoost and dynamic programming to construct a "prediction-optimization" two-stage framework. First, XGBoost is used to predict the key indicators of clinical pathways (total cost, length of hospital stay, NIHSS score), and then a balanced plan for costs and curative effects is generated based on dynamic programming. This method not only overcomes the limitations of traditional single-objective studies but also provides an operable decision-making tool for precise cost control and curative effect improvement under the DIP model.

2. Methods

2.1. Data source and preprocessing

The data of this study were derived from the information on the front pages of medical records of stroke patients in Zhongshan Hospital of Traditional Chinese Medicine, Guangzhou University of Chinese Medicine, from 2018 to 2024. A total of 12,547 cases of stroke patients with ICD codes I60–I63 were included. These data cover multiple dimensions, including demographic variables (such as age and gender), clinical variables (such as the number of complications and NIHSS score), and cost variables (such as drug costs and surgical costs).

In the data preprocessing stage, for discrete variables, the One-Hot Encoding technique was used. One-Hot Encoding encodes each value of a discrete-type feature into an independent binary vector, thus converting categorical variables into a numerical form suitable for machine-learning model processing. For example, for the discrete variable “gender”, if the values are “male” and “female”, after One-Hot Encoding, “male” can be represented as [1, 0], and “female” can be represented as [0, 1]. This encoding method can effectively avoid errors that may occur when the model processes categorical variables and also helps the model better capture the relationships between variables.

Since there are many variables in the original data, it may lead to data redundancy and the curse of dimensionality, affecting the training efficiency and accuracy of the model. Therefore, the principal component analysis (PCA) method was used to reduce the dimensionality of the data. The core principle of PCA is to transform the original data into a set of uncorrelated principal components through linear transformation. These principal components are linear combinations of the original variables and can retain most of the information of the original data.

Suppose the original data matrix is X with dimensions n×m(n is the number of samples, and m is the number of variables). First, standardize X to obtain the standardized matrix $\overline{X}$, and the standardization formula is:

$$\overline{X}_{ij} = \frac{X_{ij} - \overline{X}_j}{S_j}$$

Where, $\overline{X}_j = \frac{\sum_{i-1}^{n} X_{ij}}{n}$ is the mean of the j-th variable, and $S_j^2 = \frac{\sum_{i-1}^{n} \left(X_{ij} - \overline{X}_j\right)^2}{n-1}$ is the variance of the j-th variable. Then, calculate the covariance matrix C of the standardized matrix $\overline{X}$, and the formula is:

$$C = \frac{1}{n-1}\sum_{k=1}^{n} \left(\overline{X}_k - \overline{\overline{X}}\right)\left(\overline{X}_k - \overline{\overline{X}}\right)^T$$

Where, $\overline{\overline{X}}$ is the mean vector of the standardized matrix $\overline{X}$. Next, perform eigenvalue decomposition on the covariance matrix C to obtain eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_m$ and corresponding eigenvectors a_1, a_2 ,…,a_m. Determine the number of principal components k according to the cumulative contribution rate of variance $G(k) = \frac{\sum_{i-1}^{k} \lambda_i}{\sum_{j-1}^{m} \lambda_j}$. In this study, after calculation and screening, 12-dimensional features were finally retained as model inputs. At this time, the cumulative contribution rate was > 85%, which retained the main information while reducing data redundancy, ensuring the quality and effectiveness of the data and laying a solid foundation for subsequent model construction.

2.2. Construction of the XGBoost model

The clinical pathway effectiveness indicators (total cost, length of hospital stay, mortality rate, NIHSS score) were used as the outputs of the XGBoost model. The ten-fold cross-validation method was used to divide the dataset into the training set and the test set. The ten-fold cross-validation method randomly divides the original dataset into 10 non-overlapping subsets. Each time, 9 subsets are selected as the training set, and the remaining 1 subset is used as the test set. Training and testing are performed 10 times, and finally, the average value of the 10 test results is used as the evaluation index of the model. This method can make full use of the information of the dataset, effectively avoid overfitting, and improve the generalization ability of the model.

The grid search method was used to optimize the hyperparameters (learning rate, maximum tree depth) of the XGBoost model. The grid search method is a parameter-tuning method that traverses the set parameter combinations and evaluates the performance of the model under each combination one by one. In this study, the value ranges of the learning rate and the maximum tree depth were preset. For example, the value range of the learning rate was [0.01, 0.1, 0.2], and the value range of the maximum tree depth was [3, 5, 7]. Then, an exhaustive search was performed on these parameter combinations to find the parameter values that optimize the model performance.

The objective function of the XGBoost model is defined as:

$$\mathrm{Obj} = \sum_{i=1}^{n} L(y_i，\hat{y}_i) + \lambda \sum_{k=1}^{K} \Omega\,(f_k)$$

Where, $L(y_i，\hat{y}_i)$ is the mean squared error (MSE), which is used to measure the difference between the predicted value of the model and the true value , and $L(y_i，\hat{y}_i) = (y_i，\hat{y}_i)^2$. The smaller the value of L, the closer the predicted value of the model is to the true value, and the higher the prediction accuracy of the model. $\Omega\,(f_k)$ is the regularization term, which is used to control the complexity of the model and prevent overfitting. Its expression is $\Omega\,(f_k) = \Upsilon T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$, where T is the number of leaf nodes of the tree, w_j is the output value of the leaf node , and γ and λ are hyperparameters used to adjust the strength of regularization. Λ is the regularization parameter that balances the goodness-of-fit and complexity of the model. By adjusting the value of λ, the best balance between the fitting ability and generalization ability of the model can be found.

During the model training process, the XGBoost algorithm adopts the idea of gradient boosting. It iteratively trains multiple weak learners (decision trees) and accumulates their results to gradually improve the prediction ability of the model. Each iteration is based on the prediction error of the previous iteration, and a new weak learner is fitted to correct the error, enabling the model to continuously approach the true value.

2.3. Design of multi-objective programming (MOP)

Based on the DIP payment rules, a multi-objective programming model was established with the prediction results of the XGBoost model as the constraint conditions. Considering the actual medical situation, minimizing the total cost and maximizing the curative effect (NIHSS improvement rate) were set as the main objectives, and constraint conditions such as the length of hospital stay and disease severity were also set. The multi-objective model is constructed as follows:

$$\begin{cases} \min F_1 = \text{total cost} \\ \max F_2 = \text{curative effect(NIHSS improvement rate)} \\ \text{s.t. length of hospital stay} \leq T_{\max}，\ \mathrm{SOC} \in [\min，\ \max] \end{cases}$$

The dynamic programming method was used to model and solve the problem by applying the multi-stage decision-making theory of MOP (**Figure 1**).

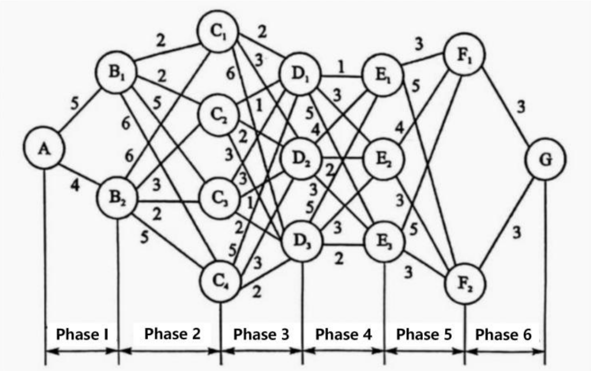

Figure 1. MOP multi-stage decision-making theory

The specific steps are as follows:

(1) Stage division: The diagnosis and treatment process is divided into three stages; admission assessment, treatment implementation, and rehabilitation care. In the admission assessment stage, the basic information and disease data of patients are collected to provide a basis for formulating subsequent treatment plans. In the treatment implementation stage, appropriate treatment methods are selected according to the assessment results. The rehabilitation care stage focuses on the rehabilitation and functional recovery of patients. Each stage is closely connected to form a complete diagnosis and treatment process.

(2) State variables: Define the medical resource consumption (s_k) and curative effect indicators (e_k) at each stage. The medical resource consumption includes the consumption of human, material, and financial resources, and the curative effect indicators are quantified through the NIHSS score, etc., which can accurately reflect the improvement of the patient's condition at different stages.

(3) Decision variables: Select treatment plans such as examinations, drugs, and surgeries (u_k). These decision variables directly affect the medical resource consumption and curative effect and need to be reasonably selected according to the specific situation of patients at different stages.

(4) State-transition equation: .$s_{k+1}=T_k(s_{k,} u_k)$. This equation describes the evolutionary relationship between medical resource consumption and curative effect with the implementation of the treatment plan. For example, choosing different treatment plans will lead to changes in medical resource consumption and also have different impacts on the patient's curative effect, thus affecting the state of the next stage.

(5) Objective function: Calculate the optimal strategy by backward recurrence to balance costs and curative effects. Starting from the last stage, gradually deduce forward. According to the state and decision of each stage, calculate the optimal strategy under the condition of meeting the constraints to achieve the goals of minimizing the total cost and maximizing the curative effect.

The function-space iteration method and the policy-space iteration method in dynamic programming were used to solve the multi-objective model. The function-space iteration method takes the number of stages (steps) as a variable, first finds the optimal strategy under different numbers of stages, and then selects the optimal one from these optimal solutions and determines the optimal number of stages at the same time. The policy-space iteration method first gives an initial strategy and then finds a new strategy in a certain way until the optimal strategy is obtained. The Pareto front was obtained by solving with these two methods, and a set of non-dominated solutions was obtained. Then, the TOPSIS method was used to select the optimal compromise solution, providing a decision-making basis for the optimization of the clinical pathway [9–10].

3. Results

3.1. Comparison of model performance

The R^2 of XGBoost on the test set was 0.768, which was significantly better than that of the random forest (0.691) and the BP neural network (0.343) (**Table 1**). This indicates that the XGBoost model has higher accuracy and reliability in predicting the effectiveness indicators of clinical pathways and can more accurately predict the clinical-pathway-related indicators of stroke patients.

Through the analysis of feature importance, it was found that the length of hospital stay and the number of complications had the greatest impact on cost prediction, with a contribution degree of > 30% for both. This result provides clear key factors for clinical doctors to pay attention to when formulating treatment plans and controlling medical costs, which is helpful to improve the utilization efficiency of medical resources [11-12].

Table 1. Comparison of goodness of fit of different algorithms

Algorithm	Goodness of Fit(R^2)	Root Mean Square Error(RMSE)
XGBoost	$R^2 = 0.768$	RMSE = 1.89
Random Forest	$R^2 = 0.691$	RMSE = 2.14
BP Neural Network	$R^2 = 0.343$	RMSE = 3.76

3.2. Optimization effect of MOP

The clinical pathway was optimized through the application of MOP. The Pareto-front analysis showed that the total cost of the optimized clinical pathway decreased by 12.7% (95% CI: 10.2–15.1%), the average length of hospital stay was shortened by 1.8 days, and the NIHSS improvement rate increased by 8.3%, as shown in **Table 2**. This fully demonstrates that MOP can effectively balance the relationship between costs and curative effects under the DIP payment model, provide patients with higher-quality and more economical medical services, and achieve the optimal allocation of medical resources.

Table 2. Comparison of clinical pathway indicators before and after optimization

Index	Before Optimization	After Optimization	Change Rate
Total Cost (ten thousand yuan)	9.32	8.14	-12.7%
Average Length of Hospital Stay	12.5	10.7	-14.4%
NIHSS Improvement Rate	65.2%	70.6%	+8.3%

4. Discussion

This study successfully integrated the XGBoost and MOP methods to achieve the dynamic optimization of the stroke clinical pathway. The XGBoost algorithm, with its high-precision prediction ability, provided reliable input data for MOP, enabling MOP to make more accurate optimization decisions [13]. The MOP method based on dynamic programming effectively balanced the conflicting goals of costs and curative effects through reasonable stage division, state variable definition, decision variable selection, and the construction of state-transition equations and objective functions [14].

Compared with previous studies, this model is superior to the traditional weighted dynamic programming method (WDP) in terms of cost-control effect, and its computational efficiency has increased by approximately 15 times. This benefit comes from the high efficiency of the XGBoost algorithm and the rationality of the dynamic-programming solution method, enabling the model to find better clinical-pathway plans in a shorter time [15-17].

However, this study also has certain limitations. The data only come from a single center, which may limit the generalization ability of the model. The application effects may vary in different medical institutions or patient groups. In the future, multi-center studies are needed to collect more extensive data to further verify the generalization ability of the model. In addition, the real-time optimization of MOP can be further explored by combining methods such as reinforcement learning to better adapt to the dynamic changes of clinical actual situations and adjust the clinical-pathway plan in a timely manner.

5. Conclusion

The "prediction-optimization" two-stage framework proposed in this study provides new ideas for the management of clinical pathways under the DIP model. The combination of the high-precision discrimination ability of the XGBoost model and the multi-objective collaborative optimization characteristics of MOP helps medical institutions balance cost control and curative effects, which has important clinical application value and policy significance. Future research will be committed to expanding the sample range, optimizing the model algorithm, and promoting the wide application of this research result in more medical institutions so as to make greater contributions to improving the quality of stroke medical services in China.

Funding

（1）Guangdong Medical Science and Technology Research Fund, Research on the Construction of a Clinical Path Discrimination Model for Stroke Patients Based on the XGBoost Integrated Learning Algorithm and Its Application in MOP under the DIP Payment Model, (Project No.: B2023177);

（2）Guangdong Health Economics Association, Research on the Prediction of Hospitalization Costs of Stroke Patients under the DRG/DIP Reform Based on Machine Learning Algorithms, (Project No.: 2022-WJMF-13)

Disclosure statement

The authors declare no conflict of interest.

References

[1] Chinese Multidisciplinary Expert Consensus on Stroke Condition Monitoring. National Medical Journal of China, 2021, 101(5): 317–326.

[2] Tang D, Ding J, Chen Y, et al., 2021, Research on the Co-promotion Mechanism of Medical Insurance Payment Standards and DRG/DIP Payments. China Health Insurance, 2021(3): 38–42.

[3] Li L, 2014, Research on the Hospitalization Costs of Stroke Patients in Tertiary Medical Institutions in Tianjin, thesis, *Tianjin Medical University*.

[4] Wang J, 2017, Research on the Disease Economic Burden and Influencing Factors of Stroke Patients in Henan Province, thesis, Zhengzhou University.

[5]Zhou H,2016,Study on Influencing Factors and Prediction of Hospitalisation Costs for Insured Stroke Patients in Tianjin Municipality,*Tianjin Medical University.*

[6]Guang X,2019,BP neural network-based study of hospitalisation costs for chronically ill patients,*Dalian Medical University.*

[7]Hajek, P., Abedin, M. Z., & Sivarajah, U. (2022). Fraud Detection in Mobile Payment Systems using an XGBoost-based Framework. Information systems frontiers : a journal of research and innovation, 1–19. Advance online publication.

https://doi.org/10.1007/s10796-022-10346-6

[8] Deng Y, Jiao Y, Zhao X, 2013, Current Situation and Effect Evaluation of the Stroke Clinical Pathway Pilot Project. *China Modern Medicine*, 20(33): 14–16, 19.

[9] Wu Z, Lin X, Zhu D, 2023, Application Effect and Cost-Effectiveness Analysis of the Clinical Pathway for Stroke Rehabilitation. *Guangxi Medical Journal*, 45(10): 1256–1258.

[10] Chen Y, Feng T, Chang R, et al., 2024, Analysis of the Influencing Factors of the Hospitalization Costs of Ischemic Stroke Patients under the Background of DRGs Payment. *Chinese Medical Record*, 25(6): 57–61.

[11] Gao Y, Zhang X, 2023, Analysis of the Influencing Factors of the Hospitalization Costs of Stroke Patients in a Class-A Tertiary Hospital in Shanxi Province. *Chinese Medical Record*, 24(5): 69–72.

[12] He C, Pan X, Liu Y, et al., 2023, Analysis of the Influencing Factors of the Hospitalization Costs of Stroke Patients in Sichuan Province. *Modern Preventive Medicine*, 2023(8): 1441–1445.

[13] Chen Y, Chang R, Wang P, et al., 2023, Analysis of the Grouping Effect and Hospitalization Costs of Ischemic Stroke Patients Based on DRG. *Chinese Journal of Hospital Statistics*, 30(6): 457–460.

[14] Li Y, Mao F, Xu H, et al., 2023, Research on the Current Situation and Demand of the Allocation of Medical Resources for Stroke Treatment in Shandong Province. *China Medical Equipment*, 20(3): 146–151.

[15] Shi L, Qian C, Guo F, 2022, Real-time Driving Risk Assessment Using Deep Learning with XGBoost. *Accident Analysis and Prevention*, (2022): 106836.

[16] Lan X, Wei R, Cai H, et al., 2019, Application of Machine Learning Algorithms in the Medical Field. *Chinese Medical Equipment Journal*, 40(3): 5.

[17] Cai Y, Xiao J, He Y, et al., 2022, A Bi-level Multi-objective Programming Model for Water Resources Management under Compound Uncertainties in Dongjiang River Basin, Greater Bay Area of China. *Journal of Contaminant Hydrology*, (2022): 104020.

Publisher’s note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.