ICIC Express Letters ICIC International ©2016 ISSN 1881-803X
Volume **10**, Number **11**, November **2016** pp. **2601–2606**

LOGISTIC REGRESSION MODELING METHOD FOR THE OPTIONAL TIME EVALUATION OF THE INITIATION OF HEMODIALYSIS BASED ON DATA MINING

Jinyu Song[1,2], Xudong Song[1], Zhanzhi Qiu[1] and Wei Wang[1]

[1]Software Institute
Dalian Jiaotong University
No. 216, Xingfa Road, Dalian 116052, P. R. China
xudongsong@126.com

[2]Dalian No. 8 High School
No. 12, Fushun Street, Dalian 116021, P. R. China
1065711627@qq.com

Received March 2016; accepted June 2016

Abstract. *The optional time evaluation of the initiation of hemodialysis has a key effect on survival quality of life in patients with stage V chronic kidney disease. How to determine the optimal time of the initiation of the hemodialysis has always been a controversial issue in the world. In this paper, a Logistic regression modeling method is provided for the optional time evaluation of the initiation of hemodialysis based on data mining. By extracting 113 patients' clinical data between January 2008 and October 2012 from a blood purification center in China, using correlation analysis, decision tree and association rule mining algorithm, find out key factors of the optional time of the initiation of hemodialysis, and establish a Logistic regression model for the optional time evaluation of the initiation of hemodialysis by WEKA data mining software. The model shows good accuracy and availability.*

Keywords: Chronic kidney disease, Initiation of hemodialysis, Data mining, Logistic regression model, WEKA

1. **Introduction.** Uremia is a serious disease which threats human health, and is the top 1 in 8 kinds of common major diseases in China. It is expected to have more than 1 million uremic patients in China. Blood purification is the main clinical treatment measure. How to determine the optimal time of the initiation of the hemodialysis has always been a controversial issue in the world. So far there is no uniform standard [1-3]. In the past years, in order to estimate glomerular filtration rate (GFR) and determine the optional time of the initiation of hemodialysis for patients with stage V chronic kidney disease, many Modification of Diet in Renal Disease (MDRD) equations were established by using multivariable statistical regression method in China and abroad. Due to the different serum creatinine measurements and also muscle's capacity, there will produce a certain error for evaluation of GFR, and will lead to evaluation bias for the optional time of the initiation of hemodialysis. In this paper, the survival quality life in patients with stage V chronic kidney disease is regarded as modeling goal, and we put forward a Logistic regression modeling method [4-6] for the optional time evaluation of the initiation of hemodialysis based on data mining. By using correlation analysis, decision tree and association rule mining algorithm, find out key factors of the optional time of the initiation of hemodialysis, and establish the optional time evaluation Logistic regression model by WEKA data mining software. As an example, we extract 113 patients' clinical data between January 2008 and October 2012 from a blood purification center in China, and establish a Logistic regression model for the optional time evaluation of the initiation of hemodialysis. The model shows good accuracy and availability. The Logistic regression

modeling method provided by this paper has a certain reference value for solving similar medical problems. The paper is organized as follows. In Section 2, we briefly review the logic regression model and data mining technology, including decision tree and association rule. In Section 3, we give Logistic regression modeling method and processing description for the optional time evaluation of the initiation of hemodialysis based on data mining. In Section 4, we extract 113 patients' clinical data from a blood purification center in China and give the specific modeling practice process. The conclusion of this paper is given in Section 5.

2. **Logistic Regression Model and Data Mining Technology.**

2.1. **Logistic regression model.** In medical research, we usually want to figure out the relationship between a dependent variable y and a set of m independent variables $x_1, x_2, x_3, \ldots, x_m$. If y is a logic variable (corresponding to a classification of two events), we use p to represent the probability of events. When the event occurs, $y = 1$; otherwise $y = 0$. We use $p(y = 1)$ to represent the probability of occurrence events, and $p(y = 0)$ to represent the probability of events which do not occur. Here $p(y = 1) = 1 - p(y = 0)$, and $p \in [0, 1]$. The Logistic regression model p is defined as follows.

$$p = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_mx_m}}{1 + e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_mx_m}} \tag{1}$$

or

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_mx_m \tag{2}$$

2.2. **Data mining technology.** Data mining is a process of discovering knowledge or patterns from a large amount of data. The commonly used data mining techniques include association, classification and clustering. This paper mainly used association analysis and decision tree classification data mining technology to discover key factors of blood purification time.

Association analysis is a method for discovering interesting association rules between variables. An association rule is defined as an implication of the form: $X \rightarrow Y$, where X is called rule antecedent and Y is called rule consequent. We often use Support$(X \rightarrow Y) = P(X \cup Y)$ and Confidence$(X \rightarrow Y) = P(Y|X)$ to discover interesting rules from the set of all possible rules.

Decision tree is a flowchart-like tree structure in which each internal node represents a test on an attribute, each branch represents an outcome of the test, and each leaf node represents classification label. Usually we select the attributes with bigger information $Gain(A)$ as key factors. Information $Gain(A)$ is defined as follows.

$$Gain(A) = Info(D) - Info_A(D) \tag{3}$$

In the above formula, $Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$, and p_i is the probability that events which record in dataset D set belong to the class label i. $Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j)$, and $Info_A(D)$ is expected information which classifies dataset D according to the attributes A.

3. **Logistic Regression Modeling Method for the Optional Time Evaluation of the Initiation of Hemodialysis Based on Data Mining.** In this paper, association analysis and decision tree data mining techniques [7] are integrated into Logistic regression modeling method process for the optional time evaluation of the initiation of hemodialysis. The whole modeling process is shown in Figure 1, and it is divided into 4 parts: data collection, data preprocessing, key factors mining, and logic regression modeling.

In the stage of data collection, select the hospital blood purification information registration system as the data source, adopt patients' selection inclusion criteria and exclusion

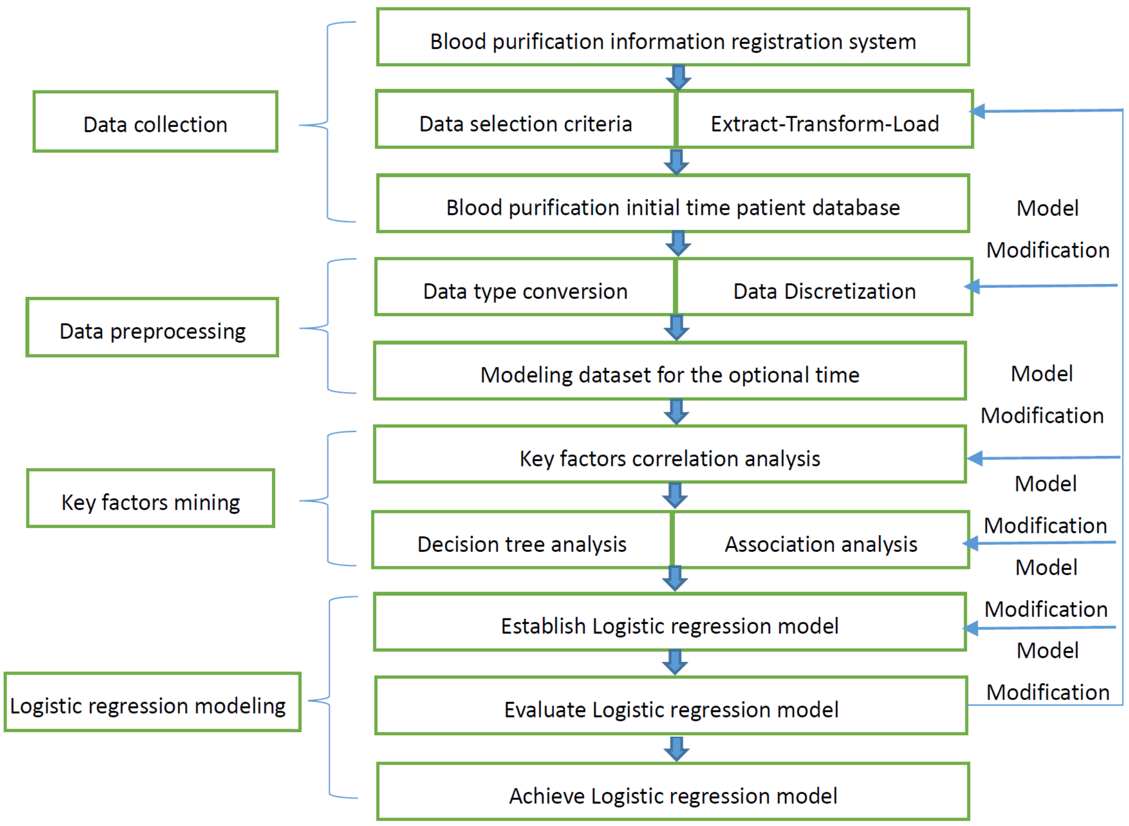
Data collection
Blood purification information registration system
Data selection criteria
Extract-Transform-Load
Blood purification initial time patient database
Data preprocessing
Data type conversion
Data Discretization
Modeling dataset for the optional time
Key factors mining
Key factors correlation analysis
Decision tree analysis
Association analysis
Logistic regression modeling
Establish Logistic regression model
Evaluate Logistic regression model
Achieve Logistic regression model
Model Modification
Model Modification
Model Modification
Model Modification
Model Modification

Figure 1. Logistic regression modeling process for the optional time evaluation

criteria defined by medical experts, and collect the required data into blood purification initial time patients database by data extract-transform-load (ETL) operations and open database connectivity data interface.

In the stage of data preprocessing, complete data preprocessing and give the definition of the modeling target variable to facilitate subsequent data mining and model establishment, identify error data by metadata and deal with the problems to ensure data integrity, and provide data type conversion and data discretization functions to implement easy variable association analysis in the following stage.

In the stage of key factors mining, select key factors for the optional time evaluation of the initiation of hemodialysis through the correlation analysis methods to ensure that the logistic regression modeling variables satisfy the constraints of correlation, and analyze the interaction relationship between variables by the methods of decision tree and association analysis to ensure that the variables satisfy the constraint of independence. If there is an interaction relationship between the two variables, the two variables can be combined to deal with the interaction relationship. If there is an interaction relationship between multiple variables, the multiple models method can be used to deal with the interaction relationship.

In the stage of establishment of Logistic regression modeling, build Logistic regression modeling for the optional time evaluation of the initiation of hemodialysis by modeling tool, analyze the model's accuracy, evaluate and validate the model. If the model does not reach the modeling goal, return to the previous modeling phase and do model modification; if it reaches the modeling goal, achieve the optional time evaluation model of the initiation of hemodialysis.

4. **Logistic Regression Modeling Case for the Optional Time Evaluation of the Initiation of Hemodialysis.** The data of all patients is obtained from a blood

purification center in China. We define the following patients' selection inclusion criteria and exclusion criteria.

Inclusion criteria: maintenance hemodialysis patients; age ranged from 18 to 75 years old, and gender is not limited; regular dialysis three times per week.

Exclusion criteria: peritoneal dialysis or kidney transplant patients; incomplete data; initiation hemodialysis patients with the malignant tumor, chronic infection, or cirrhosis of the liver; the accidental death patients caused by traffic accidents.

We define the data extraction rule is the initiation hemodialysis patients between January 2008 and October 2012, and prognosis observation time is October 2015. Data extraction variables include gender, date of birth, initiation time of hemodialysis, initiation vascular access, initiation hemodialysis symptoms (heart failure, nausea and vomiting, edema, diabetic nephropathy, and uremic encephalopathy), initiation hemodialysis laboratory indicator (hemoglobin, blood albumin, creatinine, urea, uric acid, potassium, phosphorus, parathyroid, and total carbon dioxide), and survival time. The total extraction modeling patient data is 113.

We define the effect of hemodialysis as the goal variable, use "1" to represent good effect according to survival time more than 3 years, and use "0" to represent bad effect according to survival time less than 3 years. There are 102 good effect patients and 11 bad effect patients in the modeling dataset. Gender variable and initiation vascular access variable are converted to logic type. We calculate patients' age at initiation hemodialysis by time of initiation hemodialysis minus year of birth. Because pH variable has nearly 51% missing value, we delete this variable.

By using WEKA data mining platform, we calculate Pearson correlation coefficients for the independent variables and the goal variable, and obtain the ranking results. It can be seen from the results that the independent variables have a certain correlation with the goal variable, but the correlation is not strong.

By using WEKA data mining platform, we call the J48 decision tree algorithm and obtain a decision tree. From the decision tree we can see the age variable as the first key factor, gender and edema as the other key factors. After the classification by using the first key factor age, we can see the patients whose age is less than 75 years have a very high proportion (92%), so there is no need to consider the relationship between the age variable and other variables.

By using WEKA data mining platform, we call discretize a pretreatment method to discretize the continuous variables, and call the Apriori algorithm to discover association rules by setting the minimum support 10% and the minimum confidence 90%. Through the analysis of association rules, we find that the key factors and goal variable have higher association and keep better independence between these key factors.

After the completion of the correlation analysis, based on WEKA data mining platform, we call SimpleLogistic algorithm, implement Logistic regression modeling for the optional time evaluation of the initiation of hemodialysis, and obtain the following model:

$$p = \frac{1}{1 + e^{-(4.63+0.87*[gender]-0.05*[age]+0.86*[iniVascularAccess]-0.3*[heartFailure]+0.32*[edema])}} \tag{4}$$

The model's accuracy is: 91.15%, and the error rate is 8.85%. In order to facilitate observation, we use Mathemetica software to plot the hemodialysis effect probability distribution. In Figure 2 we illustrate the hemodialysis effect probability distribution with age at initiation time between male and female for the patients of having edema but not heart failure and taking internal arteriovenous fistula at initiation time. In Figure 3 we illustrate the hemodialysis effect probability distribution with age at initiation time and edema for the patients of taking internal arteriovenous fistula at initiation time and without heart failure.

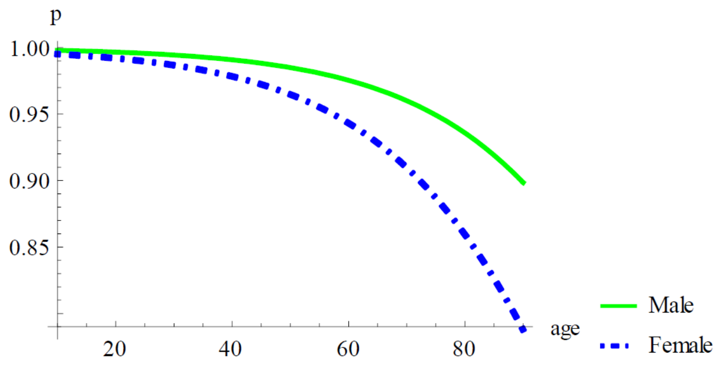

Figure 2. Hemodialysis effect probability distribution with age between male and female

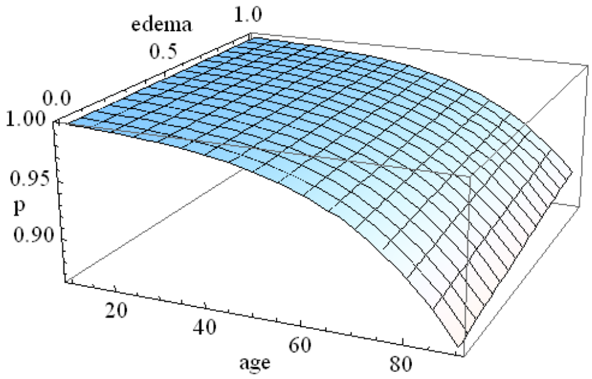

Figure 3. Hemodialysis effect probability distribution with age and edema

Using the above Logistic regression model, we can evaluate the effect of the initiation of hemodialysis. We can see that the greater age at initiation hemodialysis, the worse the effect of hemodialysis, and the hemodialysis effect of male is better than that of women. We also find that for the initiation time of hemodialysis, the effect of internal arteriovenous fistula is better than that of the central venous catheter, and the effect of patients with heart failure is worse than that without heart failure. These conclusions are consistent with the results of clinical medicine diagnosis. At the same time, we also find an interesting result: the effect of patients with edema is better than that without edema for the initiation time of hemodialysis, which seems contrary to normal sense. Although no edema is good for persons without diseases, the edema is the clinical manifestations for uremic patients, and hemodialysis treatment can improve patients' health of the body, so the effect of patients with edema is better. Also, we find the evaluation model is not associated with patients' creatinine which reflects glomerular filtration rate. The result is consistent with Cooper's conclusion [3] that early dialysis (estimated glomerular filtration rate is less than 14ml per minute) is not associated with an improvement in survival published in the New England Journal.

5. **Conclusion.** Aiming at the problem of the optional time evaluation of the initiation of hemodialysis, this paper provide a Logistic regression modeling method for the optional time evaluation of the initiation of hemodialysis based on data mining. Based on the method, use clinical dialysis medical data from a blood purification center in China to establish the optional time evaluation of the initiation of hemodialysis logic regression model for patients with stage V chronic kidney disease. Model results show that optional time of the initiation of hemodialysis is closely related to age, gender, initiation vascular access, circulatory system heart failure, and urinary system edema at initiation time of

hemodialysis. Not only have the model results provided by this paper a guiding role for clinical medical practice, but also the proposed modeling method has a certain reference value for solving the similar medical problems. In the future, we are ready to study Logistic regression modeling method for the optional time evaluation of the initiation of hemodialysis under big data medical environment, and to lay the foundation for the implementation of the logical regression model of big data medical problem in the cloud platform.

Acknowledgment. This work has been supported by Liaoning Provincial Natural Science Foundation of China (No. 201602131), and Dalian Science and Technology Project of China (No. 2014A11GX006).

REFERENCES

[1] S. Wright, D. Klausner, B. Baird et al., Timing of dialysis initiation and survival in ESRD, *Clinical Journal of the American Society of Nephrology Cjasn*, vol.5, no.10, pp.1828-1835, 2010.

[2] W. F. Clark, Y. Na, S. J. Rosansky et al., Association between estimated glomerular filtration rate at initiation of dialysis and mortality, *Canadian Medical Association Journal*, vol.183, no.1, pp.47-53, 2011.

[3] B. A. Cooper, P. Branley, L. Bulfone et al., A randomized, controlled trial of early versus late initiation of dialysis, *New England Journal of Medicine*, vol.363, no.7, pp.609-619, 2010.

[4] N. Dikaios, J. Alkalbani, H. S. Sidhu et al., Logistic regression model for diagnosis of transition zone prostate cancer on multi-parametric MRI, *European Radiology*, vol.25, no.2, pp.523-532, 2015.

[5] L.-C. Chen and L.-H. Chen, The score-type goodness-of-fit test for logistic regression models under stratified choice based sampling, *ICIC Express Letters, Part B: Applications*, vol.4, no.3, pp.603-607, 2013.

[6] H. Matsui, Variable and boundary selection for functional data via multiclass logistic regression modeling, *Computational Statistics & Data Analysis*, vol.78, no.5, pp.176-185, 2014.

[7] X. Song, Z. Qiu and J. Mu, Study on data mining technology and its application for renal failure hemodialysis medical field, *International Journal of Advancements in Computing Technology*, vol.4, no.3, pp.223-230, 2012.