**International Journal of Management, IT & Engineering**
Vol. 8 Issue 8, August 2018,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

# Using Classification and Regression Tree Techniques for Predicting Length of Stay of Diabetes Patients

**Mr.C.Natarajan M.E., (Ph.D)***

**Dr.J.M.Gnanasekar M.E., Ph.D****

**Mrs.Janorious Hermia M.E*****

## Abstract

Healthcare resource management is the imperative approaches to ensure the effective and efficient healthcare delivery to the patients. To provide the optimal utilization of hospital Resources, systematically form the healthcare resources utilization patterns which include resource planning, allocation, and management of medical needs. To provide the healthcare resource management plan as scrupulously as possible, prediction the length of stay of patients in a hospital is important in providing them with enhanced services and higher satisfaction. The proposed research shows that the classification and regression trees technique is the best tool to predict the patient's length of stay in the hospital. A medical record of the patients contains a huge number of information associated to patient conditions along with treatments and actions received. Healthcare Resource Utilization analysis based on such recorded data collected through regular track of treatment and carried out in an organized manner can be leveraged to get better treatments in several ways. This proposed method uses the classification techniques such as Support vector machine, neural network classifier and ensemble classification and Regression Tree to find the Length Of Stay for the diabetes patients. It shows the differences in the above three classification techniques. The results verified that the Classification and Regression Tree technique gives more accuracy than the other.



***Keywords:***

Support Vector Machine;
Classification;
Regression Tree;
Neural Network;
Resource Utilization;
Datamining.
.

***Author correspondence:***

C.Natarajan,
Ph.D Scholar,Department of Computer Science and Engineering,
Saveetha University, Chennai, Tamilnadu, India.
Email: vcnataraj@gmail.com

* Ph.D Scholar, Department of Computer Science and Engineering, Saveetha University, Chennai, India.

** Professor, Department of Computer Science and Engineering,Sri Venkateswara College of Engineering,India.

*** Assistant professor,Department of Computer Science and Engineering,Tamilnadu,India

## 1. Introduction

Healthcare industry engenders a huge amount of information in the form of patient's health records, disease and diagnoses of their, planning and managing the resources of hospital. To help the Hospital Resource management and its administration along with the cost savings and efficient delivery, though extracting useful information from such large datasets is not sufficient. There are so many Research has been done and applying different data mining techniques and algorithms to extract useful information and associative patterns from the outsized datasets. Data mining algorithms with most recent techniques and methodologies that can cleverly aid us in converting these enormous amounts of data into valuable information and knowledge. This extracted useful information will give extra resource of knowledge for the doctors and hospital's resource management system, for delivering effective treatments and providing cost effective healthcare services to the patients.Additionaly, those information can also be valuable in predicting patient's length of stay in the hospital that helps to evade needless usage of healthcare resources. This research proposal is to formulate a result that is competent of reducing the unnecessary usages of resources by predicting the length of patients stay in hospital. Modern healthcare institute are investing lot of resources in data mining techniques and still they are searching the better solution. This research enables the institute not only to manage efficiently but also save lot of resources.

Diabetes disease is major causes of other endanger diseases in developed countries resulting in numerous sickness, disabilities, and deaths as well. The Diabetes patient's characters regarding their length of stay (LOS) should be noted. LOS is the total number of days for the patient is undergoing treatments in a hospital or giving treatments in a similar medical facility. To identify the healthcare resource utilization, costs for the treatments, and to know the serious level of the diseases, the length of the hospital stay will be the key parameter. The LOS is used for the predictive measure of inpatient costs and highly notifies the utilization of healthcare resources.

Limited number of beds available to hold the inpatients in the hospital, and most of hospitals are facing considerable financial pressure, it is extremely important to find ways to reduce health care costs. By using data mining techniques and algorithms, predicting and finding out the discharge date and LOS of each patient will be the solution to reduce the healthcare costs. The successful consideration for a hospital administration is to predict and estimate LOS data is backbreaking process, but it is an important obsession to exact the prediction of LOS facilitates the effectiveness of resource management in hospitals. Therefore, the accurate and suitable prediction of LOS has become increasingly important for hospital management and health care systems .The awareness of factors and essentials that establish LOS could promote the advance of resourceful medical pathways and optimize resource utilization and management. Addition to that, maximum number of hospitals in the world cannot be able to predict and calculate the expectations of the outlook admission requirements. The successful prediction of discharge dates and duration of hospital stay allows the equivalent preparation of possible admissions, primary to reduce the discrepancy in resource utilization. To provide a well-organized and exact model to predict LOS for different types of diseases is one of the big tasks for the researchers. Developing models for predicting and formative LOS in hospitals can be very useful for hospital resource management, especially for scheduling the health care policies and promoting health care services, utilizing and managing the proper allotment of health care resources according to patient's diseases, LOS and patients' physical condition status. Improved prediction strategies are needed to smooth the progress of the decision making process. To improve the healthcare resource utilization profile, an efficient and accurate method to predict LOS for various types of diseases is the most needed by the researchers. Proposed research applying data mining techniques to extract useful knowledge and suggest a model to estimate length of stay for diabetes patients in medical centers.

## 2. Related Work

More studies carried over on the researches considering the LOS of the patients. In the proceedings of "International Journal of Machine Learning and Computing" M. Maggie, et al, shows that the Short-Term Mortality Prediction for Elderly Patients Using Medicare Claims Data [2]. Samaneh Aghajani et al,determined the Factors Influencing Length of Stay and Predicting Length of Stay Using Data Mining in the General Surgery Department, the factors are the average number of visits per day, number of trials, and number of days of hospitalization before surgery, the most important of these factors was length of stay. The study shows the result that overall accuracy of the decision tree was 88.9% for the training data set. This study determined that all three algorithms can predict length of stay [1]. Negassa et al. concluded that support vector machine is the best fit to predict the Length of Stay of Caronary Patients [17].In Iran, Lorestan [19] province public hospital study conducted and it is demonstrated that, the factors affecting LOS are Increase in age would lead to increase in Average LOS and the average LOS of women is lesser than that of average LOS of men. They mainly focused on descriptive analysis on traditional statistical methods and they did not provide any prediction model. Saira Seemab et al. measures the different datamining models and enhancing

the planning and management of hospital resources in the research predicting patients duration of stay by mining hospital data[12].

Rowan et al. [21] proposed and implemented a software package demonstrating that artificial neural networks could be used as an effective LOS stratification instrument in postoperative cardiac patients. Multiple linear regressions is the method used to model the linear relationship between a dependent variable and one or more independent variables. The dependent variable is rainfall and independent variables are year, area of sowing, production. Purpose of this work is to find suitable data models that achieve high accuracy and a high generality in terms of yield prediction capabilities .Blais et al. [24] designed a screening and rating tool to quantify variables related to LOS in a medical psychiatric unit. The findings from this study showed that 25 variables, including patient, illness, and treatment variables, were likely to be related to LOS. Tu and Guerriere [26] indicated that ANNs can be used as a predictive tool to identify patients at increased risk for prolonged intensive care unit LOS following cardiac surgery. They claimed that the back propagation algorithm had not previously been developed for this area. C.Combes et al, have done extensive study on predictive ability of machine learning techniques such as multiple linear regression, regression trees, artificial neural network, support vector regression for predicting lengt of stay for the patients [8], but the regression tree performs the best, Lin et al. [20] explored the prediction of hospital stays for first-time stroke patients in a rehabilitation department by a proportional hazard regression (HR) model. They proposed using the HR model to predict the mean LOS of stroke patients. Jiang et al. [18] studied the use of four data mining techniques (logistic regression, neural network, decision tree, and ensemble model) to analyze the inpatient discharge data for average LOS based on input variables. The findings from this research showed that the ensemble model was the best fit, and age and chronic disease were the important predictors. Misclassification and average squared error were used to assess the models. The ensemble model had the lowest average squared error and the decision tree had the highest average squared error (0.22). Jin oh Kang et al shows that the Artificial Neural network model is giving better prediction accuracy than Classification and Regression Tree (CART) models. However, the CART models, which serve different information from ANN model, can be used to allocate limited medical resources effectively and efficiently. For the purpose of establishing medical policies and strategies, using those models together is warranted. Wrenn et al. [22] were able to predict LOS for an emergency department through developing and validating an ANN. The results were promising and showed that ANN can predict a patient's LOS within an average of less than 1.99 hours. Using a cohort of prospectively identified heart failure patients, Wright et al. [23] found that peripheral edema, chest pain, fatigue, serum albumin, serum sodium at admission and peak creatinine could result in hospital stays longer than six days. Blais et al. [24] studied factors that differentiated psychiatry patients' short LOS (7 days or less) and long LOS (more than 14 days). Age, impairment level, and +6 independent functioning levels were all independent predictors of LOS.

Most of the research on LOS has been conducted in rehabilitation and psychiatric fields [26]. Most models in the cardiac disease area have predicted in- hospital mortality, and statistical methods, especially descriptive analyses, have been applied in that research. Mao-Te Chuang et al concluded the results indicated that the random forest method yielded the most accurate and stable prediction model. Additionally, comorbidity, body temperature, blood sugar, and creatinine were the most influential variables for prolonged LOS in the UO group, whereas blood transfusion, blood pressure, comorbidity, and the number of ICU admissions were the most influential variables in the non-UO group. This study shows that supervised learning techniques are suitable for analyzing patient medical records in accurately predicting a prolonged LOS, thus, the clinical decision support system developed based on the prediction models may serve as reference tools for communicating with patients before surgery.

## 3. Research Method

### 3.1 Support Vector Machine

Support Vector Machine(SVM) is a supervised machine leaning technique that have been practically applied to handwriting recognition, textual mining, gene prediction, remote sensing classification, patients length of stay prediction and providing cutthroat results with the accessible classification methods and with smallest amount training data sets. It is a latest method for classification of both linear and non-linear data and in terms of predictive accuracy, it is a commanding algorithm. In fact, SVM shown in Figure 1 is a linear learning machine constructed through an algorithm that uses an optimization criterion and produces good performance because it needs smallest number of parameters. To view the training data, separate the hyperplane. The hyperplane has described by $wT.x+b=0$.

Where w is a T-dimensional weight vector and b is a scalar. The vector w points perpendicular to the separating hyperplane. The offset parameter b allows increasing the margin. When the training data are linearly separable, we select these hyperplanes, so that there are no points between them and then try on maximizing the distance between the hyperplane. To find out the distance between the hyperplane as $2/|w|$.To minimize $|w|$, need to ensure that for all i either $w.xi - b \geq 1$or $w.xi - b \leq 1$.

### 3.2 Classification and Regression Tree

The CART method is able to determine the complex interactions among variables in the final tree, in contrast to identifying and defining the interactions in a multivariable logistic regression model. To begin the CART analysis, simple random sampling without replacement was used to split the sample into equal sized developmental and validation samples. CART was applied first on a developmental sample then

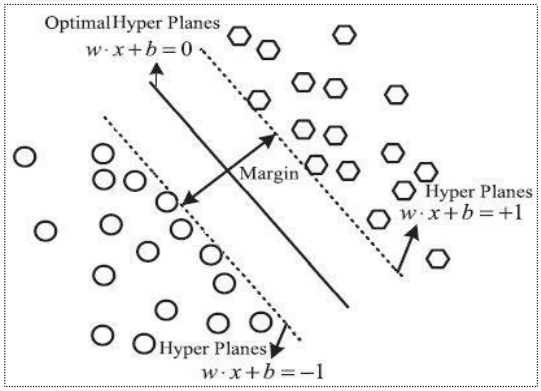



Figure 1.SVM Classification Technique

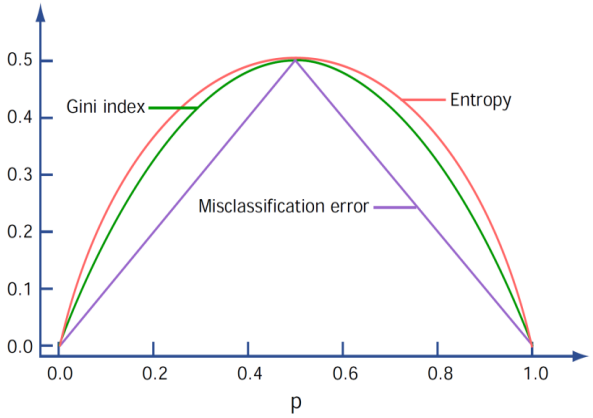



Figure 2.Gini Index for Information Gain

on a validation sample to assess the model's generalizability and to evaluate the over fitting of the model to the developmental sample.

Several sets of candidate predictors were used to build the classification trees. Using several iterations, CART models were used to determine a clinically logical fit, based on sensitivity and specificity, The Figure 2 shows the Gini Index method. It was used to split off the largest category into a separate group, with the default split size set to enable growing the tree. When the final tree was built, the tree was pruned, deleting the variables that did not further classify subjects, based on the variable importance.

CART algorithm to take advantage of the fact that cost information is typically available in utilization data, and patients with very similar utilization profiles should also have very similar cost. The vectors representing patients' utilization profiles are treated as input features and used to predict cost as the target variable. The CART algorithm constructs a rule based decision tree to segment the patient set by recursively partitioning the feature space until the patients within each partition satisfy certain purity constraint (based on cost). The final partitions correspond to the leaf nodes of the tree. CART is similar to a decision tree except at the leaf level a regression model is constructed in order to map to a continuous target variable, instead of doing a majority vote as in a decision tree classifier.Uncertainty measurement for two-class classification as a function of the proportional p in class 1.

For each leaf, we need to assign the prediction y which minimizes the loss for the regression problem.

**Algorithm for Regression Tree:**

Given the dataset D = $(x_1, y_1), . (x_n , y_n )$ where Xi € R,

$Y_i$ € Y= R

Step 1: $R1 = R^d$

Step 2: For (j =1 to d ), For (V= 1 to R)

Step 3: Split the data set:

$D_L = \{i:X_{ij} < V\}$ and $D_G = \{i:X_{ij} \geq V\}$

Step 4: Estimation of Parameters

$P_L = (\sum_i \in D_L Y_i) / |D_L|$ and $P_G = (\sum_i \in D_G Y_i) / |D_G|$

Step 5: Measurement of Squared loss

$$\sum_{i \in D_L} (Y_i - P_L)^2 + \sum_{i \in D_G} (Y_i - P_G)^2$$

Step 6: Find the Split with Minimal Loss

Step 7: Recursively call the above steps for both children nodes.

### 3.3 Neural Network

Neural network (NN) is a mathematical model or computational model based on biological neural network. Neural networks process information in a similar way the human brain does. The network is composed of a large number of highly interconnected processing neurons working in parallel to solve a specific problem. Neural networks are also used to find the duration of patients stay by using the classification technique. By adjusting the weights of neurons, minimize the errors in classification. After classifies the classes for the length of stay, you can predict the actual number of days for the patients going to stay and have to occupy the healthcare resources. In neural networks, use the multilayer feed forward networks for arranging the parameters.Mulitlayer feed forward networks consists of input layer, hidden layer and output layer.Perceptron training algorithm used to find the classes.

| Parameteres | Low | Medium | High | LOS(percentiles) | | |
|---|---|---|---|---|---|---|
| | | | | hours | days | weeks |
| Sugar level | <80 | 80 - 120 | >120 | 30 | 50 | 20 |
| Age | < 30 | 30 -45 | >45 | 32 | 45 | 23 |
| No of times pregnant | <2 | 3-5 | >5 | 29 | 44 | 27 |
| Additional co morbidities | <2 | 3-5 | >5 | 25 | 49 | 26 |
| Blood pressure | < 80 | 80-120 | >120 | 30 | 52 | 18 |
| Years of diabetes | <10 | 11-25 | 26-40 | 34 | 54 | 12 |

Table 1.Summary of Attributes

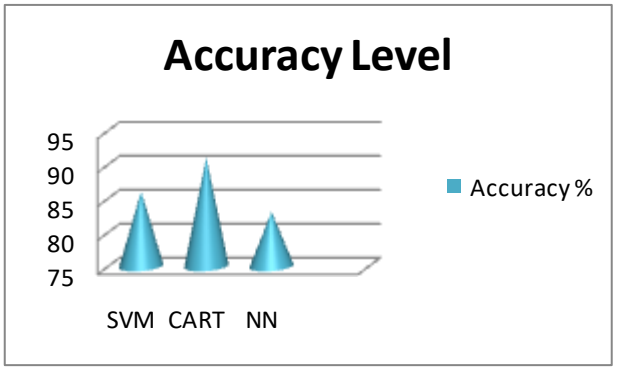
| Classification Techniques | Number of Attributes | Number of Classes | Overall Accuracy Level |
|---|---|---|---|
| SVM | 8 | 2 | 86% |
| CART | 6 | 3 | 91% |
| NN | 8 | 3 | 83% |

Table 2.Overall effiiciency of Techniques

## 4. Results and Analysis:

The proposed methods consists of Support vector machine classifier, Classification and Regression Tree, Neural network classifier are tested using the datasets collected from a network of physicians over a two year period. CART is very general technique and it can be applied to any patient population, it is useful to focus on a scrupulous use case to examine whether the results are clinically meaningful. This experiment needs the diabetes patient population which contains a total of approximately 7000 patients. From this populations, most of them from aged categories and having diabetes for more than 15 years. From the table it can be clearly observed that the majority of the patients had relatively low level of staying in number of weeks for the treatments. For example, the patients having diabetes more than 25 years needs to stay for more than 7 days .Half of the patients have to stay for less than seven days.

The above table 1 shown the summary of attributes which are used in classification and regression technique. There are totally six necessary attributes, from that it can easily predict the length of stay for the diabetes patients. From these parameters, the sugar level and the times of pregnant, additional co morbidities plays very crucial role in the prediction of duration for the patients to undergo treatments in hospital. If you have additional diseases or any additional symptoms, then you have to keep more days in observation. The following table 2 shows the accuracy level of the Support vector machine classifier, Classification and Regression Tree classifier and the neural network classifier. The proposed framework using 6978 diabetes data sets samples for the prediction of Length of stay for the patients. Among these,3078 data sets used as a Training data sets and the remaining 3900 data samples are used as the testing data sets.SVM classifier used 8 parameters to predict the Length of stay in two main classes. In Neural network classifier also uses 8 parameters to predict the duration of stay in three classes. Classification and Regression tree algorithm used only 6 efficient parameters and classify the length of stay for the patients in three classes such as hours, days, weeks. Figure: 3 shown the graphical demonstration of the accuracy level of the three techniques.




## 4. Conclusion

The proposed research shows that the classification and regression trees technique is the best tool to predict the patient's length of stay in the hospital. This proposed method uses the classification techniques such as Support vector machine, neural network classifier and ensemble classification and Regression Tree to find the Length of Stay for the diabetes patients.

It shows the differences in the above three classification techniques.

Figure 3.Accuracy level of the techniques

The results verified that the Classification and Regression Tree technique gives more accuracy than the other. The diabetes datasets are used in this paper to find the length of stay of the patients. This successful research shows that the Classification tree and Regression tree jointly used to produce the good results while predicting the Length of stay of the diabetes patients. This evaluations and case studies demonstrate the usefulness of the proposed approaches in identifying clinically meaningful instances for predicting the duration of stay for the diabetes patients while undergoing treatments, using the most basic observational data as described above. In the future, plan to expand our framework to leverage these additional data sources to provide enhanced performance and additional actionable insight, currently not consider the temporal relation-ships among different medical events or encounters.So, the future enhancements will be added those features.

## 5. References

[1] Samaneh Aghajani; Mehrdad Kargari, "Determining Factors Influencing Length of Stay and Predicting Length of Stay Using Data Mining in the General Surgery Department",Article 4, Volume 1, Issue 2, Spring 2016, Page 53-58 DOI: 10.20286/HPR-010251

[2] M. Maggie, G. Marzyeh, M. David, and O. Ziad, "Short-Term Mortality Prediction for Elderly Patients Using Medicare Claims Data," International Journal of Machine Learning and Computing, vol. 5(3), pp. 192-197, 2015.

[3] J. Q. Huang, P. M. Hooper, and T. J. Marrie, "An Example of Multiple Linear Regression, Hospital Length of Stay Among Drivers Hospitalized After a Motor Vehicle Collision," Regression Analysis Controlling for Confounding, 2015.

[4] R. Stoean, C. Stoean, A. Sandita, D. Ciobanu, and C. Mesina, "Ensemble of Classifiers for Length of Stay Prediction in Colorectal Cancer," in International Work-Conference on Artificial Neural Networks (IWANN 2015), vol. 9094, pp. 444-457, June 2015.

[5] M. Meitzner, "Inpatient Volume Forecasting Prediction Model," in DocumEnter, 2015.

[6] R. Houthooft, J. Ruyssinck, J. V. D. Herten, S. Stijven, I. Couckuyt, B. Gadeyne, F. Ongenae, K.Colpaert, J. Decruyenaere, T. Dhaene, and F. D. Turck, "Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores," Artificial Intelligence in Medicine, vol. 63, no. 3, pp. 191-207, 2015.

[7] A. Morton, E. Marzban, G. Giannoulis, A. Patel, R. Aparasu, and I. A. Kakadiaris, "A Comparison of Supervised Machine Learning Techniques for Predicting Short-Term In-Hospital Length of Stay among Diabetic Patients," in 13th International Conference on Machine Learning and Applications (ICMLA), pp. 428-431, 2014.

[8] C. Combes, F. Kadri, and S. Chaabane, "Predicting Hospital Length Of Stay Using Regression Models: Application To Emergency Department," MOSIM, 2014.

[9] P. Andreeva, "Data Modelling and Specific Rule Generation via Data Mining Techniques," in Expert Systems with Applications, pp. 1-6, 2014.

[10] Z. J. Eapen, S. D. Reed, Y. Li, R. D. Kociol, P. W. Armstrong, R. C. Starling, J. J. McMurray, B. M.Massie, K. Swedberg, J. A. Ezekowitz, G. C. Fonarow, J. R. Teerlink, M. Metra, D. J. Whellan, C. M. O'Connor, R. M. cali and A. F. Hernandez, "Do countries or hospitals with longer hospital stays for acute heart failure have lower readmission rates? Findings from ASCEND-HF. 1," Circulation: Heart Failure, vol. 6(4), pp. 727-732, 2013.

[11] X. Cheng, The Rise Of The Big Data: Why Should Statisticians Embrace Collaborations With Computer Scientists. MSc thesis, The University Of Georgia, 2013.

[12] Saira Seemab, Dr.Usman Qamar National University of Sciences and Technology, Pakistan "PEDICTING PATIENT'S LENGTH OF STAY BY MINING HOSPITAL DATA", Proceeding of the 3rd International Conference on Artificial Intelligence and Computer Science (AICS2015), 12 - 13 October 2015, Penang, MALAYSIA. (e-ISBN 978-967-0792-06-4). Organized by http://worldconferences.net.

[13] A. Azari, V. P. Janeja, and A. Mohseni, "Predicting Hospital Length Of Stay (PHLOS): A Multi-Tiered Data Mining Approach," in IEEE 12th International Conference on Data Mining Workshops, pp.17-24, 2012.

[14] J. L. Moran and P. J. Solomon, "A Review of Statistical Estimators for Risk-Adjusted Length of Stay: Analysis of the Australian and New Zealand Intensive Care Adult Patient Data-Base," BMC Medical Research Methodology, vol. 12, 2012.

[15] E. Kawaler, A. Cobian, P. Peissig, D. Cross, S. Yale, and M. Craven, "Learning To Predict Post-Hospitalization VTE Risk from EHR Data," AMIA Annual Symposium Proceedings, pp. 436-445, 2012.

[16] S. Sheikh-Nia, An Invetigation of Standard and Ensemble Based Classification Techniques for the Prediction of Hospitalization Duration. MSc thesis, University of Guelph, Ontario, Canada, 2012.

[17] Negassa A, Monrad ES. Prediction of length of stay following elective percutaneous coronary intervention.ISRN Surg 2011;2011:714935.

[18] Jiang X, Qu X, Davis L. Using data mining to analyze patient discharge data for an urban hospital. In: Proceedings of the 2010 International Conference on Data Mining; 2010 Jul 12-15; Las Vegas, NV. p. 139-44.

[19] Arab M, Zarei A, Rahimi A, Rezaiean F, Akbari F. Analysis of factors affecting length of stay in public hospitals in Lorestan Province, Iran. Hakim Res J 2010;12(4):27-32.

[20] Lin CL, Lin PH, Chou LW, Lan SJ, Meng NH, Lo SF, et al. Mode l-based prediction of length of stay for rehabilitating stroke patients. J Formos Med Assoc 2009;108(8):653-62.

[21] Rowan M, Ryan T, Hegarty F, O'Hare N. The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial post- operative factors. Artif Intell Med 2007;40(3):211-21.
[22] Wrenn J, Jones I, Lanaghan K, Congdon CB, Aronsky D.Estimating patient's length of stay in the Emergency Department with an artificial neural network. AMIA Annu Symp Proc 2005;2005:1155.
[23] Wright SP, Verouhis D, Gamble G, Swedberg K, Sharpe N, Doughty RN. Factors influencing the length of hospital stay of patients with heart failure. Eur J Heart Fail 2003;5(2):201-9.
[24] Blais MA, Matthews J, Lipkis-Orlando R, Lechner E, Jacobo M, Lincoln R, et al. Predicting length of stay on an acute care medical psychiatric inpatient service. Adm Policy Ment Health 2003;31(1):15-29.
[25] Stoskopf C, Horn SD. Predicting length of stay for patients with psychoses. Health Serv Res 1992;26(6):743-66.
[26] Tu JV, Guerriere MR. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. Proc Annu Symp Comput Appl Med Care 1992:666-72.