Original Article

Healthc Inform Res. 2013 June;19(2):121-129.
http://dx.doi.org/10.4258/hir.2013.19.2.121
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients

Peyman Rezaei Hachesu, PhD[1], Maryam Ahmadi, PhD[1], Somayyeh Alizadeh, PhD[2], Farahnaz Sadoughi, PhD[1]

[1]Department of Health Information Management, School of Health Management and Information Sciences, Tehran University of Medical Sciences, Tehran; [2]Industrial Engineering Faculty, Khajeh Nasir Toosi University of Technology, Tehran, Iran

Objectives: Predicting the length of stay (LOS) of patients in a hospital is important in providing them with better services and higher satisfaction, as well as helping the hospital management plan and managing hospital resources as meticulously as possible. We propose applying data mining techniques to extract useful knowledge and draw an accurate model to predict the LOS of heart patients. **Methods:** Data were collected from patients with coronary artery disease (CAD). The patient records of 4,948 patients who had suffered CAD were included in the analysis. The techniques used are classification with three algorithms, namely, decision tree, support vector machines (SVM), and artificial neural network (ANN). LOS is the target variable, and 36 input variables are used for prediction. A confusion matrix was obtained to calculate sensitivity, specificity, and accuracy. **Results:** The overall accuracy of SVM was 96.4% in the training set. Most single patients (64.3%) had an LOS ≤5 days, whereas 41.2% of married patients had an LOS >10 days. Moreover, the study showed that comorbidity states, such as lung disorders and hemorrhage with drug consumption have an impact on long LOS. The presence of comorbidities, an ejection fraction <2, being a current smoker, and having social security type insurance in coronary artery patients led to longer LOS than other subjects. **Conclusions:** All three algorithms are able to predict LOS with various degrees of accuracy. The findings demonstrated that the SVM was the best fit. There was a significant tendency for LOS to be longer in patients with lung or respiratory disorders and high blood pressure.

Keywords: Length of Stay, Data Mining, Coronary Artery Disease, Patients, Extract

I. Introduction

Coronary artery disease (CAD) is a major cause of disability in adults and a major cause of death in developed countries resulting in several illnesses, disabilities, and deaths as well. It should be noted that cardiovascular diseases are characterized by prolonged length of stay (LOS) [1]. LOS is defined as the number of days that a patient is hospitalized in a hospital or a similar medical facility. There has been considerable interest in controlling hospital costs, particularly in cardiac diseases; thus, hospitals try to make LOS as short as possible [2]. The length of hospital stay is an actual parameter applied to identify health care resource utilization, health cost, and severity of illness [3]. The use of LOS is highly predictive of

Submitted: February 6, 2013
Revised: 1st, March 14, 2013; 2nd, March 28, 2013
Accepted: April 1, 2013

Corresponding Author
Maryam Ahmadi, PhD
School of Health Management and Information Sciences, Tehran University of Medical Sciences, No. 6, Rashid Yasemi st., Tehran 199671388, Iran. Tel: +98-2188665052, Fax: +98-2188793805, E-mail: m-ahmadi@tums.ac.ir, prhim88@gmail.com

inpatient costs as a marker of resource utilization [4]. Hospitals have severely limited beds to hold inpatients, and as most of them are facing substantial financial pressure, it is extremely important to find ways to reduce health care costs [5]. One solution is to predict and determine the discharge date and LOS of each patient by a number of complementary techniques and technologies, such as data mining [6]. For a hospital administrator to be considered successful, predicting and evaluating LOS data is laborious but essential [7]. Precise prediction of LOS facilitates the efficiency of bed occupancy management in hospitals. Therefore, exact and proper prediction of LOS has become increasingly important for hospital management and health care systems [4]. Meanwhile, awareness of factors and elements that determine LOS could promote the development of efficient clinical pathways and optimize resource utilization and management [8]. In addition, many hospitals cannot predict and measure future admission requests. Many hospitals have no ability to predict and measure future admission requests. Also, successful prediction of discharge dates and duration of hospital stay allows the corresponding scheduling of elective admissions, leading to diminished variance in bed occupancy [9]. Providing an efficient and accurate model to predict LOS for different types of diseases is one of the issues considered by researchers. Obviously, developing models for predicting and determining LOS in hospitals can be very useful for hospital management, particularly for prioritizing health care policies and promoting health services, comprising the appropriate allocation of health care resources according to differences in patients' LOS along with considering patients' health status and social-demographic features [3]. Ideally, better prediction models are needed to facilitate the decision-making process and cannot be replaced by judgment. For these reasons, providing an efficient and accurate model to predict LOS for various types of diseases is one of the issues considered by researchers. However, there has been relatively little research related to LOS prediction. Therefore, we applied data mining techniques to extract useful knowledge and suggest a model to estimate length of stay for coronary artery patients in cardiovascular centers.

1. Literature Review

Studies on factors contributing to LOS have regularly appeared in the literature. One study conducted to determine the factors affecting LOS in public hospitals in Lorestan Province, Iran demonstrated that, first, an increase in age would lead to an increase in average LOS and, second, the average LOS of men is longer than that of women. The t-test, one-way ANOVA, and multifactor regression were used for the analysis. They did not provide any prediction model, because they focused on descriptive analysis based on traditional statistical methods [10]. Rowan et al. [8] proposed and implemented a software package demonstrating that artificial neural networks (ANNs) could be used as an effective LOS stratification instrument in postoperative cardiac patients. Blais et al. [11] 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 [12] 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. Lin et al. [13] 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. [5] 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 (0.21), and the decision tree had the highest average squared error (0.22). Wrenn et al. [14] 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 <1.99 hours. Using a cohort of prospectively identified heart failure patients, Wright et al. [1] 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. [11] 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.

As previously mentioned, most of the research on LOS has been conducted in rehabilitation and psychiatric fields [15]. Most models in the cardiac disease area have predicted in-hospital mortality [16], and statistical methods, especially descriptive analyses, have been applied in that research. Hence, raising our awareness of factors that have an impact on cardiac patients' LOS is essential in order to determine and develop a useful and efficient model to predict LOS. This

 http://dx.doi.org/10.4258/hir.2013.19.2.121

research aims to investigate the important factors that can be assessed to predict the LOS of patients with coronary artery diseases.

2. Data Mining Algorithms

Finding undiscovered information and useful patterns in a database is often referred to as data mining [17]. Data mining is heavily used in the health and medical field in applications such as disease prediction and patient management [18]. Relationships, rules, and essential information about or from the data cannot be easily extracted because of database size and other features. We used some of the most common predictive data mining methods for our goals as follows.

1) Artificial neural networks

ANNs are used to perform multivariate analysis to identify both linear and non-linear patterns among data variables [19]. Due to their good predictive performance, ANNs are the most popular method in various areas of medicine [20] and lead to appropriate decisions. An ANN consists of many connected processing elements, including multiple input nodes and weighted interconnections. The radial-basis-function (RBF) ANN was developed to recognize CAD.

2) Support vector machines (SVMs)

A category of classification that has received increasing attention in recent years is the SVM. It is a new method for classification of both linear and non-linear data [21], and in terms of predictive accuracy, it is a powerful algorithm . In fact, SVM is a linear learning machine constructed through an algorithm that uses an optimization criterion. We apply RBF kernel mode because of its good general performance and because it has the smallest number of parameters [22].

3) Decision tree

Generally, a decision tree as a visual and analytical decision support tool is a graphic representation of obtained knowledge in the form of a tree (flow chart like structure), where each non-leaf node denotes a test on an attribute, and each branch indicates an output of the test [23]. It uses a combination of mathematical and computational techniques to aid description and classification, and to extract knowledge of data set [24]. Because nodes and branches are organized hierarchically, they are easy to understand and interpret. They are reliable and have better accuracy in clinical decision-making [25]. C5.0 decision trees are the most current decision tree algorithms. The C5.0 algorithm with 10-fold cross validation and 20 trials using boosting was applied in this research.

4) Ensemble models

The ensemble method creates a new model by combining SVM, C5.0, and ANN models.

II. Methods

1. Patient Population

The cohort consisted of hospitalized patients during the study period, which started on July 18, 2006 and ended on December 30, 2011. We identified 4,948 patients who were admitted to the Academic and Educational Hospital of Rajaei Cardiovascular Medical & Research Center in Tehran, Iran with heart disease-related diagnoses. Only CAD data were included in the study (n = 3,512). Significant CAD was defined as at least one point of 50% or greater diameter stenosis in at least one coronary artery vessel [26].

2. Data Set

The data sets were stored in a database management system of Microsoft structured query language (MS-SQL) database. We extracted and constructed a new data set for LOS of CAD. However, 246 patients were removed from the analyses, because patient records data, such as identification, were unavailable from the data set. Thus, 3,266 patients were included in the final data set for further analysis. Table 1 shows features with acceptable class and values. The data set contained 36 attributes. Finally, we organized the data set into two groups, including categorical and numerical features. We categorized data values and derived new fields from existing data in the following features: ejection fraction, diastolic blood pressure, systolic blood pressure, smoking, triglyceride, low-density lipoprotein, high-density lipoprotein, hemoglobin, serum cholesterol, and fasting blood sugar. These features were changed to categorical attributes for better analysis and to obtain good results.

3. Data Pre-processing

Data cleansing and preprocessing are essential to have optimal results [27]. Therefore, we performed the following cleansing and preprocessing: repeated records, fields with spelling errors, additional tokens, other irregularities, and irrelevancies were deleted. The next step of pre-processing was the handling of patient records with missing and outlier data.

4. Dealing with Missing Values

The hospital data set had many features with missing values. Several replacement strategies were adopted to fill the missing values. First, if a feature was encountered in more than 50%

Table 1. The demographic and clinical characteristics of the length of stay data set (n = 2,064)

Variable	Value
Pulse rate (bpm)	Numerical (30–150), mean ± SD (75.9 ± 10.2)
Age (yr)	Numerical (15–94), mean ± SD (58 ± 3)
Serum creatinine (mg/dL)	Numerical (0.2–11.6), mean ± SD (1.2 ± 0.55)
Gender	1, male; 0, female
Fasting blood sugar (mg/dL)	1, 70–100; 2, 101–126; 3, ≥127
Serum cholesterol (mg)	1, ≤200; 2, 200–239; 3, ≥240
Hemoglobin (gm/mL)	1 (normal), ≥13.5 and ≤18 (men & age >17 yr) or ≥12 and ≤16 (women & age >17 yr) or ≥11 and ≤16 (age <17 yr); 2 (low level), <13.5 (men & age >17 yr) or <12 (women & age >17 yr) or ≤11 (age <17 yr); 3 (high level), >18 (men & age >17 yr) or >16 (women & age >17 yr) or >16 (age <17 yr)
High-density lipoprotein (mg/dL)	1 (best), ≥60 yr; 2 (poor), men & ≤40 yr or women & ≤50 yr; 3 (better level), men & 40–59 yr or women & 50–59 yr
Low-density lipoprotein (mg/dL)	1 (optimal), ≤100; 2 (near optimal), 100–129; 3 (border line high), 130–159; 4 (high), 160–189; 5 (very high), ≥190
Triglyceride (mg/dL)	1, <150; 2, 150–199; 3, 200–499; 4, ≥500
Marital status	0, single; 1, married
Diabetes mellitus	1, history of diabetes; 0, no such history
Hypertension (mmHg)	0, no; 1, yes
Family history of coronary disease	0, no; 1, yes
Past history of heart disease	0, no; 1, yes
Dyslipidemia	0, no; 1, yes
Smoker or not	2, current; 3, past; 4, recent; 5, never
Ejection fraction	1 (good), 50–75; (fair), 30–49; (poor), <30
Chest pain	2, yes; 3, no
Systolic blood pressure (mmHg)	1 (hypotension), <90; 2 (desirable), 90–119; 3 (border line hypertension), 120–139; 4 (hypertension), ≥140
Diastolic blood pressure (mmHg)	1 (hypotension), ≤60; 2 (desirable), 61–79; 3 (border line hypertension), 80–89; 4 (hypertension), ≥90
Exercise stress test	0, normal; 1, abnormal
Absence or presence of one or more disorders as well as a primary disease	0, no; 1, yes
Valvular heart disease	0, no; 1, yes
ST segment and T wave of electrocardiogram changes	0, normal; 1, having ST-T wave abnormality
Coronary artery disease diagnosed by physicians (diagnosis)	0, no; 1, yes
Drug category[a]	0, not used; used
Type of medical insurance used by the patient	1, medical services insurance; 2, insured rural; 5, social security
Length of stay (LOS, day)	1, if LOS ≥ 0 and LOS ≤ 5; 2, LOS between 6–9; 3, LOS > 10

SD: standard deviation.

aStatin, nitrates, inotropic, diuretic, calcium, channel blocker, beta blocker, antiplatelet, anticoagulant, angiotensin-converting-enzyme inhibitor.

 http://dx.doi.org/10.4258/hir.2013.19.2.121

Table 2. Attribute with missing and alternative value

Attribute value	Missing data (%)	Method; alternative value
Age	0	-
Sex	0	-
Marital status	0	-
Exercise stress test	0	-
Diabetes	0	-
Hypertension	0	-
Dyslipidemia	0	-
Family history	0	-
Comorbidity	5.24	Mode; 0
Ejection fraction	4.36	Mode; 1
Diagnosis	2.17	Mode; 1
Hemoglobin	8.86	Mode; 2 (for class 1) Mode; 2 (for class 2) Mode; 2 (for class 3)
Creatinine	11.2	Mean; 1.90 (for class 1) Mean; 1.25 (for class 2) Mean; 1.18 (for class 3)

of records with missing values, that characteristic was determined not to be an effective feature in the analysis. As a result, such a feature, such as weight or job, was removed. Second, if a feature was encountered in less than 12% of records with missing values, the mean values of records were replaced instead of missing values in the numeric features. For example, creatinine showed 11.2% missing data. The mean value was replaced according to its accepted class (Table 2). If the feature was in nominal or ordinal type, mode values were replaced. The comorbidity, ejection fraction, hemoglobin, and diagnosis features followed the mentioned rule (Table 2).

In the third strategy, the C5.0 algorithm was applied to those features showing missing values in more than 10% of records. We filled the missing values of these features using this algorithm with the highest accuracy according to Table 3.

To resolve outliers in each feature, we transformed the data to Microsoft Excel format and detected outlier data that was clear using methods such as sorting. Otherwise, the nearest acceptable non-outlier value for that feature was used to replace the outlier value [28].

5. Attribute Coding

Data was coded by some valid resources, such as heart disease associations and the Wikipedia Website. Scaling and coding features are given in Table 1.

Table 3. Number of features with missing data values and accuracy results

Feature	Missing data (%)	Accuracy result (%)
High-density lipoprotein	32.7	96.8
Low-density lipoprotein	16.6	94.2
Pulse rate	25.4	Mean error 0.7[a]
Systolic blood pressure	24.6	97.9
Diastolic blood pressure	24.6	96.9
Chest pain	21.1	97.6
Cholesterol	19.7	97.9
Fasting blood sugar	18.1	91.7
Triglyceride	16.9	95.8
Smoking	26.6	97.2

[a]Calculated by linear regression.

6. Training and Test Data Sets

After cleaning and preprocessing, 2,064 completed records were extracted and obtained for data mining tasks. Separating the data into training and testing sets is an important part of evaluating data mining models. We partitioned the data set into a training set and a testing set; 80% of the data (1,643 records) was used for training, and 20% of the data (421 records) was used for testing. The training set was used to adjust the parameters of the models, and the testing set was used to evaluate its predictive ability.

III. Results

1. Statistical Analysis

The mean age of 2,064 patients was 58.2 ± 13.0 years (aged 15–94) with most subjects between 54–64 years old. The sample was composed of 1,266 (61.3%) men and 798 (38.6%) women. 1,264 (61.2%) patients were diagnosed with CAD and 800 (38.8%) without CAD. LOS class 3 included most patients (39.3%); LOS class 1 and LOS class 2 comprised 35.8% and 24.9%, respectively. Table 4 demonstrates statistical results of the data set.

These data mining models were developed by a data mining classification tool. We evaluated the model created using training data and then applied test data to compare the results. We used SPSS Clementine 12 (SPSS Inc., Chicago, IL, USA) to build mining models.

The performance of a diagnostic method is usually evaluated in terms of classification accuracy, sensitivity, and specificity. In fact, accuracy is the percentage of correct decisions if CAD is predicted when the test is true and a non-CAD is

predicted when the test is false [29]. Sensitivity is the true positive, and specificity is the true negative rate of CAD.

Table 5 shows the sensitivity, specificity, and accuracy of different classification techniques. A confusion matrix was obtained to calculate sensitivity, specificity, and accuracy. The overall accuracy of SVM was 96.4% in the training set. The ensemble algorithm showed a stronger performance than other algorithms with a sensitivity of 98.2%.

Table 4. Statistical result of some important features

Variable	Value	Proportion (%)	LOS 1	LOS 2	LOS 3
Marital status	0	8.6	63	38	77
	1	91.4	675	476	735
Past history of CAD	0	72.4	543	356	595
	1	27.6	195	158	217
Diabetes	0	71.5	373	199	228
	1	28.5	365	315	584
Diagnosis	0	38.8	373	199	228
	1	61.2	365	315	584
Insurance	5	46.5	341	255	364
	1	13.5	11	66	101
	2	12.1	75	54	121
Smoking	2	33.1	305	176	203
	3	19.7	153	99	154
	4	4.6	25	31	50
	5	42.6	255	208	405
Comorbidity	0	58.2	513	297	391
	1	41.8	225	217	421
Fasting blood sugar	1	42.3	322	204	348
	2	31.1	242	167	232
	3	26.6	174	143	232
Diastolic blood pressure	2	37.9	275	189	319
	3	53.2	396	277	426
	4	7.9	59	44	62
Ejection fraction	1	25.3	250	144	128
	2	74.4	483	369	68
Chest pain	2	74.9	585	386	576
	3	25.1	153	128	236
Hemoglobin	1	25.3	250	144	128
	2	74.4	483	369	684

LOS: length of stay, CAD: coronary artery disease.

Table 5. Analysis of length of stay data set with classification techniques

Algorithm	Accuracy (%)	Specificity (%)	Sensitivity (%)
Decision tree (C5.0)	83.5	65.2	97.1
Neural network	53.9	65.1	72.2
Support vector machine	96.4	97.3	98.1
Ensemble algorithm	95.9	93.4	98.2

2. Important Features

The relative importance of each variable in evaluating the model is associated with the importance of each feature in making a prediction, and it does not relate to the model accuracy [30]. Also, the sum of the values for all variables in algorithms is 1.0. The SVM model, with earlier parameter setting, was used to extract important factors. In Table 6, features with great impact on CAD are listed in order of variable importance. The most significant variables were drug categories, such as nitrates and anticoagulants as well as CAD diagnosis. Comorbidity is also a strong predictor of prolonged LOS. Sex was significant in predicting LOS since men had longer LOS than women. Age played a notable role as well since analysis revealed that patients aged <50 and ≥80 statistically had increased mean LOS. LOS class 1 comprised mostly single patients (64.3%) and 24.5% of married patients, 41.2% of married patients were in LOS class 3 and 34.3% were married in LOS class 2. Furthermore, insurance type had a predictive power. Patients with social security and rural medical insurance were in LOS class 3. Thus, the most notable factors influencing LOS obtained by algorithm

Table 6. Important features extracted by support vector machine model

Features	Relative weight
Anticoagulant drugs	0.1824
Nitrate drugs	0.1033
Diagnosis	0.9200
Diastolic blood pressure	0.7870
Ejection fraction	0.1280
Comorbidity	0.0586
Marital status	0.0424
Chest pain	0.0272
Sex	0.0340
High-density lipoprotein	0.0310
Hemoglobin	0.0870
Smoking	0.0780
Insurance type	0.0350
Cholesterol	0.0460
Age	0.2130
ST-T change	0.0320

http://dx.doi.org/10.4258/hir.2013.19.2.121

Table 7. Important significant of extracted rules with using of C5.0 algorithm

	Antecedent	Consequent
1	If diagnosis = 1.0, comorbidity = 1.0, Hgb = 2.0, smoking in [2.0, 3.0], triglyceride in [1.0 and 2.0][a]	then LOS = 3
2	If Hgb = 2.0, creatinine ≤ 1.80, diagnosis = 1.0, comorbidity = 1.0, smoking in [2.0] and EF ≤ 2.0	then LOS = 3
3	If Hgb = 1.0, EF ≤ 1.0, smoking [2.0, 3.0, 4.0] and chest pain = 3.0	then LOS = 1
4	If comorbidity = 0.0 and smoking = 5.0	then LOS = 2
5	If diagnosis = 0.0, chest pain = 2.0, and EF in [2.0]	then LOS = 3
6	If diagnosis = 0.0 and EF = 1.0 and triglyceride = 2.0 and creatinine >1.1 and comorbidity = 0.0 and insurance = 5.0	then LOS = 1
7	If EF = 2.0 and ST = 1.0 and smoking = 2.0 and comorbidity = 0.0 and insurance = 9.0	then LOS = 2
8	If diagnosis = 1.0 and diastolic BP = 3.0 and EF = 1.0 and FBS = 2.0 and triglyceride = 3.0 and chest pain = 2.0 and insurance = 5.0	then LOS = 3
9	If diagnosis = 0.0 and EF = 2.0 and triglyceride = 3.0 and creatinine ≤ 1.1 and past history = 0.0 and smoking in [2.0, 3.0, 5.0] and marital = 1.0	then LOS = 1
10	If diagnosis = 1.0 and diastolic BP = 3.0 and EF = 1.0 and FBS = 2.0 and triglyceride = 3.0 and chest pain = 2.0 and insurance = 5.0	then LOS = 3
11	If comorbidity = 0.0 and insurance = 1.0 and marital = 0.0	then LOS = 1

BP: blood pressure, EF: ejection fraction, Hgb: hemoglobin, FBS: fasting blood sugar, ST-T: ST segment and T wave of electrocardiogram changes.

[a]Values are according Table 1 (1, if LOS ≥ 0 and LOS ≤ 5; 2, LOS between 6–9; 3, LOS > 10).

are drugs (nitrates), being diagnosed with CAD, comorbidity, hemoglobin, ejection fraction, insurance type, smoking, family history, and sex. These factors were extracted with 99.1% accuracy by the SVM.

3. Extract Rules

This section describes how these significant rules were extracted. Based on the C5.0 algorithm with the previously mentioned parameter setting, 11 rules, which are interpreted as IF (antecedent) and Then (consequent) in Table 7, were generated with a mean estimated accuracy of 95.3%. The presence of comorbidities such as lung and digestive disorders, ejection fraction <2, currently being a smoker, and insurance type = 5 in coronary artery patients was associated with them having a longer LOS than other subjects. Absence of comorbidities, being a nonsmoker, being single, not having chest pains, and having medical service insurance had a positive effect on decreasing LOS. The extracted rules are interpreted in Table 7. However, more investigation with more features and larger data sets is still required.

IV. Discussion

This study investigated the determinants of length of hospital stay in patients' representative of CAD admitted to a cardiovascular center. Many studies of length of hospital stay predict the duration of stay based on laboratory parameters or other quantifiable variables [1]. Our findings showed that a LOS greater than 10 days was associated with comorbidity and diastolic blood pressure features. There was a significant tendency for LOS to be longer in patients with lung or respiratory disorders and high blood pressure. Hence, comorbidities such as lung disorders and hemorrhage have an impact on long LOS and are important features in predicting LOS. However, Appelros [31] in his study demonstrated that comorbidities do not significantly influence LOS, while smoking has an inverse effect on acute LOS. They claimed that stroke severity is an important predictor of both acute and total LOS. Some studies have reported that patient demographics and hospital attributes were the two major factors that contributed to identifying patient LOS [3], and the most useful patient feature for predicting LOS was patient's age [32]. In many studies, age has been found to be a very significant predictor of LOS [5]. Our results from the retrospective study of LOS replicate a number of previously reported findings.

In this study, the extracted rule demonstrated that patients with normal levels of hemoglobin, medical services insurance, ejection fraction with class 1 (good level, 50–75), and those with no past history of cardiac disease and comorbidities and also non-smokers had a normal LOS in hospital. We

found that married men had longer LOS than single men and women. Patients who were using statins and blood anticoagulant drugs had a prolonged LOS. These findings did not conform to those of other study [2].

Our promising results indicate that SVM and an ensemble model, which applies three data mining algorithms, can predict patient's LOS, but still SVM is the best fit. We suggest that the SVM model is optimal for predicting mean LOS of CAD. According to other studies, SVM has the highest forecasting accuracy among other data mining algorithms. Today, this algorithm is becoming increasingly common in the medical and health field [33].

In addition to disease-related factors, LOS may be affected by factors unrelated to the disease, such as availability of hospital beds and rehabilitation facilities as well as discharge possibilities [31]. Due to the lack of medical facilities, staff shortages, and the increasing cost of health care, it is extremely important to optimize LOS and to identify factors affecting it. Note that LOS is influenced by individual characteristics such as weight, disease status, patient management style, hospital management, organizational characteristics, and other features [34,35].

This study was confined to the exploration of length of hospital stay of CAD patients with no consideration of other factors, such as the ethnic and socio-cultural environment of each patient and admission status. Health care data is generally not structured, and it is distributed over various locations. In terms of practicality, the primary limitation of this study is that all data were obtained from a specialized cardiovascular center. The factors selected for LOS prediction tended to be less social and more condition specific (e.g., presence of cardiac coronary artery bypass graft). Some important variables could not be considered, including factors such as alcohol consumption, other comorbidities, distance between patients' place of residence and the hospital, and admission type (elective and urgent). However, we attempted to identify the primary factors related to longer LOS. These data should also be collected uniformly to increase prediction accuracy.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This study was a part of PhD thesis supported and funded by Tehran University of Medical Sciences (Grant No. TUMS/SHMIS-1390/624). We would like to express our gratitude for the Dr. Rahim Firuzi and Azadeh Olyaei for their skillful and invaluable assistance. The author would like to thank Dr. Gholpira, Dr. Moghaddan, and other staff members of the information tech unit of the Rajaei Cardiovascular Medical & Research Center for allowing us to use the data and their assistance in retrieving patients' data.

References

1. 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.
2. Gomez V, Abasolo JE. Using data mining to describe long hospital stays. Paradigma 2009;3(1):1-10.
3. Lim A, Tongkumchum P. Methods for analyzing hospital length of stay with application to inpatients dying in Southern Thailand. Glob J Health Sci 2009;1(1):27-38.
4. Chang KC, Tseng MC, Weng HH, Lin YH, Liou CW, Tan TY. Prediction of length of stay of first-ever ischemic stroke. Stroke 2002;33(11):2670-4.
5. 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.
6. Isken MW, Rajagopalan B. Data mining to support simulation modeling of patient flow in hospitals. J Med Syst 2002;26(2):179-97.
7. Walczak S, Scorpio RJ, Pofahl WE. Predicting hospital length of stay with neural networks. In: Cook DJ, editor. Proceedings of the Eleventh International FLAIRS Conference; 1998 May 18-20; Sanibel Island, FL. Menlo Park, CA: AAAI Press; 1998. p. 333-7.
8. 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 postoperative factors. Artif Intell Med 2007;40(3):211-21.
9. Robinson GH, Davis LE, Leifer RP. Prediction of hospital length of stay. Health Serv Res 1966;1(3):287-300.
10. 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.
11. 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.
12. Tu JV, Guerriere MR. Use of a neural network as a pre-

 http://dx.doi.org/10.4258/hir.2013.19.2.121

dictive instrument for length of stay in the intensive care unit following cardiac surgery. Proc Annu Symp Comput Appl Med Care 1992:666-72.
13. Lin CL, Lin PH, Chou LW, Lan SJ, Meng NH, Lo SF, et al. Model-based prediction of length of stay for rehabilitating stroke patients. J Formos Med Assoc 2009;108(8):653-62.
14. 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.
15. Stoskopf C, Horn SD. Predicting length of stay for patients with psychoses. Health Serv Res 1992;26(6):743-66.
16. Negassa A, Monrad ES. Prediction of length of stay following elective percutaneous coronary intervention. ISRN Surg 2011;2011:714935.
17. Jilani TA, Yasin H, Yasin M, Ardil C. Acute coronary syndrome prediction using data mining techniques: an application. Int J Inf Math Sci 2009;5(4):295-9.
18. Liu P, Lei L, Yin J, Zhang W, Naijun W, El-Darzi E. Healthcare data mining: predicting inpatient length of stay. In: Proceedings of the 3rd International IEEE Conference Intelligent Systems; 2006 Sep 4-6; London, UK. p. 832-7.
19. Kudyba S, Gregorio T. Identifying factors that impact patient length of stay metrics for healthcare providers with advanced analytics. Health Informatics J 2010;16(4):235-45.
20. Rani KU. Analysis of heart diseases dataset using neural network approach. Int J Data Min Knowl Manag Process 2011;1(5):1-8.
21. Kamath C. Scientific data mining: a practical perspective. Philadelphia (PA): Society for Industrial and Applied Mathematics; 2009.
22. Son YJ, Kim HG, Kim EH, Choi S, Lee SK. Application of support vector machine for prediction of medication adherence in heart failure patients. Healthc Inform Res 2010;16(4):253-9.
23. Kajabadi A, Saraee MH, Asgari S. Data mining cardiovascular risk factors. In: Proceedings of International Conference on Application of Information and Communication Technologies; 2009 Oct 14-16; Baku, Azerbaijan. p. 1-5.
24. Suryawanshi RD, Thakore DM. Classification techniques of datamining to identify class of the text with fuzzy logic. In: Proceedings of 2012 International Conference on Information and Computer Applications; 2012 Feb 17-18; Hong Kong. p. 263-7.
25. Sitar-Taut DA, Sitar-Taut AV. Overview on how data mining tools may support cardiovascular disease prediction. J Appl Comput Sci 2010;4(8):57-62.
26. Wexler L, Brundage B, Crouse J, Detrano R, Fuster V, Maddahi J, et al. Coronary artery calcification: pathophysiology, epidemiology, imaging methods, and clinical implications. A statement for health professionals from the American Heart Association. Writing Group. Circulation 1996;94(5):1175-92.
27. Kang JO, Chung SH, Suh YM. Prediction of hospital charges for the cancer patients with data mining techniques. J Korean Soc Med Inform 2009;15(1):13-23.
28. Yaghini M. Data mining SPSS Clementine 12.0: 4. Handling missing and Clementine outliers values. Tehran, Iran: IUST; 2010.
29. Mullins IM, Siadaty MS, Lyman J, Scully K, Garrett CT, Miller WG, et al. Data mining and clinical data repositories: insights from a 667,000 patient data set. Comput Biol Med 2006;36(12):1351-77.
30. Malliaris ME, Pappas M. Revenue generation in hospital foundations: neural network versus regression model recommendations. Int J Manag Inf Syst 2011;15(1):59-66.
31. Appelros P. Prediction of length of stay for stroke patients. Acta Neurol Scand 2007;116(1):15-9.
32. Ghoson AM. Decision tree induction & clustering techniques in SAS Enterprise Miner, SPSS Clementine, and IBM Intelligent Miner: a comparative analysis. Int J Manag Inf Syst 2010;14(3):57-70.
33. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 2008;77(2):81-97.
34. McMullan R, Silke B, Bennett K, Callachand S. Resource utilisation, length of hospital stay, and pattern of investigation during acute medical hospital admission. Postgrad Med J 2004;80(939):23-6.
35. Vahidi R, Kushavar H, Khodayari R. Factors affecting coronary artery patients hospital length of stay of Tabriz Madani hospital 2005-2006. J Health Adm 2006;9(25):63-8.