Analytics for the Sustainable Use of Resources in Hospitals: LOS classification for nephropathy patients

Xinli Zhang a, Xue Shen b
Sichuan University, Business school, Chengdu, China
a zhangxinli@scu.edu.cn, 13980876069,86
b 979237665@qq.com, 17882247596,86

ABSTRACT

Aim: Exploring the impact of nephropathy patient characteristics on length of stay (LOS) grading, proposing a path of LOS classification based on the characteristics of patients, and providing suggestions for the accurate management of shunting of patients with nephropathy patient to promote the sustainable development of the hospital.

Methods: The data of inpatients from the Department of Nephrology of a large hospital in 2016 were used, including five variables: gender, age, patient type, medical insurance type, and LOS. Based on quantifying patient attribute variables, we use three steps to finish the grading. Firstly, using the factor analysis to extract the common factors of patient characteristics. Secondly, according to the results of factor analysis, using k-means clustering analysis to classify the patients. Finally, According to the characteristics of different types of patients and the law of LOS differences, a LOS classification path based on patient characteristics is proposed.

Results: The factor analysis shows that the LOS common factor characteristics are disease characteristics, attribute characteristics and reimbursement ratio characteristics. The k-means clustering indicates that the patients are divided into 5 categories: the mean LOS in category 1 is 15.78, Patient characteristics: Mostly elderly women with the blood resuscitates patients (38.2%) or tumor recovery patients (30.3%), city medical insurance (50%); the mean LOS in category 2 is 10.5, Patient characteristics: Mostly strong men with the ordinary patients (62.5%), City medical insurance (79.2%); the mean LOS in category 3 is 7.62, Patient characteristics: Mostly young men with the other patients (99.7%), Provincial medical insurance (73.1%); the mean LOS in category 4 is 13.7, Patient characteristics: Mostly women in pre-old age with the Ward daytime patient (38.9%) or other patients (31.8%), Urban rural medical insurance (60.6%); the mean LOS in category 5 is 6.73, Patient characteristics: Mostly young men with the other patients (99.3%), Provincial medical insurance (54.4%). According to the characteristic differences among patients and the law of LOS differences, a model of patients’ LOS classification path was proposed.

Conclusion: The LOS classification path based on patient characteristics can realize the pre-classification management of patients, which has practical significance for early intervention of hospital resources.

1. INTRODUCTION

Hospitals objectives require balance in treating each patient's condition effectively while efficiently distributing healthcare resources to patient populations over time [1], driving hospitals toward more sustainable operations. Indeed, sustainable health care may also transpire in other forms [2-4]: customer-oriented sustainability (such as an enhanced quality of patient care, increased patient satisfaction,
reduced medical bills). For patients, prolonged hospital stays increase the risk of adverse events, such as hospital-acquired infections, adverse drug events, poor nutritional levels, and other complications[5–8]. Employee-oriented sustainability (such as improvement in professionals’ job satisfaction) is important to improve sustainability awareness and enhance staff involvement to sustain the practice of sustainability in healthcare[9]. Programs such as motivation, training, and development of instructional resources or materials need to be increased continuously[10,11]; or community-oriented sustainability (such as saving energy and materials and reducing pollution). In recent years, a major contributor to realizing these improvements in health care outcomes has been the increasing use of [big] data analytics[12,13].

To predict LOS of patients, Wrenn et al[14,15] proposed an artificial neural network based prediction model for an Emergency Department. Azari et al.[16] proposed a multi-tiered data mining approach for predicting LOS. They used clustering technique to create a training dataset to train several classification algorithms for prediction of LOS. Hachesu et al[17,18] demonstrated multiple classification algorithms (decision tree, support vector machines, logistic regression) with varied level of accuracy. Most of the previous works emphasized on the use of novel or hybrid classification algorithms or complex ensemble models [18,19]. However, the performance of any prediction model depends on the number and type of its inputs variables as well [20,21,22].

By making an impact on all of these dimensions, exploring the influence of patient characteristics on LOS classification is one way that the health care industry can enhance patient outcomes and move toward more sustainable operations[23]. The use of patient characteristics to explain the effects of LOS has not been carried out, but this study is precisely a simple way to provide early intervention to promote sustainable hospital development. In this work, we proposed to solve the resource allocation problem, using a novel approach combining multi-attribute and multi-level thinking with k-means clustering. Our aim is to provide a LOS grading path based on the characteristics of patients for promote hospital sustainability. This method can not only reduce the requirement of data, but also meet the needs of hospital operation and management.

The following chapters of the paper are arranged: the second part gives the research data and methods, including data sources, factor analysis and sample clustering methods; The third part gives the empirical analysis process and results of nephropathy research, and tests the clustering results. The fourth part is the results and discussion; the fifth part is the conclusion.

2. DATA AND METHODS

2.1 Data Source and Quantitative Processing

In this paper, we use nephropathy patients’ data in the year 2016 from a large hospital in the Department of Nephrology to applying the estimation method. After eliminating the missing variables and abnormal samples, 3,556 samples were finally included. There were 5 variables including gender, age, medical insurance, patients type and LOS. In order to facilitate quantitative analysis, variables are assigned to be quantified, and the results of each variable are shown in table 1.

| Variable name       | The quantitative transformation results of variables                          |
|---------------------|------------------------------------------------------------------------------|
| Gender              | “Male”=“1”, “Female”=“2”                                                     |
| Age                 | “Ward daytime patient”=“1”, “ordinary patient”=“2”, “Special patients”=“3”, “Patients with blood resuscitates”=“4”, “Patients with tumor recovery”=“5”, “other patient”=“6” |
| Patients type       | “No medical insurance”=“1”, “New rural cooperative medical insurance”=“2”, “Urban rural medical insurance”=“3”, “City medical insurance”=“4”, “Provincial medical insurance”=“5” |
| Medical insurance   | Actual value                                                               |
| Length of stay      | Actual value                                                               |
2.2 Methods and Steps
(1) Extract common factors of inpatient characteristics. Factor analysis program in SPSS 24.0 software was used to extract the common factors of patient hospitalization characteristics based on four variables including gender, age, type of medical insurance and type of patient.
(2) According to the results of factor analysis, using k-means clustering to classify the patients.
(3) Using mean multiple comparisons to explore the importance of various patient characteristics to each major factor. The flow chart of the analysis method is shown in figure 1.

3. EMPIRICAL RESEARCH
The classification by gender, age, patient type and medical insurance type is shown in table 2. It shows the statistical characteristics of patient Los, where all patients are divided into different categories for each variable. We see that the proportion of women (55%) is higher than that of men (45%), and that the mean value of Los for women (7.84 days) is also higher than the mean value of Los for men (7.26 days). Patients aged 40-69 years accounted for 61.4%, and they consumed approximately 57.04% of total Los. The results of gender and age mean that older female patients may have longer Los than other patients. The average Los of patients with blood resuscitation was 14.26 days, which is the main force occupying the medical resources for a long time.

Table 2. Comparison of different gender, age group, patient type and medicare type Los

| Variable name | Variate meaning and description | Sample size | Mean | Std. deviation | Maximum | Minimum | Skewness | Kurtosis |
|---------------|---------------------------------|-------------|------|----------------|---------|---------|----------|----------|
| Gender        | Male (45%)                      | 1599        | 7.26 | 8.064          | 57      | 1       | 1.816    | 3.596    |
|               | Female (55%)                    | 1957        | 7.84 | 9.162          | 61      | 1       | 2.095    | 5.215    |
| Age           | <40years old (26.6%)            | 947         | 6.73 | 7.835          | 57      | 1       | 2.18     | 5.828    |
|               | 40-69years old (61.4%)          | 2183        | 7.44 | 8.714          | 61      | 1       | 2.118    | 5.530    |
|               | >70years old (12%)              | 426         | 10.21| 9.821          | 48      | 1       | 1.351    | 1.582    |
| Patients type | Ward daytime (0.5%)             | 19          | 14.11| 9.267          | 34      | 2       | 0.85     | -0.110   |
|               | Ordinary (1.5%)                 | 54(20.6%)   | 13.43| 8.341          | 42      | 1       | 0.662    | 1.097    |
|               | Special patient (0.6%)          | 2          | 7.5  | 7.778          | 13      | 2       | -        | -        |
|               | Patients with blood resuscitates| 88(2.5%)    | 14.26| 9.692          | 47      | 1       | 1.054    | 0.686    |
|               | Patients with tumor recovery    | 118(3.3%)   | 3.69 | 4.943          | 30      | 1       | -        | -        |
|               | Other (92.1%)                   | 3275        | 7.41 | 8.626          | 61      | 1       | 2.095    | 5.325    |
| Medical insurance | Provincial medical insurance | 1591      | 4.93 | 6.031          | 30      | 1       | 2.235    | 4.375    |
|               | City duty (12.1%)               | 430         | 13.87| 8.864          | 54      | 1       | 1.425    | 2.63     |
Table 3. Describes the statistical analysis

| Number of records | Effective | Missing |
|-------------------|-----------|---------|
| Urban rural       | 130 (3.7%)| 13.15   | 8.857  | 45 | 2 | 1.542 | 2.595 |
| New rural Cooperative medical insurance | 113 (3.2%) | 11.34 | 6.984 | 35 | 1 | 0.892 | 0.885 |
| No medical insurance | 1292 (36.3%) | 7.87 | 9.933 | 61 | 1 | 2.141 | 5.035 |

Table 3 shows the basic statistics of the patients. The patient's standard deviation was 8.689 days and the mean was 7.58 days, from which we could obtain a coefficient of variation of 114.63%. The positive difference between the mean (7.58 days) and the median (2 days) indicates that the data is qualitatively skewed, that the distribution to the right is the long tail, and that the quartile range between the 25th and 75th percentiles is 9 days. Figure 2 shows the highly skewed distribution of Los data, where the initial peak of the data is 2 days (approximately 51.6% of all patients) and most of the patients are less than 20 days (approximately 90.6% of all patients). It can be concluded that 73.8% of patients with Los are 10 days, 16.8% of patients with Los are longer, and 9.4% of patients with Los greater than 20 contribute to the distribution of the long tail to the right. Obviously, Los longer patients are going to be taking up resources, and those resources are going to be allocated to new patients.
3.1 Factor Analysis Results

| Table 4. KMO and bartlett tests |
|--------------------------------|
| Quantity of suitability KMO sampling | .682 |
| The approximate chisquare Bartlett test for sphericity | 81.392 |
| Degrees of freedom | 6 |
| Significant | .000 |

As can be seen from table 4, KMO value is 0.682, greater than 0.5, sig value is less than significance level, and there is a correlation between variables, indicating that the original variables can be used for factor analysis.

| Table 5. Variance of common factors |
|-------------------------------------|
| Variable name | The initial extract |
| Gender | 1.000 | .990 |
| Age | 1.000 | .868 |
| Patients type | 1.000 | .621 |
| Medical insurance | 1.000 | .678 |

It can be seen from table 5 that the table of variance of common factors indicates the degree to which each factor can be extracted by common factors. The degree to which gender, age, patient type and medical insurance type can be extracted by common factors is 0.990, 0.868, 0.621 and 0.678, respectively, all of which are greater than 0.5, indicating that most of the information in variables can be extracted by common factors.

| Table 6. Total variance explanation table |
|------------------------------------------|
| Composition | Initial eigenvalue | Extract the sum of the squares of the loads | Sum of the squares of the rotating loads |
|--------------|-------------------|-----------------------------------------------|----------------------------------------|
|              | Total | Percentage variance | Cumulative % | Total | Percentage variance | Cumulative % | Total | Percentage variance | Cumulative % |
| 1            | 1.134 | 28.352 | 28.352 | 1.134 | 28.352 | 28.352 | 1.123 | 28.072 | 28.072 |
| 2            | 1.039 | 25.984 | 54.336 | 1.039 | 25.984 | 54.336 | 1.031 | 25.779 | 53.85 |
| 3            | 0.983 | 24.584 | 78.921 | 0.983 | 24.584 | 78.921 | 1.003 | 25.07 | 78.921 |
| 4            | 0.843 | 21.079 | 100 | 0.843 | 21.079 | 100 | |

Principal component analysis (pca) was used to extract common factors, and the total variance interpretation table showed the contribution rate of extracted factors. The cumulative contribution rate of the first three factors was 78.921%, which could explain most of the information, so the first three factors were extracted as the main factors.

| Table 7. Composition matrix after rotation |
|-------------------------------------------|
| Composition | 1 | 2 | 3 |
| Gender | 0.009 | 0.995 | 0.01 |
| Age | 0.002 | 0.931 | 0.015 |
| Patients type | -0.718 | 0.077 | -0.316 |
| Medical insurance | -0.253 | 0.084 | 0.779 |

According to table 7, on the first major factor, the patient type is larger than the load value of other factors, and is classified as the first category. On the second main factor, the load value of gender and age is large, indicating that these two indicators are strongly correlated and can be classified into the second category. On the third main factor, the type of medical insurance has a large load value and can be classified into the third category. On the fourth major factor, the load value of hospitalization days is large, which can be classified into the fourth category.

Through the results of factor analysis, the four variables were refined into three relatively independent comprehensive factors, as shown in table 8.
Table 8. Principal factor naming table

| Variable                  | The main factor of 1 | The main factor of 2 | The main factor of 3 |
|---------------------------|----------------------|----------------------|----------------------|
| Principal factor name     | Disease characteristics | Attribute characteristics | Reimbursement ratio characteristics |

3.2 Cluster Analysis

K-means clustering was adopted to cluster 3556 samples into 5 categories. The specific clustering results are shown in table 9. Where A, B and C represent the importance of each category to the principal factor.

Table 9. Clustering results

| Category | Category 1 | Category 2 | Category 3 | Category 4 | Category 5 |
|----------|------------|------------|------------|------------|------------|
| (36)     | (624)      | (974)      | (314)      | (1608)     |
| The main factor of 1 Disease characteristics | A | B | C | B | C |
| The main factor of 2 Attribute characteristics | B | B | C | B | C |
| The main factor of 3 Reimbursement ratio characteristics | C | B | B | B | B |

According to table 9, category 1 patients attach more importance to main factor 1 than the other two main factors. These patients are basically elderly and female patients with serious diseases, and attach less importance to the reimbursement ratio of medical insurance. Category 2 and category 4 patients attach the same importance to the three main factors, which are generally middle-aged patients suffering from common diseases, and will consider the reimbursement rate of medical insurance. Category 3 and category 5 patients have the same degree of attention to the three main factors, and this type of patients are basically young men with less severe disease.

3.3 Test of Clustering Results

One-way anova was used to test the clustering results and test the mean difference. According to table 10, all p values are less than the significance level (given significance level 0.05), indicating the inhomogeneity of variances.

Table 10. Homogeneity of variance test results

| Gender         | Degrees of freedom 1 | Degrees of freedom 2 | Significant |
|----------------|----------------------|----------------------|-------------|
| Gender         | 61.220               | 4                    | 3551        | .000        |
| Age            | 31.181               | 4                    | 3551        | .000        |
| Patients type  | 648.652              | 4                    | 3551        | .000        |
| Medical insurance | 487.815             | 4                    | 3551        | .000        |

It can be seen from table 11 that the p values are all lower than the significance level, indicating that each variable has a significant impact on the clustering results, both inter-group and intra-group. Therefore, the clustering results are significant.

Table 11. Test results

| Gender         | Sum of squares | Degrees of freedom | The mean square | F     | Significant |
|----------------|----------------|--------------------|-----------------|-------|-------------|
| Between groups | 103.365        | 4                  | 25.841          | 118.155 | .000        |
| Within the group | 776.625        | 3551               | .219            |       |             |
| Total          | 879.990        | 3555               |                 |       |             |
### 4. RESULTS AND DISCUSSION

According to the sample data of patients, patients were divided into 5 categories: the number of category 1 is 36, the number of category 2 is 624, the number of category 3 is 974, the number of category 4 is 314, and category 5 is 1608. According to the average LOS, the LOS of patients is categorized into three levels (Represented by A, B, C): A (LOS $\geq$ 15), B (10 < LOS < 15), and C (LOS $\leq$ 10). Other characteristics of the patient are shown in Table 12. According to the attributes and other variables of the patients, the LOS of the patients was simulated as shown in Figure 3.

#### Table 12. Characteristics of various patients

| Category | Sample | Average LOS (days) | Average Age | Attribute characteristics | Disease characteristics | Reimbursement Ratio characteristics | Los level |
|----------|--------|--------------------|-------------|---------------------------|------------------------|------------------------------------|-----------|
|          |        |                    | Male        | Femal                     | Patients type          | Medical insurance                  |           |
| Category 1 | 36    | 15.78              | 70.67       | 38.90% 61.60%            | 38.2% of the patients with blood resuscitates, 30.3% of the patients with tumor recovery | 50% of the City medical insurance | A         |
| Category 2 | 624   | 10.5               | 53.90       | 54.80% 45.20%            | 62.5% of the Ordinary patient | 79.2% of the City medical insurance, 20.8% of the others medical insurance | B         |
| Category 3 | 974   | 7.62               | 39.33       | 59.40% 40.60%            | 99.7% of the Other patient | 73.1% of the provincial medical insurance, 26.9% of the others medical insurance | C         |
| Category 4 | 314   | 13.7               | 60.49       | 43.80% 56.30%            | 38.9% of the Ward daytime patient, 31.8% of the Ordinary patient | 60.6% of the Urban rural medical insurance, 21.2% of the New rural cooperative medical care, the rest is other health insurance | B         |
| Category 5 | 1608  | 6.73               | 37.71       | 67.40% 32.60%            | 99.3% of the Ordinary patient | 54.4% of the provincial medical insurance, 30.7% of the other medical insurance | C         |

According to Figure 3, according to the attributes and other characteristics of the patient, the distribution of hospitalization after the patient arrives at the hospital is simulated. There are the
following 7 paths. Due to the characteristics of the patient, the entry path is different, resulting in the length of hospitalization of the final patient may be different.

According to the results of factor analysis and clustering, Patients in this sample were divided into 5 categories and 3 levels. Class A inpatients are characterized by serious diseases of elderly women, and most of them are covered by urban occupational insurance. Such patients are more likely to use inpatient services[24], which may be due to the difference in the physical conditions of the age group of 60-79[25] of different genders[26] among their aged peers. In addition, the severity of the disease and possible complications lead to an increase in LOS. Class B inpatients are characterized by common diseases of middle-aged men and women, and the reimbursement ratio of medical insurance is more important. In this category of patients, because they are in middle age and have better physical recovery ability than old age, the reimbursement ratio of medical insurance will be considered when the disease is not very serious. Grade C inpatients are characterized by mild young male diseases. Such patients have strong physical recovery, but their condition is not serious. Therefore, in the case of patients’ personal preference, they do not care much about the reimbursement rate of medical insurance. Considering the attributes of patients and combining with other variables of patients, the simulation of the Los distribution of patients has important practical significance for the optimization of resource scheduling.

5. CONCLUSION
In this study, on the one hand, the data-driven los attribute recognition method is presented. On the other hand, the multi-attribute los multi-level is given. After the test of clustering results, the multi-attribute multi-level los is effective. The limitation of this study lies in the limited index dimensions included in this study, and the research results have certain limitations. The results of hospitals in different regions or cities and at different levels may be different, so they only reflect the results of this sample. In the next step, multi-region, multi-hospital and multi-dimensional sample data can be further collected to further study the influence of other factors on Los to explain the predictors.

REFERENCES
[1] Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academies Press; 2001.
[2] HPOE. Environmental Sustainability in Hospitals: The Value of Efficiency. Hospitals in Pursuit of Excellence; 2014, 1-34.
[3] Malliga Marimuthu, Hanna Paulose. Emergence of Sustainability Based Approaches in Healthcare: Expanding Research and Practice[J]. Procedia - Social and Behavioral Sciences, 2016, 224.
[4] Pantartzis E, Edum-Fotwe F T, Price A D F. Sustainable healthcare facilities: Reconciling bed capacity and local needs[J]. International Journal of Sustainable Built Environment, 2017:S2212609016300838.
[5] Nguyen M H B, Rivers E P, Suzanne Havstad M A, et al. Critical Care in the Emergency Department A Physiologic Assessment and Outcome Evaluation[J]. Academic Emergency Medicine, 2000, 7(12):1354-1361.
[6] Chalfin D B, Trzeciak S, Likourezos A, et al. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit*[J]. Critical Care Medicine, 2007, 35(6):1477-1483.
[7] Lucas R, Farley H, Twanmoh J, et al. Emergency Department Patient Flow: The Influence of Hospital Census Variables on Emergency Department Length of Stay[J]. Academic Emergency Medicine, 2009, 16(8):597-602.
[8] Caccialanza R, Klersy C, Cereda E, et al. PP140 NUTRITIONAL PARAMETERS ASSOCIATED WITH PROLONGED HOSPITALIZATION IN AMBULATORY ADULT PATIENTS[J]. Clinical Nutrition Supplements, 2010, 5(2):78-78.
[9] Malliga Marimuthu, Hanna Paulose. Emergence of Sustainability Based Approaches in Healthcare: Expanding Research and Practice[J]. Procedia - Social and Behavioral Sciences, 2016, 224.
[10] Dunphy J L . Healthcare professionals’ perspectives on environmental sustainability[J]. Nursing Ethics, 2013, 21(4):414.

[11] Higuchi K S , Downey A , Davies B , et al. Using the NHS sustainability framework to understand the activities and resource implications of Canadian nursing guideline early adopters[J]. Journal of Clinical Nursing, 2012, 22(11-12):1707-1716.

[12] Nambiar, R., et al. A look at challenges and opportunities of Big Data analytics in healthcare. 2013: IEEE.

[13] Michael J. Ward,Keith A. Marsolo,Craig M. Froehle. Applications of business analytics in healthcare[J]. Business Horizons,2014,57(5).

[14] Estimating Patient’s Length of Stay in the Emergency Department with an Artificial Neural Network[J]. Proceedings of the AMIA Annual Fall Symposium,2005,2005.

[15] Michael Rowan,Thomas Ryan,Francis Hegarty,Neil O’Hare. The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors[J]. Artificial Intelligence In Medicine,2007,40(3).

[16] Michael Rowan,Thomas Ryan,Francis Hegarty,Neil O’Hare. The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors[J]. Artificial Intelligence In Medicine,2007,40(3).

[17] Hachesu Peyman Rezaei,Ahmadi Maryam,Alizadeh Somayyeh,Sadoughi Farahnaz. Use of data mining techniques to determine and predict length of stay of cardiac patients.[J]. Healthcare informatics research,2013,19(2).

[18] Hachesu Peyman Rezaei,Ahmadi Maryam,Alizadeh Somayyeh,Sadoughi Farahnaz. Use of data mining techniques to determine and predict length of stay of cardiac patients.[J]. Healthcare informatics research,2013,19(2).

[19] L. Lella, A. di Giorgio, and A. F. Dragoni, Length of Stay Prediction and Analysis through a Growing Neural Gas Model, 2015.

[20] Igor Kononenko,Se June Hong. Attribute selection for modelling[J]. Future Generation Computer Systems,1997,13(2).

[21] Yvan S , Inza Inaki, Larrañaga Pedro. A review of feature selection techniques in bioinformatics[J]. Bioinformatics,2007,(19):19.

[22] Duggal R , Shukla S , Chandra S , et al. Impact of selected pre-processing techniques on prediction of risk of early readmission for diabetic patients in India[J]. International Journal of Diabetes in Developing Countries, 2016, 36(4):469-476.

[23] Hamed M. Zolbanin,Behrooz Davazdahemami,Dursun Delen,Amir Hassan Zadeh. Data analytics for the sustainable use of resources in hospitals: Predicting the length of stay for patients with chronic diseases[J]. Information & Management,2020.

[24] Ren Jiaojiao,Ding Ding,Wu Qunhong,Liu Chaojie,Hao Yanhua,Cui Yu,Sun Hong,Ning Ning,Li Ye,Kang Zheng,Shan Linghan,Zhao Miaomiao,Liu Baohua. Financial Affordability, Health Insurance, and Use of Health Care Services by the Elderly: Findings From the China Health and Retirement Longitudinal Study.[J]. Asia-Pacific journal of public health,2019,31(6).

[25] Chun Luo.Re-partitioning Population Age Group and Its Implications[J].Population Research,2017,41(05):16-25.

[26] Youhua Chen,Su Xu. Health Status, Welfare Demand and Prospects of Elderly Population in China[J].Journal of population,2011(02):34-39.