Economic Impact Analysis of Hospital Readmission Rate and Service Quality Using Machine Learning

bailek, Alexandra

University of Windsor

11 October 2018

Economic Impact Analysis of Hospital Readmission Rate and Service Quality Using Machine Learning

Abstract: The hospital readmission rate has been proposed as an important outcome indicator computable from routine statistics. The purpose of this research is to investigate the Economic Impact of service in hospitals and integrated delivery networks in the United States based on the readmission rates as the target variable. The data set includes information from 130 hospitals and integrated delivery networks in the United States from 1999 to 2008 to investigate significance of different factors in readmission rate. The dataset contains 101,766 patients' encounters and 50 variables. The 30-day readmission rate is considered as an indicator of the quality of the health providers and is used as target variable in this project. Preliminary data analysis shows that age, admission type, discharge disposition etc. is correlated to the readmission rate and will be incorporated for further data analysis. Data analysis are performed on the diabetic patient dataset to develop a classification model to predict the likelihood for a discharged patient to be readmitted within 30 days. KNN, Naive Bayes and Logistic Regression algorithm were used to classify data and KNN appears to be the best approach to develop the model. Hospitalisations and drug prescriptions accounted for 50% and 20% of total readmission expenditure, respectively. Long term nursing home care after hospital admission cost an additional £46.4 million. With the ability to identify those patients who are more likely to be readmitted within 30 days, we can deploy the hospital resources more economically affordable while improving services. Based on the results it can be concluded that the direct cost of readmission rate for hospitals rose to £459 million in 2000 and nursing home costs rose to £111 million. Also, it can be perceived that a reduced length of hospital stay was associated with increased readmission rates for jaundice and dehydration.

Keywords: Predictive Modeling, Re-admission, Simulation, Healthcare Service, Classification Modeling, Health care Quality Indicator

1. Introduction

Services are the most visible function for healthcare organization and different parameters are considered to measure the service level of the healthcare. Assessing service level is a major challenge in health care organization as different variables can be linked to show the result. For instance, waiting time in hospital, early re-admission rate, average cost per discharge, and occupancy rate are some of the commonly used variables to estimate the service level of hospital. Many hospitals have started to collect wide variety of data to measure the service level but it is very difficult to interpret result from the large set of raw data. Similarly, this project involves the large set of data collected by Center for Clinical and Translational Research, Virginia Commonwealth University from 130 hospitals for a period of 1999 to 2008. The given data comprise of 50 different variables with 101766 observations and it is difficult to identify target variable, and other key variables that have a significant interaction with the target variable. Further, developing model to evaluate the relation of target variable with other key variables and link the obtained result with the service level of hospital was a major problem associated with the project.

The balance between improving hospital efficiency while simultaneously improving quality of care is a priority in our health care system. However, a marker of efficiency, hospital length of stay (LOS), may conflict with a marker of quality: hospital readmission rates. The balance between improving hospital efficiency while simultaneously improving quality of care is a priority in our health care system. However, a marker of efficiency, hospital length of stay (LOS), may conflict with a marker of quality: hospital readmission rates.To analyze the given data, certain assumptions were made, such as elimination of less significant variables to target variable, variables with large missing values and variables with higher distinct levels. Also, readmission rate, which was identified as the target variable, with three levels of data was categorized into two groups, one within 30 days of readmission and other with no readmission within 30 days which are discussed in detail in section 2. In addition, 30248 duplicate data were not taken into consideration for the classification algorithm.

Readmission within 30 days for medical conditions is common and costly. In studies of Medicare patients, 30-day readmission rates range from 8% (8) to 21%, depending on diagnosis, with annual estimated costs of $17.4 billion. The primary objective of this project is to show linkage of early

readmission rate with other key attribute variables like admission type, discharge disposition, age etc. which is achieved by developing a classification model. This model shows whether the patient with certain type of information for input variables are likely to be readmitted within 30 days or not. Hence, this prediction might help healthcare organizations to take several precautions to improve the service level of the hospital with limited resources.

2. Literature Review

A cornerstone of proposed health care reforms in the United States is competition among providers based on cost and quality [1]. This push for competition has increased interest in external quality indicators that consumers or designated purchasing cooperatives could use to choose providers. Risk adjusted outcome measures, such as mortality and early re-admission rates, have been proposed as quality indices for use by large systems of hospitals and buyer [3-6]. In a survey of 250 Fortune 1000 companies, the benefits managers replied that they used hospital re-admission rates as a measure of provider quality in their health plans more often than any other measure.12 Blue Cross/Blue Shield of Michigan has considered using risk-adjusted hospital outcomes, including mortality and re-admission rates, to adjust hospital payment. Because ERRs have been commonly tabulated and reviewed by health care organizations for many years, why is so little known about their validity? Assessing the performance of quality indicators is difficult, in no small part of due to the absence of widely accepted gold standards for defining poor-quality events [7-8].

Validation studies also are expensive and time consuming, and require large sample sizes. Yet, we might expect assessments of the feasibility of ERRs as performance measures to be undertaken before promoting their use. How large do the quality differences need to be, how many patient discharges, hospitals, and years of data are needed, and how good does the case-mix adjustment need to be for it to be feasible to use hospital ERRs as an accurate measure of quality? A review of the literature revealed no studies that addressed these questions.

To examine these issues, we developed a simulation model for evaluating potential quality indicators in hospital systems of differing size and quality. In a simulation, as in an experiment, we can modify the assumptions regarding the impact of quality- and non-quality-related factors on variations in hospital ERR differences, and examine the importance of adequate case-mix adjustment. These factors are almost impossible to manipulate experimentally, making simulation a particularly useful tool for this. Simulation models are well established for use in many fields [9] with different approaches for use in predictive modeling, policy analysis, and planning. Evolutionary algorithms have also been used to solve complex data analytical problems [10]. A novel way of predictive modeling has been employed using system dynamics modeling which extensively has been utilized in many fields [11-12].

2. Data Description and Preparation

The data set consists of 50 variables with 101766 total observations and each entry provides a patient information about age, gender, race, discharge nature, type of diagnosis, medications used, readmission rate etc. Both numerical and categorical variables can be found in the data set with some variables have missing values more than 50%.

The primary objective of the project is to determine quality or service of health service based on the given data set. From our literature review and brainstorming, readmission rate was selected as the target variable because it can be easily linked to other key input variables like age, discharge disposition, admission source etc. and ultimately expressed to service level. Also,

research articles have shown that 30 days readmission rate are generally used to examine the quality of the health care [1]. Therefore, it was decided to use early readmission rate as a target variable and categorize it into two groups, one within 30 days of readmission and other with no readmission within 30 days. That means the data sets with patients who are not admitted or readmitted after 30 days of their treatment are grouped together to create a distinct level of no readmission within 30 days. Additional analyses expanded the models for LOS and readmissions to control for patient variables described previously as well as the admitting hospital, in accordance with previous studies based on administrative data.

Also, the variables with large missing data, variables with higher distinct levels, and variables that are less significant to early readmission rate are eliminated because it might only complicate the interpretation of result in the later phase. For instance, time in hospital, diagnosis 1, 2, 3 etc., and variables like race, gender, number of lab procedures, encounter ID are removed from the analysis due to higher distinct levels and less significant nature to early readmission rate respectively. Further, variables like weight, payer code and medical specialty have missing values greater than 50% of the total data set and are not taken into consideration for further analysis. In addition, recent research have shown that variables like age, admission type, discharge disposition, admission source can be strongly associated with readmission rate which helped to finalize the ultimate variables for an analysis using R, Microsoft Excel and Minitab[2].

While observing the general trend of the data, 30248 duplicate entries were found with some patients visited hospital for the multiple times. Also, while analyzed, 11357 patients out of 101766 were readmitted within 30 days after their treatment which is 11.2% of the total data set. And, it cannot be ignored that more than 50% of the data set relate to emergency cases. Further analysis have shown 53990 cases in admission type ID, and 57494 cases in admission source ID were related to emergency cases. Similarly, when discharge disposition data were observed, majority of the data were associated to discharge to home despite it contains 30 distinct levels. Also, while comparing patient readmission rate with age group, readmission rate was particularly recognized higher at the elderly age group with 70-80 age group as the most significant one followed by 60-70 and so on. More trends and results can be concluded through detailed analysis which are described in later sections.

3. Modeling Approach

The following is the methodology and modeling approach in analysis of data set ‘diabetic_data’. The general analysis of the data set ‘diabetic_data’ was performed in order to investigate the importance of readmission rate as our target variable in the data set. In general, the data set includes information from 130 hospitals and integrated delivery networks in the United States from 1999 to 2008. This point needs to be mentioned that not all 55 existed variables are used in our analysis since some of the variables contain huge number of missing values that incapacitate us to extract useful information from them. Besides that, there were some variables that, though didn't contain missing values, were cumbering in using them in classification or clustering. Hence, we decide to

eliminate them from the data set ‘diabetic_data’. More explanation in terms of our assumptions is going to be presented later on the modeling approach.

Early Readmission Rate is considered as our target variable since it is a reliable pattern to measure the effectiveness the service in the hospitals. In the process of evaluating the data set ‘diabetic_data’, we have utilized different software such as R Studio, Excel, and Minitab. R Studio was conducted as one of our main tool to perform analysis and obtain useful information, as well as the general behavior of the dataset. R Studio was primarily used for classification, clustering,and logistic regression. In terms of classification we have performed ID3 algorithm. By doing so, we came to know that ID3 algorithm didn't provide our desired tree because the levels for some variables were so high that the classification became useless. Therefore, we decided to apply KNN and Naive Bayes to classify the data to see which one could be better applied to our work. By doing so, the classification became more solid and made more sense. More details on comparison of these algorithm in terms of classification is going to be elaborated in analysis and results section. In terms of clustering, we have used K-means algorithm to identify the best number of main clusters regarding our target variable. Further, logistic regression was utilized to investigate the correlation between early readmission rate and other independent variables since most of our selected variables contain categorical data rather continuous data. Also, Excel was primarily used to plot visual graphs in terms of statistical relationship between readmission and other independents variables. Several charts in terms of ‘diabetic_data’ variables such as number of people readmitted to the hospitals, the age range readmitted to the hospitals, and number of people discharged to home after readmission, etc was also visualized using Excel. Moreover, Minitab was utilized in order to investigate regression models and possible correlation between the target variable (readmission) and other independent variables.

In the phase of extracting useful information from the ‘diabetic_data’, we have utilized different functions to extract useful information in RStudio from the ‘diabetic_data’ in terms of readmission and important variables that affect our target variable. Some of the useful functions that we could elucidate central tendency, number of variables, and the number of levels from the dataset were str(), summary(), levels(), sum(), mean(), median(), quantile(), and etc. In that process, we have found the central tendency, the number of variables, and the number of levels under each variable, and the total number of observations etc.

In the second phase of analyzing the ‘diabetic_data’, we have classified the target variable using KNN. Results of using KNN algorithm multiple times according to Figure 1 below show that K=8 is the best number for classification.

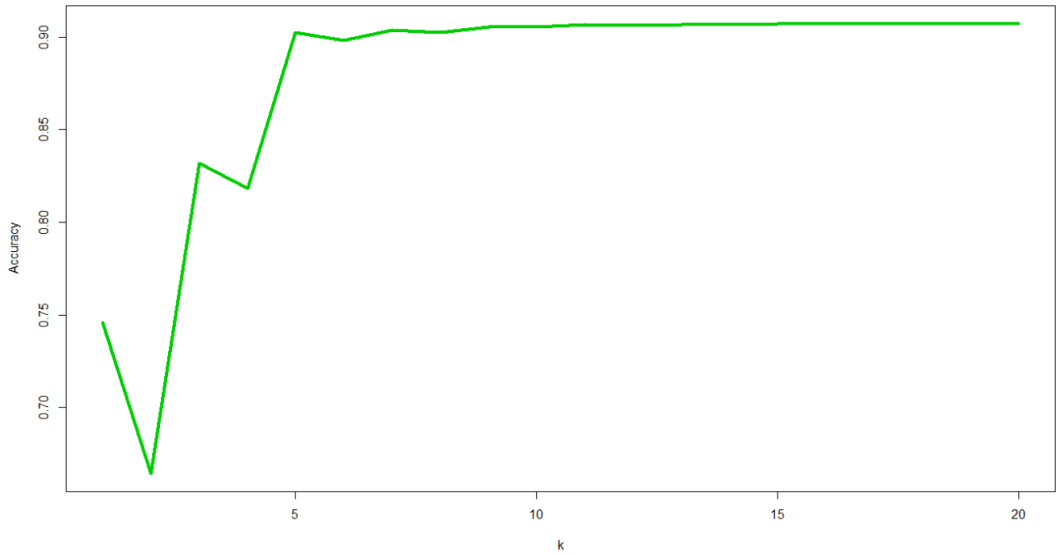

Figure 1. Accuracy curve of kNN algorithm for different k values

We also have tested different scenarios to investigate which K is the best one. As we tried different number of Ks, we came to know that after K=8 the slope line is going in constant way; therefore, K=8 is chosen as the best K number in KNN algorithm. Moreover, we have utilized K-means algorithm for clustering. The analysis of K-means algorithm shows that according to the majority rule, the best number of clusters is 2 which is the best in clustering the data set. We also have tried different training set and test set in order to find the best accuracy and to come up with the best prediction of ‘diabetic_data’. We have started from training set =50% and test set =50% as our baseline and continue to find the best match in terms of our data set. It turned out that training set =80% and test set =20% is the best match for achieving the best prediction of ‘diabetic_data’.

As we mentioned before, for analyzing the ‘diabetic_data’ we consider some assumptions in order to have a better perspective of the dataset. First of all cases, we have assumed early readmission rate as our target variable since early readmission rate is a good indicator to evaluate how effective the hospital's service were. Moreover, this target variable is classified into two subsets of readmitted patients within 30 days, and patients not readmitted to hospitals within 30 days. The latter includes both patients readmitted after 30 days and those who are not readmitted to hospitals

at all. Also, the assumptions for Logistic regression are: The target variable should be measured on a dichotomous scale; there are one or more independent variables, which can be either continuous or categorical.

Secondly, not all variables in terms of readmission rate can help us to elucidate useful information out the 'Diabetic_Data'. Therefore, we have condensed the variables to the most important ones which include: patient_nbr, age, admission_type_id, discharge_disposition_id, admission_source_id, number_emergency, diabetesMed, readmitted. Other variables either contain many missing data or they didn't contain useful information that we could extract from. The variables that contain more than 50% missing values were eliminated from the data set. For example, Weight as a variable could help us investigate if there is any relationship between numbers of diabetics readmitted patients and their Weights, but since 97% of data of weight is missed this variable is eliminated from the data set. Besides, Payer code and Medical specialty contain more than 50% of missing values; therefore, they are eliminated from the data set as well. There are other kinds of variables that we consider ignoring them. Although they didn't consist of any missing data, they contain meaningless values or variables that were hard to classify them. Further, there are about 30,000 people in the dataset who readmitted to the hospitals more than once. These duplicated records were eliminated from the dataset in order to reduce bias and have a better classification of readmission rate as our target variable.

4. Estimate of the Readmission Demand to Hospital

This section presents estimates of several variants of the admission and mean stay equations. These estimates serve three distinct purposes. First, they establish the importance of the traditional role for price as a determinant of the demand for hospital care. Second, they examine the effect on demand of the availability of physicians and hospitals beds. Third, they provide parameter values for use in the next section to test and estimate the price adjustment mechanism. The equations are estimated by an instrumental variable procedure that yields consistent parameter estimates [14]. Although a separate constant term for each state is not included, the lagged dependent variables are treated as endogenous to obtain consistent estimates even if the disturbances contain a systematic "state effect" or are otherwise serially correlated.

4.1 Estimated Cost of Community Based Health Care

About 534 000 people (281 000 men and 253 000 women) were being treated for atrial fibrillation (AF) in the UK in 1995—that is, 0.9% of the whole population and 5% of those > 65 years. There were about 590 600 male and 984 800 female consultations with general practitioners for AF in 1995, costing £29.5 million. We assumed that one quarter of patients with AF (about 134 000 people) attended for an average of two hospital outpatient visits in 1995, costing £28.8 million.

4.2 Estimated cost of hospital Readmission

Hospital admissions (principal diagnosis) There were about 31 000 hospitalisations in 1995, involving 26 000 men, associated with a principal diagnosis of AF (these accounted for a total of 193 000 bed-days). For women, the equivalent figures were 28 000 hospitalisations, 24 000 people, and 159 000 bed-days. Figure 1 shows the outcome of the patients who were hospitalised. By stratifying costs according to the proportion of bed-days spent in different types of wards (general or specialty), we calculated that the total expenditure on AF hospitalisations was £121.7 million in 1995. The greatest expenditure was for men aged 55–64 years (£20.6 million) and for women aged 75–84 years (£17.1 million).

5. Results:

An initial analysis of the dataset concerning different age groups and corresponding readmission rates is shown in the following figure 2.

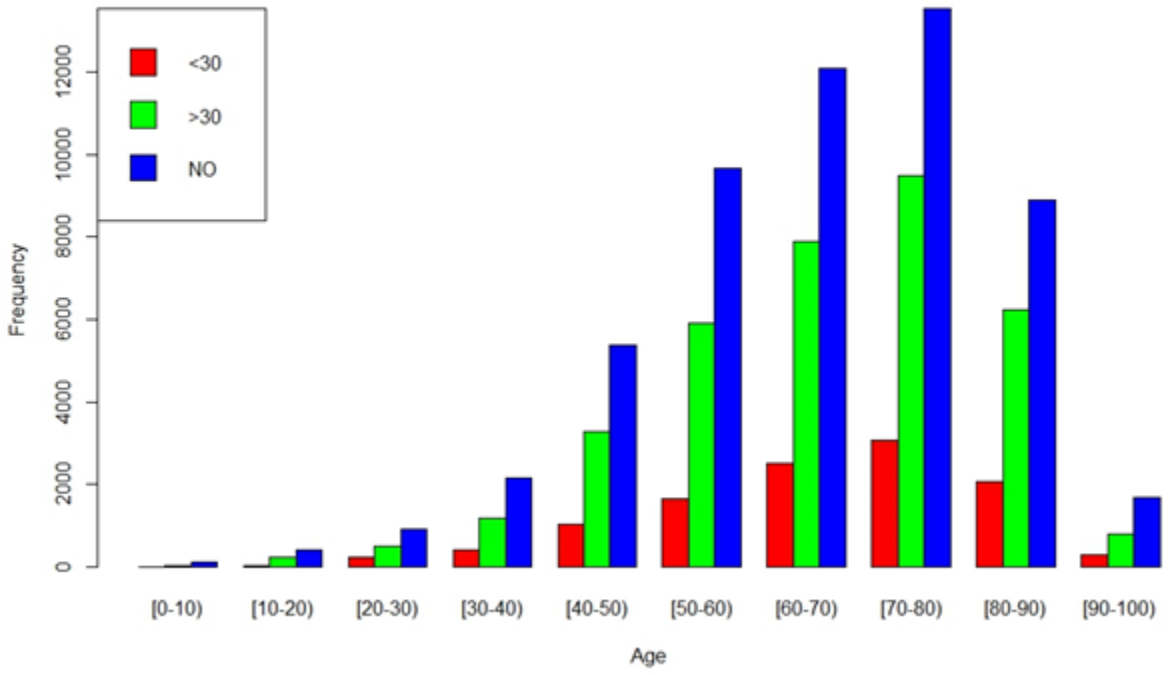

Figure 2. Initial grouped bar plot of age and readmission rates

KNN and Naive Bayes and binary logistic regression algorithm were used to classify data so that predictions can be made on the new observations. KNN is a non-parametric method used for classification and K needs to be chosen carefully during analysis. If K is overly high, the algorithm can be too sensitive to noise and overfit the new observations. If K is overly low, the algorithm can lose the true pattern of the data. It is found that the best fitting K equals 8 during analysis as explained above.

Naive Bayes doesn't require to use a predetermined parameter such as K, but it assumes that each attribute is conditionally independent of every other attribute giving the class label. It is also sensitive to correlated variables and will double count the effects if we incorporate two or more attributes that are correlated. Naive Bayes approach provided good results in terms of accuracy and other performance criteria. It is to be mentioned that the same training set and test set were used across all algorithms so that the comparison is not biased. In the appendix section, the details of both training and test sets can be obtained.

Binary logistic regression is a regression model where the dependent variable is binary (0 or 1) in nature. Since we have two levels for the target variable, binary logistic regression was used to classify the data and predict it for the test sets.

Accuracy, false positive rate (FPR), false negative rate (FNR), true positive rate(TPR) and precision are typically used to evaluate the effectiveness of the classification models. The performance metrics for KNN, Naive Bayes and Logistic Regression methods were calculated and listed in Table 1.

Table 1. Performance metrics for the KNN, Naive Bayes and Logistic Regression algorithms

Performance criteria	KNN	Naïve Bayes	Binary Logistic Regression
Accuracy	90.44%	87.76%	90.52%
True Positive Rate	36.14%	16.67%	23.08%
False Negative Rate	63.86%	83.33%	76.92%
False Positive Rate	9.15%	9.02%	9.32%
Precision	2.93%	7.72%	0.59%

The very low values of precision might seem odd. In fact, we defined the level "<30" (0 for binary logistic regression) to be positive and "NO" (1 for binary logistic regression) to be negative. If we had defined otherwise, the precision values would increase. But the accuracy would remain the same. That is why we used accuracy as the principal indicator for the algorithm's success.

6. Discussion and Interpretation of Data

Table 1 shows that the accuracy for Logistic regression is the highest among three, although the precision is not as good as the other two models. False Negative Rate appears to be higher than 50% for all three models. In this project, however, the FNR measures the percentage of those patients who are incorrectly classified as low risk for Early Readmission. The hospital may not

allocate additional resource to these patients to reduce the risk by following up more frequently or keep the patients in the hospital for longer period of time. But the consequences associated with FNR are not catastrophic since the hospital and emergency room are accessible to these patients at anytime after they are discharged.

The binary logistic regression analysis shows that the Early Readmission Rate (ERR) is correlated with age, discharge disposition and number of emergencies. The P-values for these variables are less than 0.001 as shown in table 2. Although the hospital has no control over the age of the patients and number of emergencies, it does make decision on where the patients should be discharged to after the surgical procedures or treatments.

Table 2.Output from the ANOVA of binary logistic regression analysis

Coefficients	**Estimate**	**Std. Error**	**Z Value**	**Pr(>|z|)**
(Intercept)	3.129315	0.086999	35.970	< 2e-16 ***
age	-0.101369	0.011037	-9.184	< 2e-16 ***
admission_type_id	0.017668	0.011271	1.568	0.117
discharge_dispo_id	-0.033859	0.002743	-12.343	< 2e-16 ***
admission_source_id	-0.001377	0.004062	-0.339	0.735
number_emergency	-0.126594	0.028119	-4.502	< 6.73e-06 ***

Further analysis can be performed on discharge disposition to determine which groups (for example, patients discharged to ICF or patients discharged to another inpatient care institution) have higher risk of Early readmission. Once those groups are identified, more resources can be allocated to those group of patients to improve medical service.

There are several possible explanations for the finding of improvement in LOS without an adverse effect on 30-day hospital readmission. First, the VA health care system may have had inefficiencies in care that resulted in prolonged hospitalizations beyond what was needed for the

care of patients. Thus, a reduction in LOS would not lead to a premature discharge resulting in hospitalization. Second, the VA initiated a national effort in 2006 to improve hospital flow (for example, LOS and discharge time of day) with the Flow Improvement Inpatient Initiative. By evaluating the performance metrics of these methods, KNN appears to be the best approach although it requires some efforts to determine K. The KNN model allows us to make predictions on new observations (test set) with the accuracy of 90.44%. If a patient is identified as a high-risk patient who is likely to be readmitted within 30 days, more resources will be allocated to the patient to reduce the potential risk. Although, the binary logistic regression shows slightly higher accuracy, it is not completely unbiased. The threshold value is assumed to be 0.8 (It means that the algorithm will assign 1 to the probabilities which are greater than 0.8 and 0 otherwise.) which is highly dependent on the training set and the entire data (or its subsets).

7. Conclusions and Next Steps

The hospital readmission rate has been proposed as an important outcome indicator computable from routine statistics. Despite these limitations, we have shown that hospitals readmission rate imposes a substantial economic as well as a health burden on Hospitals in the USA with low service quality. Adjusted rates of potentially avoidable readmissions are scientifically sound enough to warrant their inclusion in hospital quality surveillance. Data analysis are performed on the diabetic patient dataset to develop a classification model to predict the likelihood for a discharged patient to be readmitted within 30 days. KNN, Naive Bayes and Logistic Regression algorithm were used to classify data and KNN appears to be the best approach to develop the model. With the ability to identify those patients who are more likely to be readmitted within 30 days, we can deploy the hospital resources more efficiently while improving services. The correlation between clearly and potentially avoidable readmissions rates argues for its use to monitor the quality of the discharge process, especially in the context of resource constraints.

Further analysis might be conducted in terms of other classification algorithms. The readmission rate interval (30 days) might be changed for a better prediction. But it will require some efforts to collect the data. Some variables may be defined to relate service quality directly to the early readmission rate and make the prediction much better.

6. References

[1] Hofer, T. P., & Hayward, R. A. (1995, March). Can Early Re-Admission Rates Accurately Detect Poor-

Quality Hospitals? *Journal Storage (JSTOR), 33*(3), 234-245. Retrieved April 13, 2017, from http://www.jstor.org/stable/3766825

[2] Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014, April 3). Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International*, 1-11. doi:https://www.hindawi.com/journals/bmri/2014/781670/

[3] Rontal R, Kiess MJ, DesHarais S, Reutter K. Applications for risk-adjusted outcome measures. Qual Assur Health Care 1991; 3:243.

[4] Brook RH, Lohr KN. Monitoring quality of care in the Medicare program. JAMA 1987; 258:313.

[5] Holloway JJ, Thomas JW. Factors influencing readmission risk: implications for quality monitoring. Health Care Financ Rev 1989; 11:19.

[6] Sisk JE, Dougherty DM, Ehrenhaft PM, Ruby G, Mitchner BA. Assessing information for consumers on the quality of medical care. Inquiry 1990; 27:263.

[7] Pallarito K. Expert sees quality of care becoming important factor in hospital credit ratings. Modern Healthcare 1990; 20:48

[8] Perry L. Michigan Blues plan initiates bonuses penalties ties tied to standards of quality. Modern Healthcare 1989;19:58.

[9] Sarwar, F., & Zaman, M. A. U. (2017). A comparative study on the application of evolutionary algorithms to multi-objective, multi-stage supply chain network design. *International Journal of Supply Chain and Inventory Management*, *2*(2), 143-161.
doi: https://www.doi.org/10.1504/IJSCIM.2017.092325

[10] Zaman, M. A. U. (2018). *Bicubic L^1 Spline Fits for 3D Data Approximation* (Master's thesis, Northern Illinois University).

[11] Moraga, R. J., & Rabiei Hosseinabad, E. (2017). A System Dynamics Approach in Air Pollution Mitigation of Metropolitan Areas with Sustainable Development Perspective: A Case Study of Mexico City. *Journal of Applied Environmental and Biological Sciences*, 7(12), 164-174.

[12] Hosseinabad, E. R., & Moraga, R. J. (2017). Air Pollution Mitigation in Metropolitans Using System Dynamics Approach. In IIE Annual Conference. Proceedings (pp. 638-643). Institute of Industrial and Systems Engineers (IISE).

[13] H. R. Feili, P. Ahmadian, and E. Rabiei, "Life Cycle Assessment Of Municipal Solid Waste Systems To Prioritize And Compare Their Methods With Multi-Criteria Decision Making," *The Open Access Journal of Resistive Economics (OAJRE)*, vol. 2(1), pp. 38-46, 2014.

[14] Feldstein, M. S. (1971). Hospital cost inflation: A study of nonprofit price dynamics. The American Economic Review, 61(5), 853-872.