2016

Prediction modeling and pattern recognition for patient readmission

Davood Golmohammadi, PhD

Int. J. Production Economics 171 (2016) 151–161

Contents lists available at ScienceDirect

Int. J. Production Economics

journal homepage: www.elsevier.com/locate/ijpe

Prediction modeling and pattern recognition for patient readmission

Davood Golmohammadi *, Naeimeh Radnia

University of Massachusetts Boston, United States

ARTICLE INFO

Article history:
Received 18 May 2014
Accepted 21 September 2015
Available online 13 October 2015

Keywords:
Prediction modeling
Pattern recognition
Hospital readmission

ABSTRACT

Readmission is a major source of cost for healthcare systems. Hospital-specific readmission rates are considered an indicator of hospital performance and generate public interest regarding the health care quality. We aimed to identify those patients who are likely to be readmitted to the hospital. The identified patients can then be considered by health care personnel for application of preventive alternative measures such as: providing intensive post-discharge care, managing the conditions of the most vulnerable in their home, supporting self-care, and integrating health services and information technology systems to avoid unnecessary readmissions. Neural Network, Classification and Regression model and Chi-squared Automatic Interaction Detection models were used for the readmission prediction. All models were able to perform with an overall accuracy above 80%, with the latter two models having the advantage of providing the user with the opportunity of selecting different misclassification costs. We employed C5.0 algorithm to search for recurring pattern in the history or demographics of patients who have been readmitted and explored if a rule of thumb can be derived to predict those at risk of future readmissions. Moreover, the key variables influencing readmission were studied based on a large data set. The most important factors contributing to readmission were determined such as age, sex, number of previous prescriptions and length of previous stays, place of service, and number of previous claims.

1. Introduction

Re-hospitalization is a major issue of concern for the U.S. healthcare system. According to the American Hospital Association's (2013) annual survey of U.S. hospitals, the total number of admissions exceeded 36 million in 2011 and as the numbers from the Agency for Healthcare Research and Quality (AHRQ) show, about one in 10 hospitalizations in 2008 was potentially unnecessary (AHRQ, 2010). This problem is costly. For example, in 2008, $12 billion of Medicare spending went towards potentially preventable readmissions according to the Medicare Payment Advisory Commission (MedPAC)'s report to Congress (Glenn and Hackbarth, 2009). Carey and Stefos (2015) found that,overall, hospitals could expect to save $2140 for the average 30-day readmission avoided. For heart attack, heart failure, and pneumonia patients, expected readmission cost estimates were $3432, $2488, and $2278, respectively. For high-risk patients, including those with severe illnesses and complications, those expected costs more than doubled.

Hospital-specific readmission rates are considered an indicator of hospital performance and generate public interest regarding the health care quality. Some policy makers are considering either rewarding hospitals with low readmission rates or penalizing those with high rates, and studies have been done to introduce a consensus method of calculating the readmission rate in order to make it a more reliable quality-of-care indicator for comparison and ranking of hospitals (van Walraven et al., 2012).

Research has been done to determine which factors contribute to early re-hospitalization, mostly limiting the population under study to the elderly, patients with a specific disease or a certain ethnicity (Marcantonio et al., 1999; Philbin and DiSalvo, 1999; Ottenbacher et al., 2000; Shyu et al., 2002; Reuben et al., 2002; Hamner and Ellison, 2005; Wong et al., 2010; Allaudeen et al. 2011; Dunlay and Gersh, 2013; Engoren et al., 2013; Njagi et al., 2013). Much attention has been given to predicting hospital readmission within a thirty-day time frame after discharge as a binary value (Marcantonio et al., 1999; Lagoe et al., 2001; Shyu et al., 2002; Wong et al., 2010; van Walraven et al., 2012).

Applying a comprehensive dataset that make generalization more reasonable, we aimed to identify those patients who are likely to be readmitted to the hospital. The identified patients can then be considered by health care personnel for application of preventive alternative measures such as: providing intensive post-discharge care, managing the conditions of the most vulnerable in their home,

* Corresponding author.
E-mail addresses: davood.golmohammadi@umb.edu (D. Golmohammadi), naeimeh.radnia001@umb.edu (N. Radnia).

http://dx.doi.org/10.1016/j.ijpe.2015.09.027
0925-5273/

supporting self-care, and integrating health services and information technology systems to avoid unnecessary readmissions. In this research, we investigated the following research questions:

- What are the key variables influencing readmission?
- What is the recurring pattern in the history or demographics of patients who have been readmitted?
- What rule of thumb can be derived from these patterns to predict those at risk of future unnecessary readmissions?

We employed secondary data from a local hospital that recorded medical data over three years (in contrast with the extant literature, which is usually based on a short period of time) for an exceptionally comprehensive population of more than 113,000 patients, inclusive of a range of ages and health conditions. We developed predication models to identify patients who are more likely to be readmitted and compared the results.

The remainder of this paper is organized as follows. A detailed background of the research already carried out on risk-predicting models for hospital readmissions is provided in Section 2. Section 3 explains the methodology and Section 4 describes the data and explains the process of data preparation. Section 5 elucidates the modeling techniques. A discussion of the results is presented in

Table 1
Risk factors considered as the most important contributors to readmission in previous studies.

	Literature									
Factors	Boult et al. (1993)	Marcantonio et al. (1999)	Lyon et al. (2007)	Reuben et al. (2002)	Billings et al. (2006)	Bottle et al. (2006)	Donnan et al. (2008)	van Walraven et al. (2012)	Donzé et al. (2013)	Total number of factors
Age	X	X			X	X	X			5
Ethnicity					X	X				2
Sex	X			X	X	X	X			5
Availability of an informal caregiver	X									1
Local admission rate						X				1
Not working				X						1
Area-level socio-economic data						X				1
High social deprivation							X			1
Sodium level at discharge									X	1
Hemoglobin at discharge									X	1
Discharge from an oncology service									X	1
Lack of documented patient or family education		X								1
Depression		X								1
Religious participation				X						1
Low iron level				X						1
Low serum albumin				X						1
Taking loop diuretics				X						1
Number of respiratory medications							X			1
Previously prescribed analgesics, antibacterials, nitrates & diuretics							X			1
Procedure									X	1
Admission for an ambulatory Care sensitive condition						X				1
Acuity of admission								X		1
Source of admission						X				1
Type of admission									X	1
Clinical condition					X					1
Leg ulcers			X							1
Heart problem			X							1
Diabetes	X			X						2
More than six doctor visits	X									1
ER admission during the last year/last 3 years/ last 6 months			X			X		X		3
Hospital admission during the last year/last 2 years/ 30 days				X	X		X		X	4
Having ever had coronary artery disease	X									1
Length of stay							X	X	X	3
Comorbidity		X				X		X		3
Unable to walk 0.5 mile				X						1
Needing help bathing				X						1
Ability to go out of the house without help			X							1
Self-rated general health	X		X	X						3
Memory loss			X							1

Section 4, and Section 7 gives a summary of our research contributions and findings.

2. Background

A growing amount of literature is devoted to developing new tools for predicting readmission risks (Kansagara et al., 2011), and these studies vary in target population, objectives and outcome variables, as well as in the statistical approaches used for prediction. Most previous studies have focused exclusively on elderly populations (Marcantonio et al., 1999; Shyu et al., 2002; Reuben et al., 2002; Allaudeen et al., 2011; van Walraven et al., 2012). Although the underlying hypothesis of these studies is that most readmissions affect individuals who are over 65 years old, substantial resources are devoted to younger people being readmitted as well. Our findings support that elderly people is not the only group who should be under scrutiny in understanding readmissions risk.

Some studies focus on predicting readmission for patients with a particular disease (Philbin and DiSalvo, 1999; Ottenbacher et al., 2000; Hamner and Ellison, 2005; Dunlay and Gersh, 2013; Engoren et al., 2013; Njagi et al., 2013), while others constrain the studied population to a specific ethnicity or race (Shyu et al., 2002; Wong et al., 2010). We could not find a comprehensive study with a large data set in the literature to investigate related research questions.

Another interesting point we found in the literature was that definitions and terms interpretations diverge. Marcantonio et al. (1999), Lagoe et al. (2001), Shyu et al. (2002), Wong et al. (2010) and van Walraven et al. (2012) define re-hospitalization as a return to hospital within a time frame of one month or less. Their assumption is that choosing a thirty-day time frame increases the likelihood that poor outcomes are related to the index admission or the discharge process and are more likely remediable. Lagoe et al. (2001), on the other hand, handles this issue by defining readmission as a non-elective return for inpatient care that occurs for a Diagnosis-related group (DRG) within the same major diagnostic category as the DRG of the initial admission.

The most common methods used to predict re-hospitalization are stepwise logistic regression (Lagoe et al., 2001; Billings et al., 2006) and multivariate logistic regression (Donnan et al., 2008; van Walraven et al., 2012). And the most common technique to evaluate the model performance is to use C-statistic, the area under the Receiver Operating Characteristic curve (Donnan et al., 2008; van Walraven et al., 2012; Donzé et al., 2013). Table 1 shows the different factors previous studies have identified as important contributors to readmission.

Reviewing previous studies clarifies that most previous studies in predicting hospital readmissions are limited to applying a single predictive model, while our research is fairly unique in employing four different techniques as well as using an exceptionally large population under study with a long previous history of three years. In terms of population size, for instance 264, 233, 164 and 80 patients were studied by Shyu et al. (2002), Allaudeen et al. (2011) and Njagi et al. (2013), respectively, which makes their results hard to generalize. Moreover, not limiting our data to a specific age, ethnicity, or disease group makes our analysis more readily generalizable. Additionally sensitivity analysis on important factors contributing to readmission is not covered by previous works. Finally, using the longer readmission period of one year, as opposed to the thirty-day time frame often used, makes our models more practical and useful for health care workers who can thus create long-term plans in advance.

3. Methodology

To conduct the analysis in this study, secondary data from a local hospital consisting of 113,000 individuals' medical histories over three years were used. The data from five datasets of Patients data, Claims data, Laboratory data, Drug data, and Outcome data were modified and combined to serve our research.

We aimed to predict the patients' readmission status in the third year as well as the key driver factors involved in the readmission process. The pattern recognition in history of readmitted patients and possible rules of thumb for spotting most vulnerable patients to readmission in future were also studied.

To achieve a prediction model that determines which patients are at greatest risk of being readmitted, we used three models Neural Network, Classification and Regression model (C&R), and Chi-squared Automatic Interaction Detection (CHAID). In each case:

- The overall accuracy was considered as a measure of models performances; and
- The models were compared to find out which produces the highest accuracy.

To address the pattern recognition question, a C5.0 algorithm was employed to search for some rules of thumb that can describe readmission patterns in the data. Finally, a sensitivity analysis was carried out to find out how the output of the Neural Network model changed with respect to variation in input variables. See Fig. 1 for a display of the steps taken.

4. Data description and preparation

In the following, more information about the data and datasets are presented:

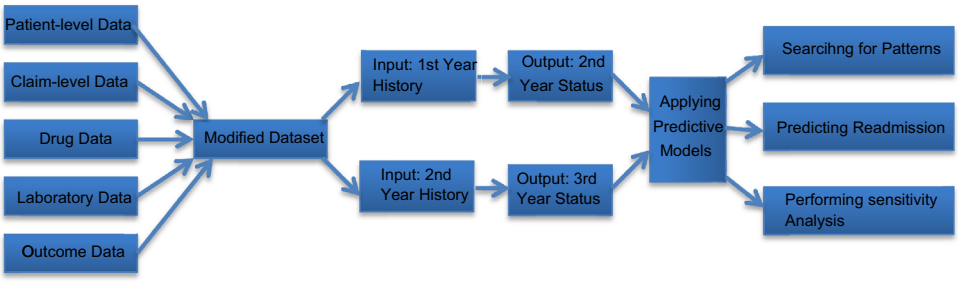

Fig. 1. The steps of our analysis.

Table 2
List of the field in our 5 datasets.

Dataset	Fields	Notes
Patient data	Member ID	
	Age at first claim	10 categories
	Sex	
Claims data	Member ID	
	Provider ID	14,700 categories
	Vendor ID	6388 categories
	Primary care physician (PCP)	1360 categories
	Year in which the claim was made	3 categories
	Generalized Specialty groups.	13 categories (refer to Table A1 for details)
	Place of service	8 categories (refer to Table A1 for details)
	Pay delay	
	Length of stay	7 categories
	Days since first claim (DSFS)	12 categories
	Primary condition	46 categories (refer to Table A1 for details)
	Charlson Index	A measure of the affect, diseases have on overall illness with 4 categories (Charlson et al. 2008).
	Procedure Groups	17 categories (refer to Table A1 for details)
Drug Count data	Member ID	
	Year	The year in which the drug prescription was filled
	Days since first claim (DSFS)	12 categories
	Drug Count	Count of unique prescription drugs filled by DSFS.
Lab Count data	Member ID	
	Year	The year in which the Laboratory or pathology test was ran.
	Days since first claim (DSFS)	12 categories
	Lab Count	Count of unique laboratory and pathology tests.
Outcome data	Member ID	
	Days in hospital year2	Days in hospital, the main outcome, for members with claims in Y1.
	Days in hospital year 3	Days in hospital, the main outcome, for members with claims in Y2.
	Claimed truncated	2 categories

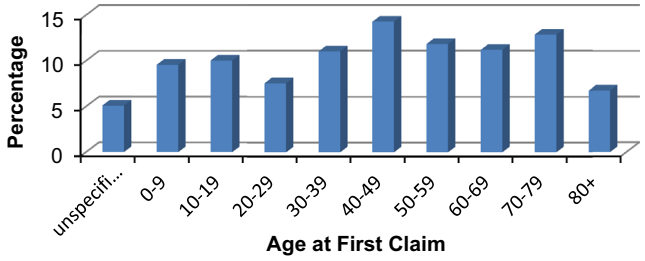

Fig. 2. Members' distribution of age.

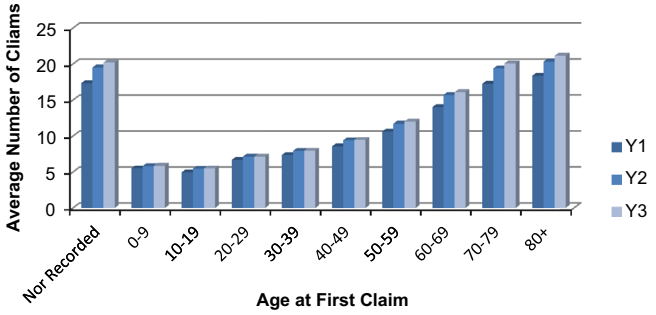

Fig. 3. Average number of claims for each age group.

4.1. Data description

Our analysis consisted of the following five datasets for 113,000 patients over three years:

- Patients data
- Claims data
- Laboratory data
- Drug data
- Outcome data

The list of fields for each dataset is shown in Table 2. As listed in the third column, many of the fields had a large number of categories, which added to the complexity of our modeling.

As an instance we explain here, the details for one field; age at first claim. Fig. 2 shows the histogram of age for all the participants, later when we combined this data set with claims data, we observed that as the age goes up, the average number of claims increases, and also the comparison of three years data showed that the average number of claims had an increasing trend over time (See Fig. 3) The age groups of over 80, 70–70 and 20–29 were the ones with highest percentage of stay in hospital in the third year with 27.08%, 19.74% and 15.37% respectively. The order was the same for the second year with the numbers being 29.2%, 20.14% and 16.4% respectively (See Fig. 4).

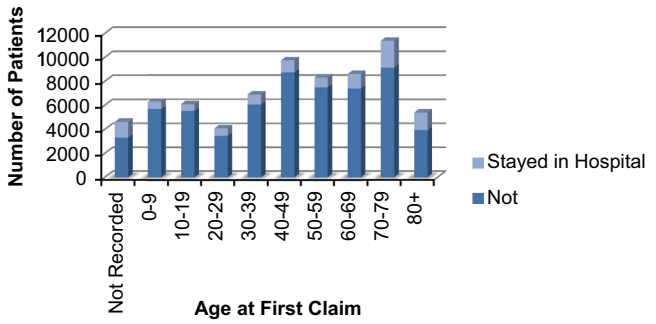

Fig. 4. Distribution of age and hospital stay in year 3.

Our dataset size was large, more than 100,000 patients who had an average of 18 claims per year, with to 44 claims per year. This does not mean a patient was admitted to hospital 44 times, because each procedure was recorded as a claim; for instance, in one reference to the hospital, if a patient was visited by a physician, then went to radiology, and then went to pathology, this would be recorded as three different claims. Fig. 5 shows the frequency of the number of claims for the first year.

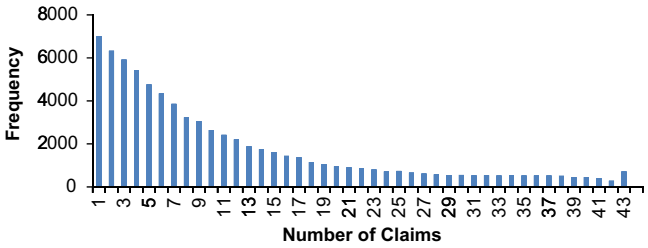

Fig. 5. Frequency of claims in first year.

Not all software can handle such a large dataset. Our search for the most effective software led to the IBM SPSS Modeler 15.0, which can handle applying different models to a large dataset and making comparisons among them.

4.2. Data preparation

We had four fields that were pseudonyms: patient ID, provider ID, physicians ID and vendor ID. As pseudonyms, they should have been treated as nominal values, not numbers, because a large or small number for a physician does not imply anything specific. Considering that different clinicians, vendors and providers may have their own particular traits that affect hospital admissions, it would have been useful to apply these fields in our models (Qu and Shi, 2011, van der Vaart et al., 2011) but due to the large number of categories, converting these fields to binary variables was not viable, since 1360 physicians would require 1359 new binary variables. We therefore used the number of unique physicians for a specific patient in a specific year as a new field, doing the same for provider and vendor. We used SQL codes for calculating. The member ID field was not used in model building as a patient is defined by his or her history rather than the assigned pseudonym.

Each nominal field with n categories, such as Place of service, diagnostic categories, procedure categories, etc., was replaced by $n-1$ binary variables. Assume the variable Z indicates the Place of service and can get four values: Ambulance, Home, Office and Hospital. We replaced this nominal variable by four binary ones with the following conditions:

Replace this nominal variable → with 3 binary variables
Replace Z with $\rightarrow X_{Amb}, X_{Hom}, X_{Ofc}$
$X_{Amb} = 1$ if Z=Ambulance, $X_{Amb} = 0$ *otherwise*
$X_{Hom} = 1$ if Z=Home, $X_{Hom} = 0$ *otherwise*
$X_{Ofc} = 1$ if Z=Office, $X_{Ofc} = 0$ *otherwise*
$X_{Amb}, X_{Hom}, X_{Ofc} = 0$ if Z=Hospital

The patient ID, which was a nominal field with 113,000 categories, was excluded from our models, which made it impossible to identify which claims belonged to which patient. The solution was to use patient-level data instead of claim-level data by aggregating all claims into one row of data. The risk of aggregating all the claims of a patient into one row of data was a potential loss of important information; we handled this risk by adding pairwise combination fields such as Specialty*Procedure, Category*Primary and Diagnostic* Category.

The final field in our primary data was Admission Risk, defined for each claim as an integer on a scale of 1 to 5 as a factor of two fields: age and Primary Condition Group.

5. Modeling

After combining the aforementioned datasets and doing the data modifications briefly explained in previous paragraphs, we reached a patient-level dataset with 130 fields. Considering that we assumed a patient's medical history in the current year could be employed to predict the individual's status in the next year, we ran each of the applied models on two sets of input and output data. In the first set, the input variables were the medical history in the first year and the output variable was readmission status in the second year. In the second set, the input variables were the medical history in the second year and the output variable was the readmission status in the third year (See Fig. 2). Employing the two datasets for each model and comparing the results helped us be assured of the models' performance. We took the following steps:

- First we used Neural Network, Classification and Regression (C&R), and Chi-squared Automatic Interaction Detection (CHAID) models to spot the patients most likely to be readmitted to hospital within the next year. The three models were employed in order to reach a model with a high level of accuracy. The two later models had the advantage of providing the user with the choice of misclassification costs. So that the user could decide how to charge the model if it misclassifies a readmitted patient as non-readmitted one or vice versa. (More explanation is provided in in the Subsection 5.1.2).
- Sensitivity Analysis was then performed to identify how the output of the Neural Network model changes with respect to variation of the input variables.
- Finally, a C5.0 algorithm was applied to search for recurring characteristics patterns among individuals with readmissions.

5.1. Predicting readmission

Three aforementioned models of Neural Network, Classification and Regression (C&R), and Chi-squared Automatic Interaction Detection (CHAID) models were developed to predict patients' readmission status in the following year based on their history.

5.1.1. Neural Network

Neural Network models simulate the way the human brain processes information by creating interconnected processing units that resemble the human body's neurons. These models collect and process data for the purpose of learning. Generally speaking, the processing units are arranged in three types of layers: input layers; one or more hidden layers; and output layers. These units are connected with varying connecting weights. Initial weights are chosen randomly and the model learns through training, by repeatedly studying individual histories, making a prediction for each record, and adjusting the weights when it makes a wrong prediction. The model continually becomes more accurate until a stopping criterion is met (Golmohammadi et al., 2009a, 2009b; Golmohammadi, 2011; IBM, 2013).

The models developed in this study employed the Multi-Layer Perceptron (MLP) form of Neural Network. MLP, unlike many other statistical methods, does not make any assumptions on the distribution of data, the linearity of the output function, or the type of predictor or output variables, so it can find more complex relationships than the Radial Basis Function (RBF), the other form of Neural Network (IBM, 2013).

In the first run of the model, we used 130 fields of the patient-level data from the first year as variables in the input layer, and a binary variable indicating readmission in the second year in the output layer (please see Subsection 4.2 for details on how the data fields were created). 75% of records were used for training the

model, while the remaining 25% were held back for testing. To achieve the lowest error rate, the model was run three times with the following three stopping rules: maximum training time; customized number of maximum training cycles; and minimum accuracy. The rule and the model with the highest accuracy were chosen. The best results were achieved when the model had one hidden layer with eight neurons and the overall accuracy was 84.2%. In the second run of the model, 130 fields of the patient-level data from the second year were used as variables in the input layer, while the output layer consisted of a binary variable indicating readmission in the third year. Again the proportion of the training to testing set was 75–25. The highest overall accuracy was 85% when the model structure had one hidden layer with 8 neurons.

5.1.2. The Classification and Regression (C&R)

The Classification and Regression (C&R) was the second model applied. C&R builds a decision tree to classify future observations by partitioning the training data into splits and trying to maximize the purity of each model. Each node in the decision tree is defined as pure if the entire node's cases fall into a certain class of the dependent variable field.

One necessary assumption for this model was misclassification costs, which were factored into the model as a way of protecting against costly mistakes. In choosing the appropriate costs, we must decide first when do we want to charge the model more; when it misclassifies a readmitted patient as non-readmitted (we call it “type 1 misclassification”), or when it misclassifies a non-readmitted patient as readmitted (we call it “type 2 misclassification”). In the case where it is more important for the health care unit to spot patients with a high risk of readmission and take preventive measures to avoid it, a higher cost should be assigned to type 1 misclassification. However, if the goal of a health care unit is to apply the model to get an estimation of the number of patients who will be readmitted and use that to provide necessary resources, a higher cost should be considered for type 2 misclassification, since overestimating needed resources would force an unnecessary cost on the health care unit, considering the importance of managing shared resources, especially beds. (Bachouch et al., 2012; Saremi et al., 2013).

In the first run of this model, the input variables were 130 fields of patient-level data from the first year, and the output variable considered to a binary variable indicating readmission in the second year. We ran the model with various assumptions where type 1 misclassification costs 1,2,…,7 times more than type 2 misclassification. We stopped at 7 since for larger numbers the chance of the model correctly classifying non-readmitted individuals were becoming too low. A sample result is shown in Table 4, where type 1 misclassification costs six times more than type 2 misclassification (as listed in Table 3). The overall accuracy (defined as the percentage of patients, either readmitted or non-readmitted, which were correctly classified by the model) was 59.44%. The best overall accuracy of 84.2% was achieved when the two misclassification costs were the same.

Table 3
A sample of considered misclassification costs, where misclassification type 1 costs six times more than misclassification type 2.

		Prediction	
		Not readmitted	Readmitted
Observation	Not readmitted	0	1
	Readmitted	6	0

Table 4
Performance results associated with the misclassifications cost in Table 3.

		Prediction	
		Not readmitted (%)	Readmitted (%)
Observation	Not readmitted	57.39	42.61
	Readmitted	29.37	70.63

Table 5
Performance results achieved when misclassification type 1 costs five times more than misclassification type 2.

		Prediction	
		Not readmitted (%)	Readmitted (%)
Observation	Not readmitted	68.79	31.21
	Readmitted	38.14	61.86

Table 6
A sample of considered misclassification costs, where misclassification type 1 costs five times more than misclassification type 2.

		Prediction	
		Not readmitted	Readmitted
Observation	Not readmitted	0	1
	Readmitted	5	0

In the next run of the model, the input variables were 130 fields of patient-level data from the second year, and the output variable considered to be a binary variable indicating readmission in the third year; seven different sets of misclassification costs were considered where type 1 misclassification costs 1,2,…,7 times more than type 2 misclassification. A sample result is shown in Table 5 where type 1 misclassification costs five times more than type 2 misclassification, with an overall accuracy (defined as the percentage of patients, either readmitted or non-readmitted, which were correctly classified by the model) of 67.75%. The best overall accuracy of 85% was achieved when both misclassification costs were the same.

Compared to the Neural Network, the C&R has the advantage of giving one a choice in selecting case-appropriate misclassification costs. Both NN and C&R resulted in fairly high levels of accuracy.

5.1.3. Chi-squared Automatic Interaction Detection (CHAID)

The third model applied to our data was Chi-squared Automatic Interaction Detection (CHAID). It creates a decision tree by identifying optimal splits through the application of Chi-square statistics. First the model studies the cross-tabulation between each pair of input variables and the target variable and uses Chi-square independent test to test the significance. The most significant relation with the lowest p Value then is selected. For inputs with more than two categories, the categories are compared and those that show no difference in outcome are merged together; this is continued repeatedly by merging the categories with the least significant difference until at a certain testing level all the remaining categories differ (IBM, 2013). The CHAID model, like the C&R, has the advantage of providing users with a choice of different misclassification cost options.

In the first run of the model, we assumed the input variables were 130 fields of the patient-level data from the first year and the output variable was readmission status in the second year. Various sets of misclassification cost were considered where type 1 misclassification costs 1,2,…,7 times more than type 2. We stopped at 7 since after that the chance of the model classifying non-readmitted patients correctly were very low. A sample result is shown in Table 7, where type 1 misclassification costs five times more than type 2; as listed in Table 6, the overall Accuracy was 67%. The highest overall accuracy of 84.2% was achieved when the misclassification costs were equal.

In the second run of the model, we considered the input variables to be the 130 fields of the patient-level data from the second year, and the output variable was readmission status in the third year. The model was run with various assumptions where type 1 misclassification costs 1,2,…,7 times more than type 2. A sample result is shown in Table 8, where type 1 misclassification costs six times more than type 2; the overall accuracy was 64.72%. The best overall accuracy of 85% was obtained when both misclassifications cost the same.

A brief summary of models comparsion is presented in Table 9.

5.2. Pattern recognition

Finally, to search for a recurring characteristics pattern among individuals with unnecessary readmissions, a C5.0 algorithm was used. It derives a set of rules to predict readmission as a binary variable. Without the need for further modeling, when a patient falls into a category of these derived rule sets, the hospital staff can determine his/her chance of readmission and take the necessary actions. The algorithm splits the dataset based on the field that delivers the biggest information gain. Each division is split again and the process is continued until the subsets can no longer be split (IBM, 2013).

In the first run of the model, when the input variables were the 130 fields of the patient-level data from the first year and the target field was readmission status in the second year, 32 rules were created to help one predict if the patient will be readmitted to the hospital in the next year. Table 10 includes the 3 most accurate ones. The third rule, for instance, implies that if an individual has equal or less than 6 referrals to hospital with Place of service recorded as Office and never had a “Miscellaneous non-cardiac congenital anomalies; miscellaneous symptoms other than fever; miscellaneous tooth and tongue disorders, miscellaneous diagnoses of pain” condition, or Surgery-Integumentary System procedure, had more than 6 claims with the Pregnancy condition, and less than 10 unique laboratory or pathology tests, then there is a 93.8% chance that she would be readmitted to hospital.

In the second run of the model, when the input variables were the 130 fields of the patient-level data from the second year and the target field was readmission status in the third year, 38 rules were created to help one predict if the patient will be readmitted in the third year. We included the 3 most accurate rules in Table 11.

In general these rules can be used by heath care units to spot the most vulnerable patients for readmission as they do not require the health care unit to employ statistical software and modeling techniques or to hire skilled staff. However, not much consistency could be found in the rules for the first and second year, suggesting that our dataset is too complex to recognize a much accurate pattern in the readmission data. We may have reached a better, more consistent pattern recognition if we had used data from a wider time frame of, for instance, 10 years.

Table 7
Performance results associated with the misclassifications cost in Table 6.

		Prediction	
		Not readmitted (%)	Readmitted (%)
Observation	Not readmitted	67.78	32.22
	Readmitted	37.28	62.72

Table 8
Performance results when misclassification type 1 costs six times more than misclassification type 2.

		Prediction	
		Non-readmitted (%)	Readmitted (%)
Observation	Non-readmitted	64.61	35.39
	Readmitted	34.66	65.34

6. Discussion

Using patients' demographics and medical history, we were able to discover patterns in readmitted patients' data and identify the significant factors in patients' readmission to hospital. In this section we discuss; selecting the proper misclassification costs, the practicability and reliability of the derived rules, the availability and generalizability of the proposed models, and the results of the performed sensitivity analysis .

For health care management, in choosing the right misclassification cost as an input to the proposed models, it is important to clarify the purpose of modeling readmission status; Is it to spot those individuals at high risk of readmission and consequently take preventive actions (hence desire a higher *B* value in Table)? Or is it to get an estimation of the number of individuals who will be readmitted, so that one can plan the required resources in advance (therefore desire a higher *A* value in Table)?

Table 9
A brief comparison of models.

	Models		
	Neural Network	Classification and Regression	Chi-squared Automatic Interaction Detection
Objective	To predict patients' readmission status in the following year based on their history		
The highest level of accuracy	85%	84.20%	85%
Overall accuracy	84%	59.4–67.7%	64.72–67%
Advantages	Powerful method to determine a complex relationship between inputs and outputs	Providing the user with the choice of misclassification costs	
Challenge	Design parameters	Misclassification costs	

Table 10
Three most accurate rule sets discovered to predict readmission in second year.

Number	Rule	Accuracy (%)
1	No claims with Specialty being Emergency And at most one claim with Specialty being Pathology And no claims with Primary Condition Group being Congestive Heart Failure And at least one claim with Primary Condition Group being Chest Pain And no claims with Primary Condition Group being Urinary Tract Infections And sum of claims with the Charlson Index of 3–4 > 2 And number of claims > 21 And Claims Truncated > 0	96.4
2	Sum of claims with Days Since First Service of 10–11 months > 4 And at least two claims with Primary Condition Group being Pregnancy And no claims with Procedure Group being Surgery-Integumentary System And Sum of Lab Count > 9	95
3	Number of claims with Place of service being Office < = 6 And no claims with Primary Condition Group being "Miscellaneous non-cardiac congenital anomalies; miscellaneous symptoms other than fever; miscellaneous tooth and tongue disorders, miscellaneous diagnoses of pain" And number of claims with Primary Condition Group as Pregnancy > 6 And no claims with Procedure Group being Surgery-Integumentary System And Sum of Lab Count < = 9	93.8

Table 11
Three most accurate rule sets discovered to predict readmission in third year.

Number	Rule	Accuracy
1	Number of claims with Specialty being Emergency > 2 And no claims with Specialty being Pathology And number of claims with Primary Condition Group being Gastrointestinal Bleeding > 3 And number of claims with Primary Condition Group being Ingestions and Benign Tumors < = 4 And Claims Truncated > 0	93.80%
2	At least one claim with Primary Condition Group being Acute Renal Failure And Sum of Lab Count > 56 And Claims Truncated > 0	92.30%
3	No claims with Specialty being Emergency And no Length Of Stay shorter than 1 day And at least one claim with Primary Condition Group being Congestive Heart Failure And at least one claim with Primary Condition Group being Pneumonia And Sum of Drug Count < = 45 And Claims Truncated > 0	91.70%

Table 12
Choosing a misclassification cost that serve our purpose.

		Prediction	
		Non-readmitted (%)	Readmitted (%)
Observation	Non-readmitted	A	(100 − A)
	Readmitted	(100 − B)	B

Choosing a misclassification cost that serve our purpose is shown in Table 12.

In the former case, we suggest selecting a high cost of type 1 misclassification (that is, charging the model more if it misclassifies a readmitted patient as non-readmitted), and in the latter case a higher cost of type 2 misclassification is recommended (that is, charging the model more if it misclassifies a non-readmitted patient as readmitted).

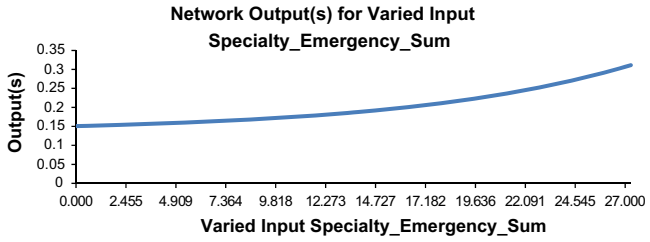

Fig. 6. Network Output for varied Input being the number of claims with "Emergency" listed as Specialty.

Apropos the pattern recognition, in terms of practicability, while most of the derived models here may require skilled staff for application on new patient data, the decisions rules suggested are quiet straightforward and can be used by management in the health care system to make decisions relating to readmission. The reliability of the achieved rules here is not high, as the rules derived from the first and second year data are not consistent. This may indicate that our data is too complex to allow for an accurate pattern recognition. Had we had access to data from a longer period of time, we may have been successful in finding a more consistent recurrent pattern.

Regarding the availability of data, the increasing trend of digitizing health care data and routinely adding new data in electronic form is likely to make the introduced models increasingly operable.

6.1. Sensitivity analysis

We performed a sensitivity analysis to find out how the output of the Neural Network model changed with respect to variation in input variables. We used Nerousoltions 6 for performing that. The network learning was disabled during this operation so that the network weights were not affected. The basic idea used by the software is that the inputs to the network are shifted slightly and the corresponding change in the output is reported either as a percentage or a raw difference. The way Neurosolutions work is to generate the input data for the sensitivity analysis by temporarily increasing the input by a small value (dither). The corresponding change in output is the sensitivity data (NeuroDimension Inc.,

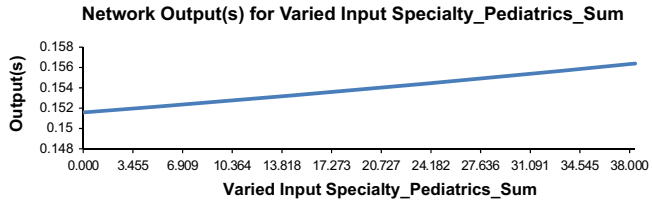

Fig. 7. Network Output for varied Input being the number of claims with "Pediatrics" listed as Specialty.

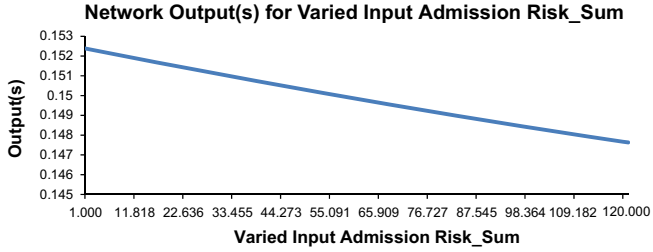

Fig. 8. Network Output for varied input being the sum of Admission Risk.

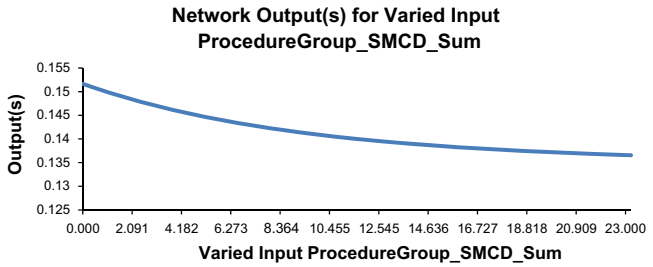

Fig. 9. Network Output for varied input being the number of Claims with SMCD listed as Procedure Group.

2014).

We performed the sensitivity analysis for all variables, due to the large number of them it is not viable to bring all of the resulted curves here, hence we categorized them based on the increasing/decreasing trend and their pattern to four groups.

The first group consisted of the following variables; Specialty group being Diagnostic Imaging or Emergency; Primary Condition Group being Atherosclerosis and Peripheral Vascular Disease; Charlson Index being 3–4 or 5+; Procedure Group being Emergency, Pathology and Laboratory, or Surgery-Integumentary System; number of claims; and number of Unique Vendors. These variables had almost the same pattern in that by increasing the input variable's value, the model results in a larger output value, and the model can lead to an acceptable result for any data outside the trained data range. The increase is small at first but as the input variable gets larger, the output increases with a higher rate. Fig. 6 demonstrates the sensitivity curve for one instance of such inputs. We see that by increasing the number of times the patient has been admitted with Specialty group being Emergency, the model results in a larger output value, and the slope is small at first but gets sharper as the inputs gets higher.

The second group included the following variables: Days Since First Service being 2–3 months, 7–8 months or 9–10 months; Primary Condition Group being Arthropathies, Ovarian and Metastatic Cancer, Gynecology, Diabetic Ketoacidosis and Related Matabolic, Perinatal Period, Pneumonia, or Acute Renal Failure, Charlson Index being 0, Procedure Group being Surgery-Cardiovascular System, or Surgery-Eye and Ocular Adnexa and the number of unique primary care physicians. These variables all had almost the same pattern in that by increasing the input variable value, the model results in a larger output value. The increase is at the same rate at all times and the model can lead to an acceptable result for any data outside the trained data range. Fig. 7 demonstrates the sensitivity curve for one instance of such variables. We see that by increasing the number of times the patient has been admitted with Specialty group being Pediatrics, the model results in a larger output value, and the slope is almost constant.

The third category consisted of the following variables; Admission Risk; Specialty being Anesthesiology, Internal, Laboratory, Obstetrics and Gynecology or Surgery; Place of service being Independent Lab or Outpatient Hospital, Length Of Stay being 1–2 weeks, 2 days, or 4 days; Days Since First Service being 0–1 month, 1–2 months, 10–11 months, 3–4 months, 4–5 months, 5–6 months or 6–7 months; Primary Condition Group being Fluid And Electrolyte, Liver Disorders, Sepsis or Ingestions And Benign Tumors; Charlson Index being 1–2; Procedure Group being Surgery-Auditory System or Surgery-Urinary System; Lab Count; and number of Unique Provider. These variables all had the same pattern in that by increasing the input variable, the model results in a smaller output value. The decrease has the same rate at all time and the model can lead to an acceptable result for any data outside the trained data range. Fig. 8 demonstrates the sensitivity curve for one instance of such variables. We observe that by increasing the sum of Admission Risk, the model results in a smaller output value, and the slope is almost constant.

Finally the last category included the following variables; Specialty being Pathology or Rehabilitation; Place of service being Ambulance, or Office; Primary Condition Group being Gastrointestinal Bleeding or Non-Malignant Hematologic; and Procedure Group being Surgery-Maternity Care And Delivery. These variables had almost the same pattern in that by increasing the input variable value, the model results in a smaller output value. The decrease is small at first but as the input variable gets larger, the output decreases at a higher rate. Fig. 9 demonstrates the sensitivity curve for one instance of such inputs. We see that by increasing the number of times the patient has been admitted with Procedure Group being Surgery-Maternity Care And Delivery, the model results in a smaller output value, and the negative slope is small at first but gets sharper as the inputs gets higher.

7. Summary and concluding remarks

The purpose of this study was to find a practical solution to reduce the readmission rate of health care units. We were able to develop some statistical models capable of predicting with high accuracy which individuals were most likely to have an unnecessary readmission in the following year based on their past medical records. Neural Network, C&R, and CHAID models were used for this purpose, and all were able to perform with an overall accuracy above 80%, with the latter two models having the advantage of providing the user with the opportunity of selecting different misclassification costs. This advantage becomes useful in cases when the user decides it is highly costly for them if the model misclassifies an actually readmitted person as not readmitted and hence desires a high value of B in Table, or in the opposite case when the user decides it is highly costly for them if the model misclassifies an actually non-readmitted person as readmitted and hence desires a high value of A in Table. In the next step we were able to perform a sensitivity analysis to investigate how the output of the Neural Network model changed, with respect to variations of the input variables, and to ascertain that no input variable

causes significant uncertainty in the output. Based on the results of the sensitivity analysis, we accomplished to recognize four different patterns in Network Output among all our variables. In the last step we succeeded in finding a recurring pattern in the history and demographics of readmitted patients. We derived a rule of thumb from those patterns to predict future unnecessary readmissions using a C5.0 algorithm, but the derived rules from the first and second year data did not match one another, implying that after all our dataset was too complicated to be used for a highly accurate pattern recognition. We also managed to introduce the most important contributing factors to readmission. A summary of this study's contributions is as follow:

- Performing the analysis on a large set of patients, increasing the generalization of the findings as opposed to most previous work.
- Not limiting the population under study to a particular age or ethnicity group or to individuals with a specific disease, unlike other studies. While most previous studies have limited their focus to the elderly, the fact that 64.16% of claims belonged to younger-than-60 age groups in our study reveals that including them in the models is important. Also our study's rareness in looking to readmission in a large time frame of one year after discharge makes it a more useful tool for health care managers to plan ahead.
- Defining readmission as a return to hospital in a wide window of time of one year as opposed to the majority of previous work, which limits the time frame to one month, thus making our modeling a more useful tool for health care management to envisage.
- Apropos the most important factors contributing to readmission, while our results did conform with previous study in some factors like age, sex, number of previous prescriptions and length of previous stays, we determined some other factors that the previous literature failed to recognize as important contributors, such as Place of service, Sum of previous laboratory tests, number of previous claims, and Number of unique providers and vendors. Also regarding Primary Condition Groups and procedure categories, our study offered a more detailed and thorough Classification than previous research. Therefore it did narrow down more accurately those primary conditions and Procedure Groups that lead to unnecessary readmission.
- While most of the previous literature limited modeling to one method, we tried different methods to achieve the highest level of accuracy, while two of the employed techniques also gave user the option to select the proper misclassification costs that serve their modeling purpose.

Our detailed analysis and findings enable us to offer some concluding recommendations for hospital leadership teams that are attempting to decrease their unit's readmission rate. The first step requires gathering patients' medical history in one organized data file, using the variables introduced by the aforementioned models as the most significant ones on readmission status, and identifying the most vulnerable patients. Afterwards the highlighted patients may get special treatments and care to reduce their chances of readmission. Although most hospitals are facing cost and productivity pressures and taking these measures will require additional resources, this investment will produce a quick payback in eliminating unnecessary claims and achieving a better level on a metric that will be an important indicator of quality of care in ranking health care units in near future.

The limitation of this study is that it considers patients' medical data and ignores other factors such as social and economic status and mental and spiritual factors. Counting such variables into the calculations may improve the accuracy of the models. Taking into account hospital-specific admission rates may also enhance the model's accuracy. Moreover the proposed models are based solely on historic data from individuals who have already experienced an admission. Hence if an individual with no previous records refers to a hospital, the proposed models are not capable of providing an estimate of her admission status based on real-time data.

For future research one promising avenue could be taking into account non-medical factors such as socio-economic variables, mental status, etc., to discover if these enhance the accuracy of the model. We propose future research to study the effect of intervention (such as additional home visits and/or regular phone calls by nurses succeeding the discharge to make sure the patient is following the physician's orders and that his/her home environment is safe) on those individuals that our model highlights as most vulnerable to be readmitted. We would then inspect whether taking these preventing steps on patients identified by our model would reduce the hospital-level readmission rate.

Appendix

Table A1

Table A1
List of categories for four categorized claim-level dependent variables.

Variable	List of categories
Place of service	Ambulance; Home; Inpatient Hospital; Independent Lab; Office; Outpatient Hospital; Urgent Care; Other
Procedure Groups	Anesthesia; Evaluation and Management; Medicine; Pathology and Laboratory; Radiology; Surgery-Auditory System; Surgery-Cardiovascular System; Surgery-Digestive System; Surgery-Eye and Ocular Adnexa; Surgery-Genital System; Surgery-Integumentary System; Surgery-Maternity Care and Delivery; Surgery-Musculoskeletal System; Surgery-Nervous System; Surgery-Other; Surgery-Respiratory System; Surgery-Urinary System
Primary Condition Group	Acute Myocardial Infarction; Acute Renal Failure; Acute Respiratory; All Other Infections; All Other Trauma; Appendicitis; Arthropathies; Atherosclerosis and Peripheral Vascular Disease; Cancer A; Cancer B; Catastrophic Conditions; Chest Pain; Chroic Obstructive Pulmonary Disorder; Chronic Renal Failure; Congestive Heart Failure; Diabetic Ketoacidosis and Related Matabolic; Fluid and Electrolyte; Fractures and Dislocations; Gastrointestinal Bleeding; Gastrointestinal, Inflammatory Bowel Disease, and Obstruction; Gynecologic Cancers; Gynecology; Hip Fracture; Ingestions and Benign Tumors; Liver Disorders; Miscellaneous #1; Miscellaneous #2; Miscellaneous #3; Miscellaneous Cardiac; Non-Malignant Hematologic; Other Cardiac Conditions; Other Metabolic; Other Neurological; Other Renal; Ovarian and Metastatic Cancer; Pancreatic Disorders; Pericarditis; Perinatal Period; Pneumonia; Pregnancy; Seizures; Sepsis; Skin and Autoimmune Disorders; Stroke; Urinary Tract Infections
Specialty Group	Anesthesiology; Diagnostic Imaging; Emergency; General Practice; Internal; Laboratory; Obstetrics and Gynecology; Pathology; Pediatric; Rehabilitation; Surgery; Other

References

Allaudeen, Nazima, et al., 2011. Inability of providers to predict unplanned readmissions. J. Gen. Intern. Med. 26, 7.

AHRQ, 2010. News and numbers Four Million Hospital Admissions Potentially Unnecessary. November 3. Agency for Healthcare Research and Quality, Rockville, MD. ⟨http://www.ahrq.gov/news/newsroom/news-and-numbers/110310.html⟩.

American Hospital Association, 2013. Fast facts on US hospitals. Retrieved 24 September from: ⟨http://www.aha.org/research/rc/stat-studies/fast-facts.shtml⟩.

Billings, John, et al., 2006. Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. Br. Med.J. 333, 7563.

Bottle, A., Aylin, P., Majeed, A., 2006. Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis. J. R. Soc. Med. 99, 8.

Boult, C., Dowd, B., McCaffrey, D., Boult, L., Hernandez, R., Krulewitch, H., 1993. Screening elders for risk of hospital admission. J. Am. Geriatr. Soc. 41, 811–817.

Bachouch, Rym Ben, Guinet, Alain, Hajri-Gabouj, Sonia, 2012. An integer linear model for hospital bed planning. Int. J. Prod. Econ. 140, 2.

Carey, K., Stefos, T., 2015. The cost of hospital readmissions: evidence from the VA. Health Care Manag. Sci., 1–8.

Charlson, M.E., Pompei, P., Ales, K.L., MacKenzie, C.R., 2008. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J. Chronic Dis. 40, 373–383.

Dunlay, Shannon M., Gersh, Bernard J., 2013. Predicting and preventing heart failure rehospitalizations: is there a role for implantable device diagnostics? Am. J. Cardiol. 111, 1.

Donzé, Jacques, et al., 2013. Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. Arch. Intern. Med. 173, 8.

Donnan, Peter T., et al., 2008. Development and Validation of a model for predicting emergency admissions over the next year (PEONY): a UK historical cohort study. Arch. Intern. Med. 168, 13.

Engoren, Milo, et al., 2013. Use of genetic programming, logistic regression, and artificial neural nets to predict readmission after coronary artery bypass surgery. J. Clin. Monit. Comput. 27, 4.

Golmohammadi, Davood, 2011. Neural network application for fuzzy multi-criteria decision making problems. Int. J. Prod. Econ. 131, 2.

Golmohammadi, Davood, Creese, Robert C., Valian, Haleh, Kolassa, John, 2009a. Supplier selection based on a neural network model using genetic algorithm. IEEE Trans. Neural Netw. 20, 9.

Golmohammadi, Davood, Creese, Robert C., Valian, Haleh, 2009b. Neural network application for supplier selection. Int. J. Prod. Dev. 8, 3.

Glenn, M., Hackbarth, J.D., (2009). Reforming America's health care delivery system. < http://www.finance.senate.gov/imo/media/doc/042109ghtest1.pdf > .

Hamner, Jenny B., Ellison, Kathy Jo, 2005. Predictors of hospital readmission after discharge in patients with congestive heart failure. J. Acute Crit. Care 34, 231–239.

IBM, 2013. Modeling nodes. Retrieved from: ⟨http://pic.dhe.ibm.com/infocenter/spssmodl/v15r0m0/index.jsp?topic=%2Fcom.ibm.spss.modeler.help%2Fmodeling_nodes.htm⟩.

Kansagara, Devan, et al., 2011. Risk prediction models for hospital readmission: a systematic review. JAMA 306, 15.

Lagoe, Ronald J., Noetscher, Cheryl M., Murphy, Mark P., 2001. Hospital readmission: predicting the risk. J. Nurs. Qual. 15, 4.

Lyon, David, et al., 2007. Predicting the likelihood of emergency admission to hospital of older people: development and validation of the emergency admission risk likelihood index (EARLI). Fam. Pract. 24, 2.

Marcantonio, Edward R., et al., 1999. Factors associated with unplanned hospital readmission among patients 65 years of age and older in a medicare managed care plan. Am. J. Med. 107, 1.

Njagi, Edmund Njeru, et al., 2013. A Joint survival-longitudinal modelling approach for the dynamic prediction of rehospitalization in telemonitored chronic heart failure patients. Stat. Model. 13, 3.

NeuroDimension Inc., 2014. What is sensitivity analysis. Retrieved from: ⟨http://www.nd.com/newsletter/Newsletter_v8_4.html⟩.

Ottenbacher, Kenneth J., et al., 2000. Length of stay and hospital readmission for persons with disabilities. Am. J. Public Health 90, 12.

Philbin, Edward F., DiSalvo, Thomas G., 1999. Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data. J. Am. Coll. Cardiol. 33, 1560–1566.

Qu, Xiuli, Shi, Jing, 2011. Modeling the effect of patient choice on the performance of open access scheduling. Int. J. Prod. Econ. 129, 2.

Reuben, David B., Keeler, Emmett, Seeman, Teresa E., Sewall, Ase, Hirsch, Susan H. M.P.H., Guralnik, Jack M., 2002. Development of a method to identify seniors at high risk for high hospital utilization. Med. Care 40, 9.

Shyu, Y., Chen, M., Lee, H., 2002. Caregiver's needs predict hospital readmission for elderly patients. Gerontologist 42, 1.

Saremi, Lireza, et al., 2013. Appointment scheduling of outpatient surgical services in a multistage operating room department. Int. J. Prod. Econ. 141, 2.

Wong, Frances K.Y., et al., 2010. What accounts for hospital readmission? J. Clin. Nurs. 19, 23.

van Walraven, Carl, et al., 2012. Comparing methods to calculate hospital-specific rates of early death or urgent readmission. Can. Med. Assoc. J. 184, 15.

van der Vaart, T., Vastag, Gyula, Wijngaard, Jacob, 2011. Facets of operational performance in an emergency room (ER). Int. J. Prod. Econ. 133, 1.