Journal of Biomedical Informatics 86 (2018) 143–148

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Predicting the risk of acute care readmissions among rehabilitation inpatients: A machine learning approach

Yajiong Xue[a], Huigang Liang[a,b,c,*], John Norbury[d], Rita Gillis[e], Brenda Killingworth[a]

[a] Department of Management Information Systems, College of Business, East Carolina University, USA
[b] Center for Healthcare Management Systems, College of Business, East Carolina University, USA
[c] Big Data and Analytics Research Cluster, East Carolina University, USA
[d] Department of Physical Medicine & Rehabilitation, Brody School of Medicine, East Carolina University, USA
[e] Vidant Health, USA

ARTICLE INFO

Keywords:
Readmission
Rehabilitation
Machine learning
Functional independence measure

ABSTRACT

Introduction: Readmission from inpatient rehabilitation facilities to acute care hospitals is a serious problem. This study aims to develop a predictive model based on machine learning algorithms to identify patients at high risk of readmission.
Methods: A retrospective dataset (2001–2017) including 16,902 patients admitted into a large inpatient rehabilitation facility in North Carolina was collected in 2017. Three types of machine learning models with different predictors were compared in 2018. The model with the highest c-statistic was selected as the best model and further tested by using five sets of training and validation data with different split time. The optimum threshold for classification was identified.
Results: The logistic regression model with only functional independence measures has the highest validation c-statistic at 0.852. Using this model to predict the recent 5 years acute care readmissions yielded high discriminative ability (c-statistics: 0.841–0.869). Larger training data yielded better performance on the test data. The default cutoff (0.5) resulted in high specificity (> 0.997) but low sensitivity (< 0.07). The optimum threshold helped to achieve a balance between sensitivity (0.754–0.867) and specificity (0.747–0.780).
Conclusions: This study demonstrates that functional independence measures can be analyzed by using machine learning algorithms to predict acute care readmissions, thus improving the effectiveness of preventive medicine.

1. Introduction

As an important indicator of healthcare quality, hospital readmission has received increasing attention from health care policy makers, payers, and providers. In 2013, 3.9 million (13.9%) patients were readmitted within 30 days of discharge in 2013, resulting in over $52 billion in healthcare expenditures [1]. In response, the Centers for Medicare & Medicaid Services (CMS) started the Hospital Readmissions Reduction Program in 2012 to penalize hospitals for excessive readmissions. In 2016, 78% of the hospitals were penalized for $420 million, and it is estimated that 79% hospitals will be penalized and the total penalties would reach $528 million in 2018 [2]. Many attempts have been made to predict readmissions with the hope that measures can be taken to avoid the readmissions, and readmission risk prediction models have received increasing attention [3].

An inpatient rehabilitation facility (IRF) is a post-acute setting for patients with medical and functional needs that cannot be met at home or at a lower level of care. Returns to the acute care hospital (RACH) occur when medical problems necessitate transfer back to the acute care hospital prior to the completion of rehabilitation at the IRF. RACH can slow patient recovery [4], and place financial burdens on the healthcare system, patients, and their families [5]. Post-acute care costed Medicare $62 billion in 2012, and the 30-day post-discharge cost for post-acute care and readmissions was almost the same as the initial hospital admission cost [6]. RACH rates are increasing in the USA [7]. Research shows that 12.4% of patients are readmitted from IRFs to acute care hospitals within 30 days of initial hospital discharge [8], and that many of these readmissions are preventable [9]. RACH is also a IRF quality metric [10]. CMS has developed the All-Cause Unplanned Readmission Measure for 30 Days Post Discharge from Inpatient Rehabilitation Facilities based on which IRFs will be compared to facilitate patients' care seeking decisions [8]. Accurate prediction of potential RACH cases can help IRFs identify high risk patients and intervene to prevent RACH.

* Corresponding author at: College of Business, Slay 303, East Carolina University, Greenville, NC 27858, USA.
E-mail address: liangh@ecu.edu (H. Liang).

https://doi.org/10.1016/j.jbi.2018.09.009
Received 8 May 2018; Received in revised form 5 September 2018; Accepted 13 September 2018
Available online 18 September 2018

While many models have been developed to predict the risk of readmissions, a review study reported that these models' discriminative ability is limited, with c-statistics ranging from 0.55 to 0.65 [3]. One possible explanation for the low predictive power is that previous models considered only demographics and comorbidities as the critical risk factors and patients' functional status variables were omitted. However, a recent study showed that even the model that considered functional status variables only improved the c-statistic to the range of 0.65–0.72 [8]. A review of the literature on RACH prediction reveals that the previous studies primarily employed traditional statistical approaches which are intended to provide explanations rather than to make predictions [8,11–14]. Although an explanatory model can identify which independent variables are significantly associated with the outcome variable, it does not necessarily make an accurate prediction of the outcome [15]. By contrast, predictive analytics develops mathematical models based on machine learning techniques with the specific goal to generate accurate predictions by selecting predictors and tuning model parameters based on various training and validation procedures [16]. If the objective is to predict RACH accurately, the machine learning approach is more appropriate than the traditional statistical approach. Unfortunately, to date research on RACH prediction by using machine learning algorithms is rare.

In addition, past research on RACH prediction has a critical limitation when evaluating model performance. The development and validation of the model are conducted by using the same data set [8,11–14]. This is problematic because the model is prone to biases due to overfitting [16]. Consequently, it is uncertain whether the model will perform well on newly unseen data. Machine learning algorithms overcome the overfitting problem by applying methods such as data partitioning and cross-validation [15]. Different models are developed on the training data set and their performance is tested against a different validation data set. The discriminative ability evaluated in this way will more realistically reflect how well the model can predict future RACH. Then the best performing model can be identified.

Given the gravity of the RACH issue, the limitation of research on RACH prediction, and the availability of large amount of digital patient data, it is imperative and feasible to conduct predictive analytics research to help IRFs predict RACH. Hence, the objective of this study is to develop a predictive model based on historical patient data so that it can be used to predict which patients are likely to be readmitted to an acute care hospital. RACH prediction is a typical classification problem. There are three major categories of classification models: linear, nonlinear, and tree-based models [16]. Since there are a number of models in each category, we will focus on three representative models – logistic regression (linear), support vector machine (nonlinear), and random forest (tree-based). Our goal is to select the best performing model and identify the optimum threshold value for classification.

2. Methods

2.1. Study setting and sample

Data were collected in 2017 from a rehabilitation center of a large health system in Eastern North Carolina (ENC). The health system is the largest provider in ENC, which includes one tertiary care medical center and seven regional community hospitals. In 2016, this health system reported 63,093 inpatient admissions; 274,039 emergency room visits; and 335,004 outpatient visits. The health system's rehabilitation center is a comprehensive rehabilitation center and the largest IRF in ENC. Its service area covers 29 counties and offers an array of rehabilitation services for patients of all ages.

We collected a 16-year (11/2001–09/2017) dataset of 16,902 patient records from the rehabilitation center. With the longitudinal data, we can train predictive models on data of early years and validate the models on data of later years. Among these patients, 1694 (10.0%) had RACH. The data set contains IRF-Patient Assessment Instrument data, including demographic, functional, and medical data. This study was approved by the relevant institutional review board.

2.2. Outcome variable and predictors

The primary outcome variable was RACH, defined as discharge from the rehabilitation center and immediate subsequent admission to an acute care hospital. Predictor variables included age, gender, race, marital status, admission impairment group, admission class, admit from, prehospital living setting, primary source of payment, number of comorbidities, and functional status at admission. Three predictors were extracted from unstructured data: age was calculated based on the difference between birth date and admission date; the impairment group was recoded into 21 groups based on the Rehabilitation Impairment Code; and the number of comorbidities was calculated by counting the recorded comorbidities. All the other predictors were in structured format. Functional status was measured within the first three days of admission to the rehabilitation center by using the standard 18-item Functional Independence Measure (FIM) [17]. Items were scored by nurses and occupational, physical, and speech therapists. Each item was rated by a 7-point ordinal scale from total assistance (1) to complete independence (7). The FIM item scores were summated into six domains, including self-care (6 items: Eating; Grooming; Bathing; Dressing – Upper; Dressing – Lower; Toileting), sphincter control (2 items: Bladder; Bowel), transfers (3 items: Bed, Chair, Wheelchair; Toilet; Tub, Shower), locomotion (2 items: Walk/Wheelchair; Stairs), communication (2 items: Comprehension; Expression), and social cognition (3 items: Social Interaction; Problem Solving; Memory). The six domain scores were also included as predictors.

2.3. Model training and validation

Data analysis was conducted in 2018. We developed three sets of predictive models (logistic regression, support vector machine, and random forest) and compared their performance. R 3.4.3 and the caret package were used for model training and validation. Logistic regression assumes a log-linear relationship between predictors and the target variable, and has been widely used to predict readmissions in the literature [3,8]. Support vector machine (SVM) is a flexible predictive model that can accommodate both linear and nonlinear decision boundaries between classes [18]. Random forest is a tree-based predictive model that constructs a number of trees using bootstrapped samples and determines the prediction probability by taking votes from each tree [19]. For each of the three models, three sub-models with different predictors was compared: the baseline model (with demographic and comorbidity variables as predictors), the FIM only model (with admission FIM variables as predictors), and the FIM plus model (with all predictors).

We split the longitudinal data into training and validation data sets based on time. The strategy was to use historical data to train the model and then use the more recent data to validate the model. This is more reasonable than a random split, because a random split could put the more current data in the training set but it is meaningless to use historical data to validate a model trained on the more current data. We plotted quarterly RTA cases over time and found no clear temporal patterns (Fig. 1), suggesting that the split based on time was justifiable. For model comparison, we split the data at Year 2016. The training data contained 2001–2015 data and the validation data contained 2016–2017 data. During model training, the model was tuned with a grid search strategy to identify the tuning parameters that led to the best model performance. For the SVM models, a grid search showed that a linear kernel outperformed nonlinear kernels. Hence, we used svmlinear and the cost parameter was set as 1. For the random forest models, the best performance was obtained when the number of trees (ntree) was 200 and the number of variables randomly sampled as candidates at each split (mtry) ranged between 12 and 16. We applied