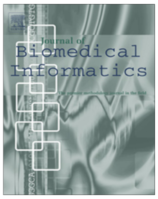
Journal of Biomedical Informatics 56 (2015) 229–238

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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



A comparison of models for predicting early hospital readmissions

Joseph Futoma [a], Jonathan Morris [b], Joseph Lucas [a,c,*]

[a] *Dept. of Statistical Science, Duke University, Box 90251, Durham, NC 27708, USA*
[b] *Quintiles, 4820 Emperor Blvd., Durham, NC 27703, USA*
[c] *Dept. of Electrical and Computer Engineering, Duke University, Box 90291, Durham, NC 27708, USA*

ARTICLE INFO

Article history:
Received 26 November 2014
Revised 13 May 2015
Accepted 24 May 2015
Available online 1 June 2015

Keywords:
Electronic Health Records
Early readmission
Penalized methods
Random forest
Deep learning
Predictive models

ABSTRACT

Risk sharing arrangements between hospitals and payers together with penalties imposed by the Centers for Medicare and Medicaid (CMS) are driving an interest in decreasing early readmissions. There are a number of published risk models predicting 30 day readmissions for particular patient populations, however they often exhibit poor predictive performance and would be unsuitable for use in a clinical setting. In this work we describe and compare several predictive models, some of which have never been applied to this task and which outperform the regression methods that are typically applied in the healthcare literature. In addition, we apply methods from deep learning to the five conditions CMS is using to penalize hospitals, and offer a simple framework for determining which conditions are most cost effective to target.

1. Introduction

Changes in federal regulation of the healthcare industry together with the novel use of payment penalties based on quality of care metrics are leading to substantial changes in business models within healthcare. The availability of large repositories of electronic health data and the continued rise of risk sharing relationships between health systems and payers have created a strong incentive to shift healthcare delivery out of the hospital setting and into lower cost, outpatient services. The double incentive of shared risk and early readmission penalties – imposed both within the United States [1] and abroad [2] – have created a strong incentive for hospital systems to identify, at the time of discharge, those patients who are at high risk of being readmitted within a short period of time.

A hospital readmission is defined as admission to a hospital a short time (typically within 30 days) after an original admission. A readmission may occur for planned or unplanned reasons, and at the same hospital as original admission or a different one. A study conducted by the Medicare Payment Advisory Committee (MedPAC) reported that 17.6% of hospital admissions resulted in readmissions within 30 days of discharge, with 76% of these being potentially avoidable [3]. In total, these readmissions accounted for $15 billion in Medicare spending. In an effort to curb hospital readmission rates, part of the Patient Protection and Affordable Care Act penalizes hospitals with excessive readmissions at 30 days through a program called the Hospital Readmission Reduction Program. In the fiscal year 2013, more than 2000 hospitals were penalized over $280 million. On October 1, 2014, the penalty increased to a minimum of 3% of a hospital's Medicare reimbursement, and also included several more conditions [1].

Hospital leaders recognize that scrutiny over readmission rates will continue to grow over the next few years, and that the financial penalties will only increase. As such, procedures for reducing readmissions have been thoroughly researched and have already started to be implemented at many hospitals. Techniques such as improving patient education, conducting followup visits or phone calls, and transferring discharge information to primary doctors may all reduce readmissions. However, individualized followups can be costly; this raises the question of which patient groups should be targeted in order to most effectively use the resources available for preventing readmissions. Methods that can accurately assess patient readmission risk are in high demand, as hospitals scramble to target the most at-risk patients and reduce their readmission rates in the most cost effective manner.

A variety of literature exists on statistical techniques for assessing patient readmission risk, using many types of available data. Some methods, such as in [4], leverage a variety of data sources, including patient demographic and social characteristics, medications, procedures, conditions, and lab tests. Other methods are

[*] Corresponding author at: Dept. of Statistical Science, Duke University, Box 90251, Durham, NC 27708, USA. Tel.: +1 919 668 3667.

E-mail addresses: jdf38@stat.duke.edu (J. Futoma), jonmorrismd@gmail.com (J. Morris), joe@stat.duke.edu (J. Lucas).

http://dx.doi.org/10.1016/j.jbi.2015.05.016
1532-0464/© 2015 The Authors. Published by Elsevier Inc.

based on only a single source of data, for instance, solely on administrative claims data, as in [5]. A thorough review of past models can be found in [6]. With the exception of [7], all of these methods are logistic regressions on independent variables typically chosen by hand.

Our aim is to compare in detail existing methods used to predict readmission with many other statistical methods. These methods include "local" models tailored to particular patient subpopulations as well as "global" models fit to the entire dataset. We compare penalized linear models as well as non-linear models such as random forests and deep learning. Due to the increased difficulty of training deep models, we conduct a smaller set of experiments to validate their performance.

The remainder of this paper will be organized as follows. Section 2 summarizes our data source. Section 3 presents a variety of statistical methods to predict patient readmissions. Section 4 introduces the experimental setup in applying these methods to hundreds of diverse groups of admissions, and summarizes the results. Section 5 compares deep neural networks to penalized logistic regression for predicting readmissions in the 5 groups that CMS (Centers for Medicare and Medicaid) is using to assign penalties. After a brief introduction to deep learning, we offer simple advice on identifying which conditions to target. We conclude in Section 6 with a brief discussion and directions for future work.

2. Data summary and processing

The dataset used is the New Zealand National Minimum Dataset, obtained from the New Zealand Ministry of Health. It consists of nearly 3.3 million hospital admissions in the New Zealand (NZ) hospital system between 2006 and 2012. New Zealand is an island nation with a national healthcare system. Because of this, we anticipate that we are losing very few patients to outside health systems. However, New Zealand uses ICD-10-AM (Australia modification) medical coding and hospitals in New Zealand are under different regulatory pressures from those in the United States. In addition, healthcare workflow and the utilization of admissions may be very different in the New Zealand healthcare environment. As such, the predictive variables and model parameters we discover will not directly translate to data from the United States. However, this paper is focused on the characteristics of the statistical models, not the learned model parameters; the results we present will be a valuable guide for modeling decisions when addressing the early readmission question with US healthcare data.

We formalize the task of predicting early patient readmissions as a binary classification task. As such, our outcome variable of interest is a binary indicator of whether or not a patient is readmitted again to the NZ hospital system within 30 days. For each visit, we have background information on the patient's race, sex, age, and length of stay. Additionally, we also know the type of facility (public or private), and whether the patient was a transfer. As noted in [5], prior admissions can be predictive of future readmissions, so we also include the number of hospital visits in the past 365 days for each patient visit.

We expect the most informative aspect of the dataset to be the large collection of ICD 10-AM codes assigned to each patient visit. Before preprocessing, this consists of 17,390 binary variables coding the precise diagnosis (12,231) and procedures (5159) relevant to each hospital admission. For each visit we also have a single Diagnosis Related Group (DRG) code, selected from a set of 815 unique DRGs which break down admissions into broader diagnoses classes than the highly specific ICD codes. Table 1 provides a brief summary of the dataset.

Before modeling, we do a small amount of preprocessing of the raw dataset. We first filter out patient visits with entry dates before 2005 and eliminate from the training/validation sets any visits that ended in the patient's death. Censored values are treated as not being readmitted within 30 days. Additionally, we combine patient visits that have overlapping admission and discharge dates; generally these represent episodes where the patient was transferred directly from one institution to another. Finally, we exclude as potential predictors in all models any ICD code that appears 10 times or fewer in the full dataset. This leaves us with a sparse 3,295,775 × 12,045 binary matrix of ICD codes, in addition to the background and demographic variables from Table 1.

Table 1
Full dataset, post-processing.

Characteristic	
Total number of admissions	3,295,775
Number of unique individuals	1,328,384
Percent readmission within 30 days	19.0
Number of unique procedures (ICD-10 AM)	3599
Number of unique diagnoses (ICD-10 AM)	8446
Number of ICD-10 AM codes per visit, mean (SD)	5.1 (3.8)
Number of unique diagnosis related groups (DRGs)	815
Variables used in prediction	
Age (years), mean (SD)	41.2 (14.1)
Male (%)	38.8
White/Islander/Asian/Hispanic/African (%)	62.6/26.9/7.1/0.2/0.5
Public facility (%)	93.9
Transfer (%)	5.6
Length of Stay (days), mean (2.5%, 25%, 50%, 75%, 97.5% quantiles)	2.9 (0, 0, 1, 3, 16)
Number of admissions in past 365 days, mean (2.5%, 25%, 50%, 75%, 97.5% quantiles)	3.7 (0, 0, 0, 1, 25)

3. Methods

3.1. DRG-specific methods

Most published approaches to the prediction of 30 day readmission focus on a single target patient population – typically those that are penalized by CMS. In order to mirror this approach and produce a large scale model comparison, we tested a variety of statistical models on 280 different patient-visit cohorts as determined by the DRGs. In the context of regression, this is equivalent to the inclusion of an interaction effect between disease groups and every predictor. Fig. 3.1 displays a histogram of sample sizes for the 280 patient cohorts we consider. In Section 3.2 we introduce methods that scale seamlessly to the entire dataset of over 3 million admissions.

For each DRG, we test 5 methods, 2 of which are novel for the task of predicting early readmission. Before modeling each group, we exclude as potential predictors any ICD code appearing 10 times or fewer in that group. Table 2 contains an abbreviation and short description of each of the DRG-specific methods considered. All models are trained on the background variables from the lower half of Table 1, as well as all the ICD codes remaining after thresholding. Note that this implies that the matrix of independent variables will be extremely sparse since on average only 5 codes are used per admission.

- **LR** The first method considered is logistic regression with a maximum likelihood estimator for the regression coefficients. Define $y_i \in \{-1, 1\}$ to indicate whether the i'th patient visit resulted in readmission within 30 days (where a 1 denotes readmission), and define $\mathbf{x_i}$ to be the sparse p-dimensional vector of independent variables for patient visit i. Maximum likelihood logistic regression involves the identification of a p-dimensional vector of regression coefficients, $\hat{\boldsymbol{\beta}}$, such that