Predicting early hospital readmissions using Electronic Health Records

Joseph Futoma
Dept. of Statistical Science
Duke University
Durham, NC 27708
jdf38@stat.duke.edu

Joseph Lucas
Dept. of Electrical and Computer Engineering
Duke University
Durham, NC 27708
joe@stat.duke.edu

1 Introduction

The United States has the third largest healthcare spending per capita in the world [1], yet this disproportionately high spending has not produced a higher quality of care. One avoidable expenditure that contributes to these skyrocketing costs is hospital readmissions. A hospital readmission is defined as admission to a hospital a short time (typically within 30 days) after an original admission. A study conducted by the Medicare Payment Advisory Committee reported that 17.6% of hospital admissions resulted in readmissions within 30 days of discharge, with 76% of these being potentially avoidable [2]. 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 2,000 hospitals were penalized over $ 280 million. Starting on October 1, 2014, the penalty increased to a minimum of 3% of a hospital's Medicare reimbursement [3].

Hospital leaders recognize that scrutiny over readmission rates will continue to grow, and that financial penalties will only increase. 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, and beg the question of which patient groups should be targeted in order to most effectively use a hospital's resources. Methods that accurately assess patient readmission risk are in high demand, as hospitals scramble to target the highest risk patients and reduce their readmission rates. A literature exists within the health informatics community on statistical techniques for assessing patient readmission risk, using different 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 based on only a single source of data, for instance, on administrative claims data, as in [5]. A thorough review of past models considered can be found in [6]. With one exception known to us [7], all methods are simple logistic regressions on independent variables typically chosen by hand. Most exhibit poor performance and would not be suitable for clinical use, unlike the methods presented here, which have better predictive power.

The remainder of this paper will be organized as follows. Section 2 describes the dataset and preprocessing. Section 3 describes work fitting multiple models to 280 disjoint patient subgroups to test model generalization. In particular, we examine the performance of models specific to each subgroup against global models trained on the entire data and tested on each subgroup. In Section 4, we train deep neural networks on the five conditions CMS is using to assess penalties. We follow this with simple advice for determining which patient groups to target in order for a hospital to best allocate its resources. Section 5 concludes with a discussion and directions for future work.

2 Data summary

The dataset used is the New Zealand (NZ) National Minimum Dataset, obtained from the NZ Ministry of Health. It consists of nearly 3.3 million hospital admissions for 1.3 million individuals in the NZ hospital system between 2006 and 2012. An important aspect of our data is that we have information on an entire healthcare system, due to the nation's small size. This, combined with NZ's status as an island nation, allows us to track patients who might be readmitted to a different hospital on a future visit, and it is likely we are missing very few admissions to other systems.

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 compile the number of hospital visits in the past 365 days of each patient, for each admission. However, the most informative aspect of the dataset is the large collection of ICD 10-AM (Australia modification) codes assigned to each patient visit. Before preprocessing, this consists of 17,390 binary variables coding the precise diagnosis (12,231) and procedures (5,159) relevant to each hospital admission. For each visit we also have the patient's Diagnosis Related Group (DRG) code, a set of 815 indicators breaking down admissions into broader diagnoses classes than the highly specific ICD codes.

Before modeling, we do a small amount of preprocessing of the raw dataset. We first filter out patient visits with entry dates before 2005 or within 30 days of Dec. 31, 2012, and remove any visits that ended in the patient's death. Censored values where a patient is never seen again are treated as not being readmitted within 30 days. Additionally, we combine patient visits that have overlapping admission and discharge dates, where the patient was transferred directly from one institution to another. Finally, we remove any ICD codes appearing 10 times or fewer in the full dataset. This leaves us with a sparse $3,295,775 \times 12,045$ binary matrix of ICD codes, in addition to the background and demographic variables mentioned.

3 Modeling a large number of patient subgroups

Initially, we fit models that include an interaction between disease groups and the predictors, where a disease group is based on DRGs. This amounts to fitting models to smaller subsets of the full data, as partitioned by DRG. We fit models to 280 DRGs, varying in size from 2,000 to 27,000 admissions, and compare their performance. We then compare these disease-specific models with "global" models fit via stochastic gradient descent on the entire data.

Table 1 contains a summary of the methods considered. All are well known in the statistics and machine learning community, although random forests and penalized approaches are novel to this application. For the logistic regression variable selection method from [5], univariate variable selection is first performed by fitting a logistic regression with each predictor, followed by a Likelihood Ratio Test. Following this, a stepwise forward variable selection is performed. For the methods trained via stochastic gradient descent (using [8]), we consider the logistic regression loss function and the modified Huber loss function [9], and use the same three elastic net penalties on coefficients as for the disease-specific penalized logistic regressions.

To evaluate each method, we use a 10-fold cross validation scheme. We train each method 10 times, each time on 90% of the DRG, withholding a different 10% for the final evaluation. Additionally, we use a separate validation set within each fold to tune parameters. Area under the ROC curve (AUC, sometimes referred to as C-statistic) is our quantitative means of comparing predictive performance among classifiers. For each DRG and for each method, we record the mean AUC across all 10 held-out test sets. We then report the mean and standard error of these 280 average AUCs for each method in Table 2. The table also reports the number of DRGs on which each method had the highest mean AUC among all methods. The random forest is overall the best method, and its mean AUC over all 280 DRGs is significantly better than all other methods (t-test, 0.01 significance level). However, both penalized logistic regression and random forests perform much better on average than the other

Table 1: Methods for predicting readmission within DRGs

Method name	Abbr.	Short Description
Logistic regression	LR	Logistic regression with no additional variable selection. Most frequently used method in literature.
Logistic regression with variable selection	LRVS	Logistic regression with variable selection, from [5]. ICD codes with low prevalence are excluded, followed by univariate and multivariate variable selection.
Penalized logistic regression	PLR	Logistic regression with elastic net regularization term, penalizing large coefficient values. Tested $\alpha = 0.01, 0.5, 1$, where $\alpha = 1$ is an L_1 penalty and $\alpha = 0$ is an L_2 penalty.
Random forest	RF	Tree-based classification method that is able to capture nonlinearity.
Support vector machine	SVM	Method maximizes margin between data points and separating hyperplane. Linear kernel and polynomial (order 3) kernel tested, following [7] who used this to predict readmissions.
"Global" methods trained via stochastic gradient descent	SGD	Logistic and modified Huber loss functions, with the three elastic net coefficient penalties from PLR, trained over the full dataset using stochastic gradient descent.

methods. There is large variability in AUC across the 280 DRGs, as it ranged from 0.57 to 0.95, with a mean of 0.69. Within DRGs there was also large variability among methods. Finally, in 221 of the 280 DRGs, the best local method outperformed the best global model fit via stochastic gradient descent, with local methods averaging 0.02 higher AUC. This indicates it is beneficial to create models tailored to specific conditions, rather than a single large model.

Table 2: Results for predicting readmission across 280 DRGs

Method name	Mean (SE) AUC across all 280 DRGs	# DRGs where method had highest mean AUC
LR	0.648 (0.004)	2
PLR (α=0.01)	0.683 (0.004)	27
PLR (α=0.5)	0.682 (0.004)	15
PLR (α=1)	0.681 (0.004)	22
RF	**0.684 (0.004)**	**80**
LRVS	0.667 (0.004)	28
SVM (linear)	0.671 (0.004)	47
SVM (poly)	0.588 (0.005)	0
SGD (log, α=0.01)	0.669 (0.004)	4
SGD (log, α=0.5)	0.670 (0.004)	3
SGD (log, α=1)	0.671 (0.004)	5
SGD (Huber, α=0.01)	0.665 (0.004)	1
SGD (Huber, α=0.5)	0.669 (0.004)	7
SGD (Huber, α=1)	0.672 (0.004)	39

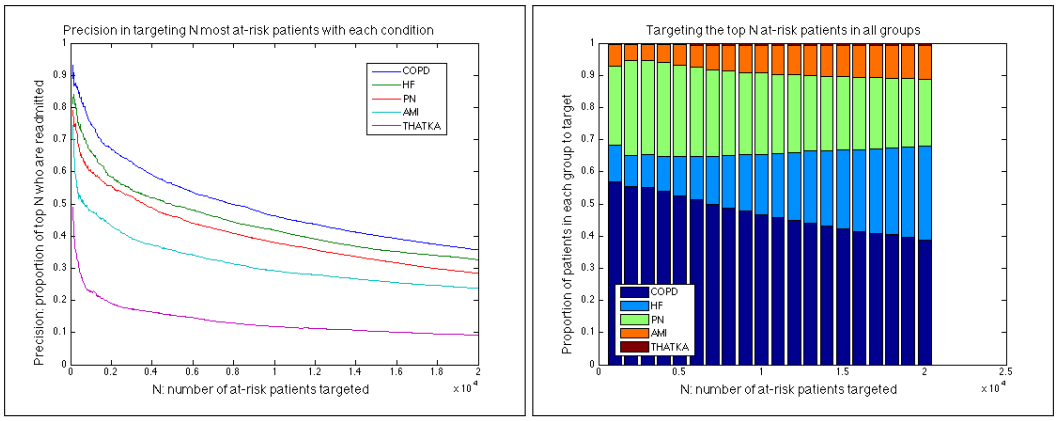

Figure 1: Left: Precision curves for penalized logistic regression models fit to the five conditions CMS penalizes. Right: Proportion of patients in each group among the overall top N at risk.

4 Readmission prevention policy

CMS penalizes hospitals that have excess readmission rates for heart failure, heart attack, and pneumonia, and starting on October 1, 2014 they also added COPD and total hip/knee replacements. This makes models for these conditions especially relevant for healthcare systems. We first fit a penalized logistic regression to each group, as it was one of the best performing and fastest models from the previous section. Additionally, for each condition we trained a deep neural network (modifying Matlab code from [10]), which are capable of capturing complicated nonlinear functions of their input. The final architectures used had three hidden layers of equal size (400 for HF, COPD, AMI, 200 for THA/TKA, and 750 for PN). We follow tips in [11], using various means (e.g. early stopping, random dropouts, sparsity and regularization penalties) with a validation set to improve training and avoid overfitting. Before fine-tuning with backpropagation, we initialized an unsupervised Deep Belief Net for 1000 epochs on each condition.

Table 3 shows the resulting AUCs per group for both methods, along with the sizes and readmission rates. The deep neural networks consistently had better AUC, but involved a substantial amount of tuning of parameters, requiring a large amount of CPU time. We also consider the precision (proportion of those designated high risk who are actually readmitted) for both methods. In the left pane of Figure 1, we plot the precision in each condition as a function of number of admissions flagged as high risk, running from 100 to 20,000. Interestingly, the precision was consistently around 1-2% higher for the penalized logistic regression, so we only display its results. Additionally, the precision curves for HF and COPD are superior to PN, which had the highest AUC, indicating AUC should not be the only metric used to evaluate a classifier. In the right pane of Figure 1, we display the proportion of patients in each condition among the overall highest risk patients. COPD makes up the largest fraction of high risk patients, followed by PN and HF. Figure 1 indicates that a hospital with limited resources to only target one condition should focus on COPD, and that it will have the most trouble targeting THA/TKA and AMI admissions.

Table 3: AUC from 10-fold cross validation on the five conditions CMS penalizes. We test a penalized logistic regression (PLR) and a deep neural network (NN).

Condition	Abbr.	Size	Readmission rate	PLR AUC: mean (SE)	NN AUC: mean (SE)
Chronic obstructive pulmonary disorder	COPD	31457	20.4%	0.703 (0.003)	**0.709 (0.003)**
Heart failure	HF	25941	19.0%	0.654 (0.005)	**0.673 (0.005)**
Pneumonia	PN	40442	27.9%	0.715 (0.005)	**0.733 (0.004)**
Acute myocardial infarction	AMI	29060	29.5%	0.633 (0.006)	**0.648 (0.007)**
Total hip arthroplasty / total knee arthroplasty	THA/TKA	23128	8.7%	0.629 (0.005)	**0.639 (0.006)**

5 Discussion

In this paper we highlighted a general framework for constructing models for assessing patient readmission risk using only ICD codes and a few background variables. Comparing our proposed methods to others that exist in the literature, we find random forests, penalized logistic regressions, and deep neural networks have significantly better predictive performance than existing methods. A future direction to explore is models that reduce the dimensionality of the original data. We tried utilizing supervised topic models as a method to reduce the dimensionality of our predictors and implicitly account for the strong correlations among ICD codes, but they didn't yield better performance. Ideally, such methods would reduce our collection of high dimensional, sparse ICD codes into a lower dimensional set of more interpretable predictors, perhaps even clustering admissions into more homogeneous subpopulations than DRGs.

There are a variety of factors involved in hospital readmissions, many of them unpredictable. Often times there may be unseen socioeconomic factors at play that are not readily available in a hospital database. However, our models do a relatively good job predicting readmissions, helping hospitals determine which patient groups to target. Future collaborations will allow us to take a model-based approach to determine useful intervention strategies for each patient cohort, extending the ideas introduced in Section 4.

References

[1] The World Bank. (2014) Health expenditure per capita. http://data.worldbank.org/indicator/SH.XPD.PCAP.

[2] Medicare Payment Advisory Committee. (2007) Report to Congress: Promoting Greater Efficiency in Medicare.

[3] Centers for Medicare and Medicaid Services. (2014) Readmissions Reduction Program. http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html.

[4] Choudhry, S.A. et al. (2013) A Public-Private Partnership Develops and Externally Validates a 30-Day Hospital Readmission Risk Prediction Model. *Online Journal of Public Health Informatics*.

[5] He, D. et al. (2014) Mining high-dimensional administrative claims data to predict early hospital readmissions. *Journal of the American Medical Informatics Association* **21**: 272-279.

[6] Kansagara, D. et al. (2011) Risk prediction models for hospital readmission: a systematic review. *Journal of the American Medical Association* **306**: 1688-1698.

[7] Yu, S. et al. (2013) Predicting Readmission Risk with Institution Specific Prediction Models. *Proceedings of the 2013 IEEE International Conference on Healthcare Informatics*, 415-420.

[8] Pedregosa, F. et al. (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**: 2825-2830.

[9] Zhang, T. (2003) Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics* **32**: 56-134.

[10] Palm, R. B. (2012) Prediction as a candidate for learning deep hierarchical models of data. *Master's thesis, Technical University of Denmark*.

[11] Hinton, G. E. (2012) Improving neural networks by preventing co-adaptation of feature detectors. http://arxiv.org/abs/1207.0580.