A Predictive Model of Patient Readmission Using Combined ICD-9 Codes as Engineered Features

Robert P. Yerex, Ph.D.,UVA Medical System

November 4, 2015

Abstract

The timing of post-discharge care is a significant factor in reducing unplanned hospital readmission. Statistical learning techniques can be applied to the development of models that predict the likelihood of patient readmission during the critical 30 day post-discharge period. The accuracy of these models is dependent on the quantity and quality of data used for training and validation. For Medicare and Medicaid patients who are members of an Accountable Care Organization (ACO), the Centers for Medicare and Medicaid (CMS) provides detailed claims based data that can be used, if appropriately collated, and transformed. This involves the identification and creation of useful features, which when included in the model, increases its predictive strength. Creation of derived features (feature engineering) is a process in which a large number of base dimensions (n) are combined to create a smaller features set $n^* \ll n$, reducing the complexity of the model while retaining its inherent information value. The inpatient admission diagnoses, in the form of ICD-9 codes, are an example of high dimensionality attributes found in the CMS claims data. Patterns inherent in the combinations of these codes can be used to create an engineered feature. In this study, a taxonomy of patterns of patient diagnoses was developed that was then used as a feature within a random survival forest model that predicts the hazard function (where the hazard event is unplanned readmission) of an individual patient for the first 30 days post discharge. Over the ensuing 30 days after release from hospital, a patient's likelihood of readmission can be dynamically estimated based on the remaining portion of the hazard curve. Inclusion of the multiple diagnoses feature increased model accuracy to the point where it could be effectively used as a tool for targeting post-discharge patient care.

1 Introduction

Reducing 30-day readmissions has become a national priority for medical personnel and government agencies. Jencks, Williams, and Coleman estimate that the cost to Medicare in 2004 of unplanned hospital readmission was $17.4 billion [19]. Thus, there have been numerous efforts to reduce readmissions, including a penalty and incentive program implemented by Centers for Medicare and Medicaid Services, or CMS [5]. Krumholz et al. examine hospital performance, as measured by 30-day readmission and mortality rates, for patients with a primary diagnosis of acute myocardial infarction (AMI) or heart failure (HF) in different parts of the country. The study concluded that 30-day rates differed among hospitals in different parts of the country, and that readmission rates in particular present a good opportunity for improvement [23].

Addressing this need, numerous studies have been conducted to develop predictive models that identify discharged patients at high risk of readmission. In a comprehensive review of 30 studies which developed predictive models for hospital readmission, Kansagara et al. conclude that most readmission risk prediction models perform poorly, corroborating the need for improvement [21]. Considering 30-day readmission rates for general surgery patients, Kassin et al. find that postoperative complications appear to drive surgical readmissions [22]. In a broader study on heart failure patients, Amarsingham et al. use a variety of predictors, such as measures of social instability and socioeconomic status, as well as EMR data and data available upon admission, to predict readmissions [1]. Additional studies into readmission targeted Veterans Affairs (VA) hospitals specifically: Kaboli et al. investigate associations between reducing length of stay for VA patients under the concern that a shorter stay would lead to increased readmissions [20]; Glasgow, Vaughn-Sarrazin, and Kaboli focus on VA patients who left against medical advice (AMA), finding that AMA patients had higher 30-day mortality and readmission rates than other discharged patients [10].

To identify discharged patients at high risk of readmission, several studies have developed risk indices. Most notably, van Walvaren et al. develop the LACE index, which uses length of stay (L), acuity of admission (A), comorbidities as calculated via Charlson score (C), and the number of ED visits in the previous six months (E) to quantify odds of readmission [29]. Gruneir et al. validate the LACE index, suggesting the tool can be used to identify candidate patients for post-discharge interventions [12]. Conversely, Cotter et al. apply LACE to a population of older patients, finding that the index was not very predictive in that population [6]. Further studies build upon LACE, including an extension by van Walvaren et al. known as LACE+ [30], and a risk score termed HOSPITAL by Donzé which includes hemoglobin, sodium at discharge, and a few other predictors [8].

In this study, we similarly attempt to identify patients at high risk of readmission. Using predictors identified solely from claims data (need more explanation here) or socioeconomic predictors available freely online, we identify which patients are at highest risk of 30-day readmission. We employ a random survival forest technique (RSF) to generate accurate predictions and provide insight into which variables are most important. Additionally, we engineer several new features from combined ICD-9 codes in a patient's claim, showing that these claims contain information that can be used in prediction.

The next section details the types of data used in the model. Section 3 discusses the feature engineering process, where we identify and create useful predictors from the data sources. We describe the modeling technique of random survival forest (RSF) in Section 4 before presenting results in Section 5. Section 6 concludes the paper.

2 Data Sources

2.1 ACO Claims Data

The Centers for Medicare and Medicaid Services, or CMS, established the Accountable Care Organization (ACO) model, where groups of health care providers could unite to form their own ACO. Providers within in an ACO work together to coordinate their care for Medicare beneficiaries. As a result of their collaboration, ACOs are eligible to receive a share of the savings that they generate for the Medicare program [5]. The University of Virginia Health System participates in the Well Virginia ACO, which serves over 20,000 Medicare beneficiaries [32].

The first set of features comes from data obtained using a patient's ACO claims. In total, ¡number¿ claims were used from ¡date1¿ to ¡date2¿. Each ACO claim contains information on a patient's stay in the hospital, including their dates of admission, dates of discharge, ages, genders, lengths of stay, discharge codes, admission codes and sequence, and other information. A complete set of information can be found in the Appendix.

2.2 Indirect Features

The second set of features in the model consists of county-level socioeconomic variables. Several studies indicate that socioeconomic status and the environments in which patients live can affect their readmission rates [3, 16, 25]. Therefore, to add additional socioeconomic data to the model, we downloaded data for the state of Virginia from the County Health Rankings website provided by the University of Wisconsin Population Health Institute [24]. These variables include demographic health data, such as the percentage of adults in each county who are obese, the percentage of adults who smoke, and the percentage of adults who are diabetic, as well as variables that relate to the level of care available in each county, such as the number of dentists and the number of primary care physicians in each county. A complete set of environmental predictors can be found in the Appendix.

3 Feature Engineering

Feature engineering, or the process of deriving features in a model from a large number of base dimensions, reduces model complexity while preserving the information value of that data. Anderson et al. provide a useful background into the importance of feature engineering, explaining common problems associated with it [2]. Guyon and Elisseeff present a strong overview of feature selection, detailing the process of selecting variables and explaining its usefulness in machine learning applications [13]. Inpatient admission diagnoses, which are provided as ICD-9 codes in the CMS claims data, are high dimensionality attributes which are candidates for mining new features. Identifying patterns or relationships among these codes can be useful to predicting 30-day readmissions.

To develop new predictors from an ACO claim, we created several features that describe the patient's stay at the hospital. In one hospitalization, patients can receive multiple diagnoses: an initial diagnosis upon admission, and subsequent diagnoses as either new claims are filed for billing purposes over an extended stay or as new diagnoses are assigned to the patients. Diagnoses are divided into admission diagnoses, which explain why a patient is admitted to the hospital, and principal diagnoses, which provide the explanatory reason for the admission after study. Studying these sequences, we identified the total number of diagnoses in each sequence and the number of unique diagnoses in each sequence as candidate predictors in the model. This totaled four additional predictors derived from the ICD-9 diagnosis codes in a patient's claim.

Many of the predictors in the ACO Claims data, such as the admission and discharge codes for the patients, contained sequences of ICD-9 codes.
Talk about ICD-9 codes and the dimensionality of them
Explain how these codes can be used in sequence
Specify the creation of the new predictor variables from the admission and discharge sequences

4 Modeling Techniques

4.1 Random Survival Forest

Random forest (RF) models, developed by Breiman [4], consist of ensembles of tree-based classifiers. In a regression or classification random forest, multiple trees are grown which output a numeric value in the case of regression, or a binary vote, 0 or 1, in a classification model. Averaging over the output of the trees gives the output of the random forest.

Random survival forests (RSF), proposed by Ishwaran et al., consist of an ensemble learning approach to right-censored survival data [18]. Essentially, RSF models are the survival analysis analog of general random forest (RF) techniques. In an RSF, each tree calculates a cumulative hazard function (CHF); averaging over the CHFs for every tree gives an ensemble CHF. Using the ensemble CHF, one can predict survival times for patients, and evaluate the model using traditional metrics such as the concordance index (CI).

Random survival forest models have been used in a variety of health applications, including in studies on esophageal cancer [27, 26], as well as studies on patients with systolic heart failure [15] and Fontan patients who undergo cardiopulmonary exercise testing [7]. RSF models have even been used outside of the health domain in a credit risk management model for small medium enterprises [9]. However, while random forest techniques have been used in a 30-day readmission setting previously [31], random survival forest techniques had not been applied to a 30-day readmission problem. In this paper, we apply RSF models to the 30-day readmission setting, using the `randomForestSRC` package in R [17].

Used RSF, subsetted the variables using a penalized survival function via L_1 regularization. (Question - to include this or not? Or include it in the Results section?)

5 Results

The initial data frame contained n rows, one for each patient, and k predictor variables. We used the `penalized` package in R to conduct an L_1 penalized survival analysis, which eliminated many features from the candidate feature

Figure 1: This frog was uploaded via the project menu.

set [11]. Since lasso (L_1) regularization tends to remove a predictor when it is highly correlated with another [28], we added predictors back into the data frame since RF models can derive additional information from correlated predictors. This generated the final data frame of m rows and i columns.

The data were trained on a 2/3 sample of the data and tested on the remaining 1/3. However, these were not random samples, as the training data ran from ¡date1¿ to ¡date2¿ and the testing data covered ¡date3¿ - ¡date4¿. Thus, we examine the model performance after training on historical data and testing on future data. We evaluate the model using an ROC curve of the predicted number of days until readmission, built with the `survivalROC` package in R [14].

6 Conclusion

7 Some examples to get started

7.1 How to add Comments

Comments can be added to your project by clicking on the comment icon in the toolbar above. To reply to a comment, simply click the reply button in the lower right corner of the comment, and you can close them when you're done.

7.2 How to include Figures

First you have to upload the image file from your computer using the upload link the project menu. Then use the includegraphics command to include it in your document. Use the figure environment and the caption command to add a number and a caption to your figure. See the code for Figure 1 in this section for an example.

7.3 How to add Tables

Use the table and tabular commands for basic tables — see Table 1, for example.

Item	Quantity
Widgets	42
Gadgets	13

Table 1: An example table.

7.4 How to write Mathematics

LaTeX is great at typesetting mathematics. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i}^{n} X_i$$

denote their mean. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.

7.5 How to create Sections and Subsections

Use section and subsections to organize your document. Simply use the section and subsection buttons in the toolbar to create them, and we'll handle all the formatting and numbering automatically.

7.6 How to add Lists

You can make lists with automatic numbering . . .

1. Like this,
2. and like this.

. . . or bullet points . . .

- Like this,
- and like this.

We hope you find Overleaf useful, and please let us know if you have any feedback using the help menu above.

References

[1] Ruben Amarasingham, Billy J Moore, Ying P Tabak, Mark H Drazner, Christopher A Clark, Song Zhang, W Gary Reed, Timothy S Swanson, Ying Ma, and Ethan A Halm. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Medical Care*, 48(11):981–988, 2010.

[2] Michael Anderson, Dolan Antenucci, Victor Bittorf, Matthew Burgess, Michael J Cafarella, Arun Kumar, Feng Niu, Yongjoo Park, Christopher Ré, and Ce Zhang. Brainwash: A data system for feature engineering. In *6th Biennial Conference on Innovative Data Systems Research (CIDR '13)*, 2013.

[3] Alicia I Arbaje, Jennifer L Wolff, Qilu Yu, Neil R Powe, Gerard F Anderson, and Chad Boult. Postdischarge environmental and socioeconomic factors and the likelihood of early hospital readmission among community-dwelling medicare beneficiaries. *The Gerontologist*, 48(4):495–504, 2008.

[4] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] cms.gov. Accountable Care Organizations (ACO) - Centers for Medicare and Medicaid Services, 2015.

[6] Paul E Cotter, Vikas K Bhalla, Stephen J Wallis, and Richard WS Biram. Predicting readmissions: poor performance of the LACE index in an older UK population. *Age and Ageing*, 41(6):784–789, 2012.

[7] Gerhard-Paul Diller, Alessandro Giardini, Konstantinos Dimopoulos, Gaetano Gargiulo, Jan Müller, Graham Derrick, Georgios Giannakoulas, Sachin Khambadkone, Astrid E Lammers, Fernando Maria Picchio, et al. Predictors of morbidity and mortality in contemporary fontan patients: results from a multicenter study including cardiopulmonary exercise testing in 321 patients. *European Heart Journal*, page ehq356, 2010.

[8] Jacques Donzé, Drahomir Aujesky, Deborah Williams, and Jeffrey L Schnipper. Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA Internal Medicine*, 173(8):632–638, 2013.

[9] Dean Fantazzini and Silvia Figini. Random survival forests models for sme credit risk measurement. *Methodology and Computing in Applied Probability*, 11(1):29–45, 2009.

[10] Justin M Glasgow, Mary Vaughn-Sarrazin, and Peter J Kaboli. Leaving against medical advice (ama): risk of 30-day mortality and hospital readmission. *Journal of General Internal Medicine*, 25(9):926–929, 2010.

[11] Jelle Goeman, Rosa Meijer, and Nimisha Chaturvedi. *penalized: L1 (lasso and fused lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model*, 2014. R package version 0.9-45.

[12] Andrea Gruneir, Irfan A Dhalla, Carl van Walraven, Hadas D Fischer, Ximena Camacho, Paula A Rochon, and Geoffrey M Anderson. Unplanned readmissions after hospital discharge among patients identified as being at high risk for readmission using a validated predictive algorithm. *Open Medicine*, 5(2):e104, 2011.

[13] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[14] Patrick J. Heagerty and packaging by Paramita Saha-Chaudhuri. *survivalROC: Time-dependent ROC curve estimation from censored survival data*, 2013. R package version 1.0.3.

[15] Eileen Hsich, Eiran Z Gorodeski, Eugene H Blackstone, Hemant Ishwaran, and Michael S Lauer. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circulation: Cardiovascular Quality and Outcomes*, 4(1):39–45, 2011.

[16] Jianhui Hu, Meredith D Gonsahn, and David R Nerenz. Socioeconomic status and readmissions: evidence from an urban teaching hospital. *Health Affairs*, 33(5):778–785, 2014.

[17] H. Ishwaran and U.B. Kogalur. *Random Forests for Survival, Regression and Classification (RF-SRC)*, 2015. R package version 1.6.1.

[18] H. Ishwaran, U.B. Kogalur, E.H. Blackstone, and M.S. Lauer. Random survival forests. *Ann. Appl. Statist.*, 2(3):841–860, 2008.

[19] Stephen F Jencks, Mark V Williams, and Eric A Coleman. Rehospitalizations among patients in the medicare fee-for-service program. *New England Journal of Medicine*, 360(14):1418–1428, 2009.

[20] Peter J Kaboli, Jorge T Go, Jason Hockenberry, Justin M Glasgow, Skyler R Johnson, Gary E Rosenthal, Michael P Jones, and Mary Vaughan-Sarrazin. Associations between reduced hospital length of stay and 30-day readmission rate and mortality: 14-year experience in 129 veterans affairs hospitals. *Annals of Internal Medicine*, 157(12):837–845, 2012.

[21] Devan Kansagara, Honora Englander, Amanda Salanitro, David Kagen, Cecelia Theobald, Michele Freeman, and Sunil Kripalani. Risk prediction models for hospital readmission: a systematic review. *JAMA*, 306(15):1688–1698, 2011.

[22] Michael T Kassin, Rachel M Owen, Sebastian D Perez, Ira Leeds, James C Cox, Kurt Schnier, Vjollca Sadiraj, and John F Sweeney. Risk factors for 30-day hospital readmission among general surgery patients. *Journal of the American College of Surgeons*, 215(3):322–330, 2012.

[23] Harlan M Krumholz, Angela R Merrill, Eric M Schone, Geoffrey C Schreiner, Jersey Chen, Elizabeth H Bradley, Yun Wang, Yongfei Wang, Zhenqiu Lin, Barry M Straube, et al. Patterns of hospital performance in acute myocardial infarction and heart failure 30-day mortality and readmission. *Circulation: Cardiovascular Quality and Outcomes*, 2(5):407–413, 2009.

[24] University of Wisconsin Population Health Institute. County Health Rankings and Roadmaps, 2015.

[25] Saif S Rathore, Frederick A Masoudi, Yongfei Wang, Jeptha P Curtis, JoAnne M Foody, Edward P Havranek, and Harlan M Krumholz. Socioeconomic status, treatment, and outcomes among elderly patients hospitalized with heart failure: findings from the national heart failure project. *American Heart Journal*, 152(2):371–378, 2006.

[26] Thomas W Rice, Valerie W Rusch, Hemant Ishwaran, and Eugene H Blackstone. Cancer of the esophagus and esophagogastric junction. *Cancer*, 116(16):3763–3773, 2010.

[27] Nabil P Rizk, Hemant Ishwaran, Thomas W Rice, Long-Qi Chen, Paul H Schipper, Kenneth A Kesler, Simon Law, Toni EMR Lerut, Carolyn E Reed, Jarmo A Salo, et al. Optimum lymphadenectomy for esophageal cancer. *Annals of Surgery*, 251(1):46–50, 2010.

[28] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011.

[29] Carl van Walraven, Irfan A Dhalla, Chaim Bell, Edward Etchells, Ian G Stiell, Kelly Zarnke, Peter C Austin, and Alan J Forster. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Canadian Medical Association Journal*, 182(6):551–557, 2010.

[30] Carl van Walraven, Jenna Wong, and Alan J Forster. Lace+ index: extension of a validated index to predict early death or urgent readmission after hospital discharge using administrative data. *Open Medicine*, 6(3):e80, 2012.

[31] Michael Vedomske, Donald E Brown, and James H Harrison. Random forests on ubiquitous data for heart failure 30-day readmissions prediction. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, volume 2, pages 415–421. IEEE, 2013.

[32] Well Virginia. Well Virginia ACO, 2015.