Prediction of High-Cost Hospital Patients

Jonathan M. Mortensen, Linda Szabo, Luke Yancy Jr.

Introduction

In the U.S., healthcare costs are rising faster than the inflation rate, and more rapidly than other first world countries. There is a national effort not only to improve medical care, but also to reduce medical care costs through use of evidence-based medicine. Electronic health records (EHRs) present an opportunity from which to find such evidence. In support of these efforts, many hospitals have recently created large de-identified data sets for use in developing methods for evidence-based practice. The Stanford Translational Research Integrated Database Environment (STRIDE) is one such EHR database, containing 35 million discharge codes and de-identified clinical notes on over 1.8 million patients who received care at the Stanford University Medical Center beginning in 1995. Recent work by Moturu and colleagues[3] uses such structured patient data to predict high-cost patients. The authors suggest these predicted patients are candidates for additional preventative interventions, thereby reducing cost.

The goal of this project is to predict the future one-year cost of a patient using features extracted from 6 months of textual clinical notes and discharge codes in the EHR. We develop a system that leverages the STRIDE data to predict high-cost patients. This system will enable hospitals and other interested parties to automatically identify patients likely to return for costly procedures in the future and invest in preventative care measures to both reduce costs and improve health.

Methods

Our system uses de-identified patient information, both textual and structured, as features to predict cost. The components of the method are (1) **feature engineering** to capture the most relevant features for the task, (2) **cost assessment** for the prediction interval, and (3) **classification that identifies high-cost patients** at a given cost threshold. *Figure 1 and Figure 2* provide an overview of this system.

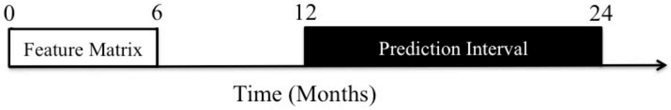

Figure 1. Month 0 indicates a patient's first encounter. The feature matrix is composed of those clinical notes and discharges codes within the 0-6 month time interval for all patients. The goal of the model is to predict the total cost associated with discharge codes found within the prediction interval (months 12 to 24).

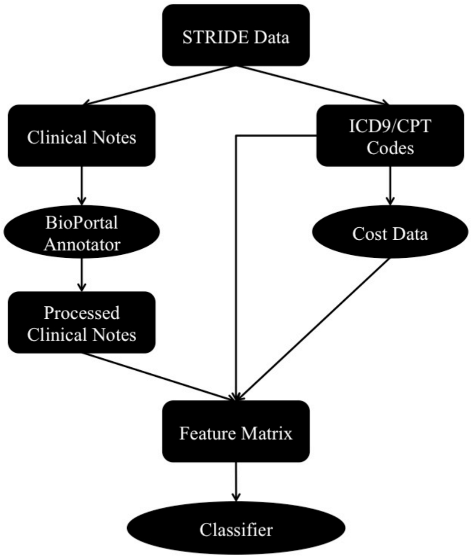

Figure 2. Here, the feature matrix is composed such that the predictors are concept frequencies (from clinical notes), in addition to ICD9 and CPT codes from the date of initial encounter to month 6. The response is the aggregated cost of all CPT codes observed during the prediction interval of a patient (months 12 to 24).

Feature engineering: Patient visits, months 0 – 6

The feature matrix contains frequencies of "concepts" and structured codes obtained from the first 6 months of clinical notes for each patient with at least 24 months of data (see *Figure 1*). We process each clinical note via the BioPortal Annotator to create a list of unique words found in the corpus of free-text clinical notes. We include words identified in the context of patient family history since we believe these provide additional predictive power. We remove negated words and words with 3 or less characters, which are often functional words. Next, we map terms to more general "concepts" from medical ontologies using the Unified Medical Language System (UMLS), a comprehensive metathesaurus of medical terminology. For example, all terms conceptually related to "diabetes" are normalized to a single concept "diabetes". In doing so, we aggregate potentially low-signal terms into a more representative single feature. Finally, we remove noisy, low-frequency concepts (occurring fewer than 50 times in the database). For structured codes, we extract CPT (Current Procedural Terminology) and ICD-9 (International Classification of Diseases) codes for each patient. The resulting feature matrix consists of 90,532 features for 96,176 patients.

Cost assessment: Patient visits, months 12 – 24

To estimate the cost of a patient, we collect all CPT codes associated with a patient's prediction interval (12-24 months from initial encounter). We skip months 6-12 since this is the timeframe for possible medical intervention. In other words, a doctor may not be able to avoid a high-cost procedure tomorrow but could initiate preventative care over a 6-month period to reduce future costs. We filter duplicate CPT codes entered on the same day for a patient or for the same visit. We then use these codes as a proxy for patient cost by mapping each patient's CPT codes to cost (in dollars) and summing up the total cost associated with these codes. We utilize cost mappings from the 2012 Medicare Physician Fee Schedules from the state of Illinois[1] since California data is unavailable. Although we cannot make conclusions about exact patient costs, we are confident that this cost data allows us to make relative comparisons to identify the most costly patients.

As shown in *Figure 3*, one-year costs range from $6 to $66,084 per person; however, 95% of the patients have costs of less than $5,000. Understanding this skew in the response variable is important when selecting and applying machine-learning methods for classification.

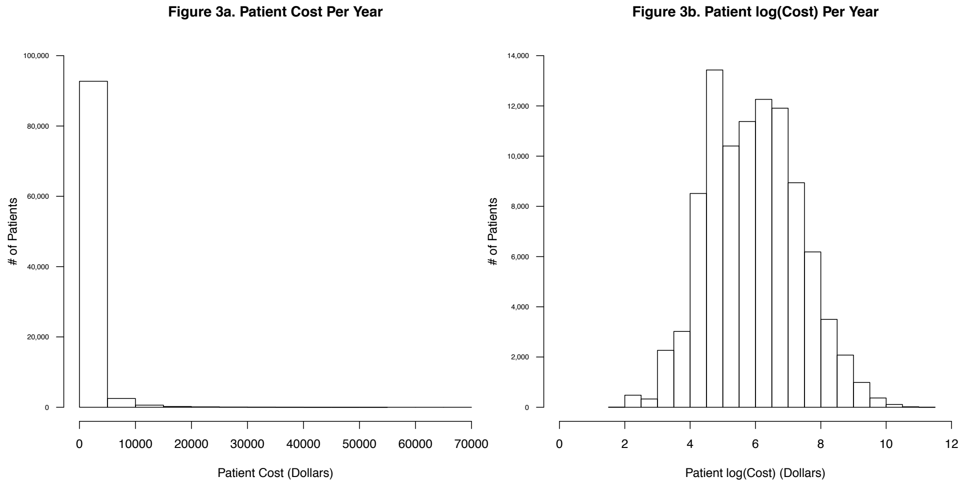

Figure 3a. Patient Cost Per Year

Figure 3b. Patient log(Cost) Per Year

Figure 3.
(a) Patient cost during prediction interval.
(b) log of patient cost during prediction interval.

Classification:
Using the textual and coded clinical information as predictors, and the patient cost data as response, we classify "high cost" patients. In each classifier, the response is the patient's total cost (or log cost) during their "prediction interval" (encounters between 12 months and 24 months; see *Figure 1*). To evaluate these methods, we train a model on a subset of the data and measure performance by the model's ability to maximize the percentage of total cost captured while minimizing the number of patients associated with high costs in the test set (i.e. identify a small number of patients who are responsible for a large percentage of the cost).

Results

To explore the space, we apply a battery of classifiers to predict cost -- Naïve-Bayes, SVM, logistic regression, regularized linear regression, and KNN. Considering the large sample size, we follow the standard practice of 60% training, 10% tuning (to select optimal parameter values), and 30% testing. Since we have 90,532 features, we implement information gain or regularization to reduce the feature space to those most informative for a given classifier. *Table 1* reports the top features. In some situations, performance (measured by accuracy of classifier as well as observation of intuitively meaningful features selected) improves when we use under sampling to obtain balanced classes. A performance summary of each classifier is shown in *Table 2*.

Information Gain	Regularized Linear Regression
Transplantation	Difficulty Kneeling
Platelet Count Measurement	Stimulant Abuse
X-Ray Computed Tomography	Large Nose
Liver	Blood Iron Measurement
Creatinine	Anorectal Abscess
Scanning	Multiple Pulmonary Embolisms
Radionuclide Imaging	Cast Brace
Albumins	Concussion, Severe
Count	Nail Problem
Phase	Non-Pyogenic Meningitis

Table 1. Top 10 features determined from the two different feature selection methods used.

	Performance	# of Features	Specificity	Patients Classified High Cost	Total Cost Identified
Naïve-Bayes	Poor	[redacted]	[redacted]	[redacted]	[redacted]
SVM	Poor	[redacted]	[redacted]	[redacted]	[redacted]
Linear Reg.	Poor	[redacted]	[redacted]	[redacted]	[redacted]
Logistic Reg.	Moderate	170	97.34%	4.28%	17.23%
Linear Reg. (log cost)	Moderate	24,219	94.49%	6.46%	14.57%
KNN	Best	300	98.00%	2.45%	11.20%

Table 2. Machine Learning methods applied to the data set. Statistics not reported for methods that did not perform well.

Naïve-Bayes We first apply a Naïve-Bayes classifier to our dataset since it often performs well for text-based classification. We expect low accuracy for patients near the "high cost" threshold as we are binning a continuous variable. However, we find that it also misclassifies patients with extreme cost values. We observe low specificity over a range of "high cost" thresholds, with a dramatic spike at a threshold of $10,500. This is likely due to the high density of patients with cost less than $10,000 and a paucity of patients at any given cost above that. Note that we use information gain to filter features before classification. We find that varying the number of features included from 50-10,000 has little affect on accuracy.

SVM The major issue with SVM is the computational complexity of selecting optimal parameters for our large dataset. Ideally, we would perform grid-search to select optimal values for parameters gamma and C. However, this process is too computationally intensive due to the large number of samples and features in our data (using R package e1071[2], tuning did not complete after 24 hours). We use information gain to reduce the number of features but the number of patients remains an issue. We try tuning on a subset of training data by spot-checking a few parameter values, but this understandably does not yield a good model.

Logistic Regression We apply regularized logistic regression at various cost thresholds to our dataset. Logistic regression develops class probabilities based on the training distribution, so classifying this dataset is especially difficult given the importance of identifying infrequent but high cost patients. To correct for this difficulty, we under-sample the training set to balance the number of low cost patients and high cost patients. Even with this additional step, regularized regression still performs poorly based on our performance metrics.

Linear Regression The skewed data is not appropriate for standard linear regression. Instead, we apply regularized linear regression to the log cost, which is approximately normal. In this model, we identify as patient as high cost if their cost is greater than 1.5 standard deviations from the mean log cost of the training data. The trained model selected only 24,219 features (the top 10 features can be found in *Table 1*). With this model, we achieve 94.49% specificity (CI: 94.21, 94.76) and 19.21% sensitivity (CI: 17.50, 21.02). The model classifies 6.47% of patients as high cost patients; these patients' costs account for about 14.57% of the total cost for a year.

K-Nearest Neighbors We apply KNN to develop a classifier using $10,500 as the "high-cost" threshold given the distribution of the cost data (*Figure 3*) and the peak in specificity observed at this threshold for Naïve Bayes. We expect this model to perform well due to the large number of patients in our training set. We use information gain to filter features, and then evaluate classification performance on the test data with values of k ranging from 1 to 140. We also vary the number of features considered, using the top 50 to 1500. Optimal performance (smallest number of patients accounting for highest percentage of cost) is achieved with 300 features and k=90, reporting 98% specificity (95% CI: 97.88, 99.12). Using this classifier, 2.45% of patients are classified as high cost, accounting for 11.2% of the total cost for patients for 1 year.

Discussion

The final two models (Linear Regression and KNN) suggest that by using our method, a hospital could identify and intervene on approximately 1 in every 40 patients to potentially reduce ~11% of its yearly cost. This serves as a baseline for addressing the cost prediction issue. However, there are many opportunities for improvement.

Complex Data A major element of this project was the process of understanding the data. As we had a large amount of data, we carefully considered what data to use. For example, when examining CPT codes marked as "billing codes," we manually reviewed a sample of the data and noted that these codes marked for billing were repeats of previous codes or that these codes only appeared after 2007. To reduce noise, we removed them. This dataset also contained many "concepts" that appeared only within a few patient's clinical notes. These concepts would often appear to be very significant features according to our model. For example, the concept "gannet" appeared as a significant feature with one of our models. Gannet, a seabird, is likely is not related to the patient's medical status. To remove such noise, we filtered low frequency concepts from the feature set. The size and complexity of the data also affected classifier selection. Many classifiers did not work "out-of-the-box" for such data; each requiring tuning, optimization, and large computational resources without which initial results are dismal.

Ontologies More than half of the top features were clinical note concepts, not discharge codes. This suggests that our results depended heavily on the ability to combine terms describing the same thing. Without using relationships defined in medical ontologies, it is unlikely we would have found any signal. In future work, the ontologies could enable further hierarchical aggregation of both concepts and codes. For example, "Diabetes I" and "Diabetes II" would both contribute to the frequency of a concept called "Diabetes", the parent of both concepts.

Applicability In this work, we addressed the task of identifying high cost patients at a reasonable threshold. We did not address the question of whether this cost could be reduced. To fully explore this problem, one would need to collaborate with a hospital and clinicians to determine the optimal cost threshold and the desired percentage of patients on which to intervene. Balancing the reduction of patient costs with added intervention costs and balancing the likelihood of successful intervention with its cost to intervene is an important situation to consider in a true application. For example, to fully address this issue, one might collaborate with a physician to identify patients that are "intervene-able" and have the potential to largely reduce their future predicted cost.

Conclusion

Considering the rapidly growing healthcare costs in the U.S., reducing such costs is paramount. In this work, we developed a method that identifies high cost patients. This method uses textual and coded data from 6 months of patient encounters to predict the future one-year cost of a patient. Using this method, we can identify 2.45% of patients that constitute 11.2% of total patient cost. In a clinical setting, a computer could automatically flag patients for physicians and care providers as opportunities for them to suggest additional preventive measures.

Acknowledgements

We would like to thank Dr. Nigam Shah for providing access to de-identified patient data from STRIDE and for assistance in defining the project goals.

References:

1. *2012 Medicare Physician Fee Schedules: Revised zero percent 2012 Medicare physician fee schedules*. (2012) Retrieved from Wisconsin Physicians Service Insurance Corporation website: http://www.wpsmedicare.com/part_b/fees/physician_fee_schedule/2012-fee-schedule.shtml
2. Meyer D, et al. (2012) Misc Functions of the Department of Statistics (e1071). R Project.
3. Moturu, S. T., Johnson, W. G., & Liu, H. (2010). Predictive risk modelling for forecasting high-cost patients: a real-world application using Medicaid data. *International Journal of Biomedical Engineering and Technology*, *3*(1/2), 114.