Casting a Wider Net: Data Driven Discovery of Proxies for Target Diagnoses

Dusan Ramljak, Adam Davey PhD, Alexey Uversky, Shoumik Roychoudhury, Zoran Obradovic PhD[1]

[1]Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA, USA

Abstract

Background: The Hospital Readmissions Reduction Program (HRRP) introduced in October 2012 as part of the Affordable Care Act (ACA), ties hospital reimbursement rates to adjusted 30-day readmissions and mortality performance for a small set of target diagnoses. There is growing concern and emerging evidence that use of a small set of target diagnoses to establish reimbursement rates can lead to unstable results that are susceptible to manipulation (gaming) by hospitals.

Methods: We propose a novel approach to identifying co-occurring diagnoses and procedures that can themselves serve as a proxy indicator of the target diagnosis. The proposed approach constructs a Markov Blanket that allows a high level of performance, in terms of predictive accuracy and scalability, along with interpretability of obtained results. In order to scale to a large number of co-occuring diagnoses (features) and hospital discharge records (samples), our approach begins with *Google's* PageRank algorithm and exploits the stability of obtained results to rank the contribution of each diagnosis/procedure in terms of presence in a Markov Blanket for outcome prediction.

Results: Presence of target diagnoses acute myocardial infarction (AMI), congestive heart failure (CHF), pneumonia (PN), and Sepsis in hospital discharge records for Medicare and Medicaid patients in California and New York state hospitals (2009-2011), were predicted using models trained on a subset of California state hospitals (2003-2008). Using repeated holdout evaluation, we used ~30,000,000 hospital discharge records and analyzed the stability of the proposed approach. Model performance was measured using the Area Under the ROC Curve (AUC) metric, and importance and contribution of single features to the final result. The results varied from AUC=0.68 (with SE<1e-4) for PN on cross validation datasets to AUC=0.94, with (SE<1e-7) for Sepsis on California hospitals (2009 – 2011), while the stability of features was consistently better with more training data for each target diagnosis. Prediction accuracy for considered target diagnoses approaches or exceeds accuracy estimates for discharge record data.

Conclusions: This paper presents a novel approach to identifying a small subset of relevant diagnoses and procedures that approximate the Markov Blanket for target diagnoses. Accuracy and interpretability of results demonstrate the potential of our approach.

Objective

Identify a small subset of diagnoses (Markov Blanket) that can serve as highly accurate proxies for the set of target diagnoses used to establish reimbursement rates under the Hospital Readmissions Reduction Program (HRRP). Identification of these subsets has applications to problems such as providing more stable hospital quality estimates, estimating the extent of diagnostic "gaming," identification of potential "upcoding" or fraudulent claims, and fuller understanding networks of diseases and medical procedures.

Introduction

As of October 1, 2012, §3025 of the Affordable Care Act (ACA) dictates that hospital reimbursements have to be based on performance relative to preventable 30-day Medicare hospital readmission rates observed in hospitals with similar predicted risk profiles. Three specific diagnoses in particular are used to track reimbursement rates: acute myocardial infarction (AMI), congestive heart failure (CHF), and pneumonia (PN). In October 2014 chronic obstructive pulmonary disease (COPD) was added to the list of diagnoses with two additional procedures, but we did not include it in our experiments since its value in this context has been criticized[29]. This policy change may have introduced an incentive for hospitals to under-diagnose these illnesses by substituting related diagnoses for which they will not be held accountable. Burgess and Hockenberry (2014)[31] provide a historical perspective on current attention to hospital readmissions and consider the potential for gaming of readmissions, but not via subtle changes in diagnoses themselves. In addition to the three diagnoses we decided to analyze (AMI, CHF, and PN), we included sepsis as a diagnosis for studying in our experiments, given that it is one of the most prevalent diagnoses, and is currently the diagnosis with the highest mortality rate in the US.

Our method is aimed at approximating Markov Blankets consisting of small subsets of diagnoses and procedures that frequently co-occur with, and can shield, each of our four target diagnoses from the rest of the disease/procedure network (in other words, the Markov Blankets for each target diagnosis node consist of diagnosis and procedure nodes that contain the only information required to accurately predict the behavior of the target diagnosis in question). We chose to use a Markov Blanket for this task since several studies have already shown that using only a small set of features that constitute the Markov Blanket for a dependent variable is sufficient to accurately predict the value of the variable[6,7,8]. Hence, each target diagnosis can be accurately identified and inferred from a small subset of related diagnoses and procedures, which can then be used to identify true cases with target diagnoses, estimate the extent of gaming via substitute diagnoses/procedures, and also suggest related sets of diagnoses and procedures which, in combination, may provide more stable methods for setting reimbursement rates. We have observed that the most frequent co-occurrence by itself is necessary, but not sufficient, to establish an accurate approximation of a Markov Blanket.

We built a directed weighted network of diagnoses and procedures from the hospital records that contain our target diagnoses, and adjusted the weights in that network according to co-occurrences of diagnoses and procedures in the records that don't contain our target diagnoses. We apply Google's PageRank algorithm and use the PageRank value as a criterion to identify important nodes which will belong to the Markov Blanket, and set the number of the "important" nodes to be selected based on the distribution of the number of diagnoses and procedures in the records.

We attained preliminary results that were accepted for presentation at SDM-DMMH 2015 Workshop[24] which were tested on a very small subset of CA data. We have since then refined our approach to make it more scalable, and evaluated its validity on a much larger set of data, spanning two states (CA and NY) and several years. We found that our approach is both accurate and stable in both states, even when trained on a relatively small sample of CA dataset. Given its simplicity and interpretability, we are confident that it can be used effectively for any target diagnosis, not just the four that we focus on in our experiments.

Background

As a direct result of the change in the structure of Medicare reimbursements, there is now more focus on problems such as the ability of health care providers to identify changing predictors of 30-day hospital readmissions[1,2,3], as well as to identify characteristics of individuals and providers associated with above-average levels of readmission risk. Hospitals that perform below expectations will see a reduction of up to 1% in Medicare-based reimbursements for services related to all diagnostic-related groups (DRGs). Based on performance levels in 2010, these targets would have placed half of all hospitals in the under-performing group. In coming years, additional diagnoses will be added to the list used to determine reimbursement rates. A focus on 30-day readmission rates has been criticized for a variety of reasons, including concerns about the validity of diagnoses, sparse evidence that decreased readmissions translate into improved health outcomes, an assumption that most readmissions are preventable, and disproportionate penalization of hospitals serving "safety-net" hospitals[29]. Joynt and Jha (2013)[32] found evidence that large teaching hospitals and safety-net hospitals are more likely to be penalized under the HRRP. Kansagara et al. (2011)[4] performed a systematic review of readmission risk models. They found that specific medical diagnoses were the most universally used predictors appearing in 24 of 26 prediction models they considered. This same study also found that the range of the Area Under the ROC Curve (hereinafter AUC) metric for predicting readmission ranged from 0.50 (nurse/case manager predicted risk of readmission) to 0.83 (administrative model plus self-report).

Several potential methods of "gaming" have been suggested, including enriching the population admitted with a target diagnosis with individuals assessed to have low readmission risk, using extended "holding areas" for patients to receive hospital care without being readmitted, and selectively coding target diagnoses among patients with expected low readmission risk. Such gaming strikes us as particularly likely given the narrow range of criteria that factor into reimbursement rates under the HRRP[29]. Because hospitals cannot be penalized for diagnoses they do not make, physicians are incentivized to choose similar, but distinct, diagnoses for criterion diagnoses. For example, a patient who is admitted to a hospital with AMI may initially be diagnosed as having chest pains or coronary atherosclerosis. If this patient was subsequently readmitted within the following 30 days, this diagnosis could not be used to penalize the hospital for poor performance. Similarly, PN may initially be diagnosed as acute bronchitis or an upper respiratory infection, and CHF may instead be diagnosed at first as chronic obstructive pulmonary disease. In addition to studying the three aforementioned diagnoses used by the ACA, we also study Sepsis using our approach, since it is one of the most prevalent diagnoses, and is currently the diagnosis with the highest mortality rate in the US. It can be diagnosed initially as a bacterial infection, pneumonia, urinary tract infection, peritonitis, or a skin ulcer. In practice, only those assessed as having the lowest risk of 30-day readmission may be likely to receive the target diagnosis.

However, several specific diagnoses are likely to co-occur with the target diagnosis, and some procedures (e.g., angioplasty) may be strongly indicative of a specific underlying true diagnoses (e.g., AMI), serving as good proxy indicators of the true diagnosis. Evidence for changes to clinical practice, diagnoses, and associated procedures in response to changes in reimbursement has been well-documented for more than 30 years[5] and there is reason to suspect that similar changes are already occurring due to the most recent changes enacted under the ACA. Rothberg et al. (2014)[19] suggested that, by more liberally applying diagnoses of sepsis and respiratory failure, hospitals might improve their reported performance under the HRRP. Another way of estimating the extent of these changes, and identifying cases that represent the true diagnoses of criterion diagnoses, is considering diagnoses as a set of connected nodes in a graph (connected by aspects such as co-occurrence). The Markov Blanket (MB) for a node is the set of nodes that shield it from the rest of the network. Previous studies have shown that knowing the Markov Blanket of a diagnosis node is all that is required in order to predict the value of the criterion, either by classification or regression[6,7,8]. If the MB of a specific diagnosis can be identified prior to a policy change, it may be used to more accurately identify the set of criterion diagnoses following the policy change, which can in turn be used to estimate true cases, as well as the extent of gaming of diagnoses which will occur due to the policy change.

Our experiments used discharge data from the California and New York State Inpatient Databases (SID), obtained from the Healthcare Cost and Utilization Project (HCUP) provided by the Agency for Healthcare Research and Quality[26]. The SID is a component of the HCUP, a partnership between federal and state governments and industry, which tracks all hospital admissions at the individual level. We included all data from January 2003 through December 2011. Patients were excluded from the analysis if they did not have Medicare or Medicaid as the primary payer and if they were younger than 19 years of age. The final dataset included 16,736,927 discharge records for CA and 12,717,787 discharge records for NY, with the primary set of features used in our experiences being the Clinical Classifications Software (CCS) diagnoses for ICD9-CM. CCS codes, developed as part of the HCUP, are designed to cluster patient diagnoses (hereinafter DX) and procedures (hereinafter PX) into a manageable number of clinically meaningful categories (272 diagnoses and 231 procedure codes).

Some prior research has examined the role of comorbid conditions with the aim of identifying longer-term effects and mortality risk with a single target diagnosis in mind. Each of our target diagnoses has been considered in this fashion: AMI[9], CHF[11], PN[14], and sepsis[15]. To the best of our knowledge, ours is the first study concerned with identifying co-occurring diagnoses and procedures that can serve as a proxy indicator of the target diagnosis, something necessary to identify potential instances of hospitals gaming the system to reduce risk exposure.

Methods

Our goal was to find a minimum subset of the most informative DX and PX (accurately predict the presence of Target DX in the records) associated with the Target DX. We decided to build a directed weighted ego-centric network for each Target DX where weights are calculated from counts of co-occurrences of DX and PX. We calculated the PageRank for each DX, PX to obtain ranked list. We used the PageRank value as a criterion to rank the importance of DX, PX in the above defined Target DX ego-centric network. Number of features for PageRank approximation of MB identified as the maximum number of DX, PX in the records. This way obtained PageRank approximation of MB then serves as feature set for logistic regression with a default setup[22,23,25]. Generated software will be available at https://github.com/dusanramljak

Target Diagnosis Ego-centric Network of Diagnoses and Procedures

We started with identifying DXs and PRs that most frequently co-occur with the Target DXs. The most frequent co-occurrence by itself is necessary, but not sufficient, to establish an appropriate approximation of the Markov Blanket. For example, some diagnoses might co-occur with a Target DX simply because they are frequently diagnosed. Meanwhile, other frequent diagnoses might co-occur with our Target DX and also contribute significantly in discriminating between the classes when viewed in combination with other diagnoses or procedures. Making that distinction is not possible by looking only at the frequencies of co-occurrences with a Target DX.

To that end we have built directed weighted networks of DX and PR from the hospital records that contained each of our Target DX. The starting weights in this network were counts of co-occurrences of the nodes. In order to build directed weights we followed the following intuition: the “level of trust” of a node (source) in a connection with another node (destination) should be scaled by the counts of occurrences of the source in the network – the source can “trust” the destination only to an extent that is proportional to the ratio between the count of the occurrence of the link between them and its own count of occurrence. Since we also have the counts of occurrences of the nodes in the records that don’t contain our Target DX, we adjusted the weights in the network according to occurrences of destination in the records that don't contain our Target DX. Following this intuition, the “level of trust” should be scaled with the count of occurrences of the destination in the records that don’t contain our Target DX. The formula for setting the weights is $w = \frac{c_l}{c_s * c_d}$ where c_l is the count of co-occurrences for the source and destination nodes, c_s is the count of occurrences of the source node in the network, and c_d is the count of occurrence of the destination node in the records that do not contain Target DX.

PageRank Approximation of Markov Blanket

Since we defined weights by co-occurrences, as well as by additional information from the structure of the network of DXs and PXs that co-occur with our Target DX, we could use the PageRank value as a criterion to identify important nodes. For a subset of highly important nodes, we could say that the nodes with highest PageRank represent an approximate Markov Blanket for our Target DXs.

PageRank gives us a ranked list of important DX and PX, but we are not able to determine if there are any redundancies. Redundancies in this context mean that several DX and PX might not provide information to discriminate between classes. In our earlier experiments we used a feature selection method to help us decide the number of DX and PX that will be provided by our PageRank approximation of MB, but our more recent experiments follow a different path. Because our goal is to have all the important DX and PX that could represent our Target DX in our MB, we determined the maximum number of DX and PX that could be present in individual records. That number is then used to determine the number of nodes with the highest PageRank included in the MB.

Results

The data we used in our experiments comes from the HCUP family of databases, and the raw data consists of patient hospital visit records from California’s and New York’s SID in the period from January 2003 up to December 2011. Each record consists of a number of attributes, which are explained in detail on the HCUP website[26] . The California and New York database contain more than 50 million inpatient discharge records over the specified 9 years. The information is not specific to a group of hospitals, but rather represents the data for the entire state.

The database also includes demographic information for each patient (such as age, birth year, sex, race), DX (primary and up to 24 secondary for CA, primary and up to 14 secondary for NY), PX (up to 21 for CA, and up to 15 for NY), information about hospital stays, and other information (including length of stay, total charges, type of payment and payer, discharge month, and survival information). For the purposes of our experiments, we used only the procedure and diagnosis information since our earlier work showed its potential, and since one of our primary concerns is preserving privacy (which could be breached if we used the more specific demographic and hospital stay information). Each of the 4 Target DXs we examined in this study was fairly prevalent. As shown in Table 1, CHF was most common (17.37% CA, 16.51% NY), followed by PN (9.43% CA, 7.77% NY), Sepsis (6.13% CA, 4.91% NY), and AMI (3.25% CA, 3.14% NY). There were also considerable seasonal and secular trends (similar in both datasets) in these Target DXs. For example, the prevalence of sepsis increased steadily across the study period. The three other Target DXs showed gradual decreases in prevalence over time, but also very strong seasonal trends.

Table 1. Target DX frequency in CA and NY state databases

	Target DX			
	Sepsis	AMI	CHF	PN
CA	1,027,088	544,228	2,907,625	1,577,822
NY	625,310	399,371	2,100,602	988,475

We chose to use 20 combined DX and PX in this study, since this covered 95% of California and 98% of New York state data. We divided the data into two parts in order to show that the model generalizes well over both space and time. We used the first part, containing data from California in the years 2003 – 2008, for training and validating our models, and report the results of using 5-fold cross-validation. The second part consists of 2009 – 2011 records for both NY and CA data, and is used exclusively for testing. Since a previous study[27] suggests that accuracy is maximized when both training and test sets are balanced, used the same number of positive (Target DX present in the record) and negative (Target DX not present in the record) examples during training and testing. Since the four Target DXs appear at different rates (as shown in Table 1), the sample sizes we chose were different for each Target DX, but in each case we chose sizes with a roughly logarithmic progression such that the largest sample size covered the majority of the available cases. We used three sample sizes of different scale for each Target DX, and refer to these sample sizes using the general terms small, moderate, and large (for example, for CHF (2003-2008) in California, these sample sizes were 100k, 300k, and 1million, respectively). We gathered 10 random samples for each sample size, and used 5-fold cross-validation for training and validation, resulting in 180 distinct training datasets per Target DX. When testing, we used the maximum possible size of the positive class for each Target DX, and performed 10 replications for each of the two states we have available, resulting in 20 distinct testing datasets for each Target DX. We opted to use 10 random samples that, when taken together, covered the entire dataset, rather than using the entire dataset as one big sample to increase the number of replications. To this end, we sampled cases with replacement with no overlaps within samples, and minimal to nonexistent overlap between samples.

For each of our Target DX, 150 out of 180 models were included in 5-fold cross-validation and tested once on an appropriate 5th fold holdout. 30 models were tested on all of the 20 test sets, and the results indicate high out–of–sample accuracy, scalability, and stability of the method. The fact that adding more features always increases the accuracy supports the idea that we are dealing with a challenging problem, though the diminishing returns on accuracy improvements suggest that using a smaller set of features can be sufficient. Furthermore, although diagnoses are the most commonly used features across different hospital readmission models[4], many other individual characteristics are also important but either not considered here (e.g., age, sex, race) or not available in the data we used (medications, laboratory tests ordered, laboratory test results). We include results using all diagnoses and procedures as features, rather than using only the significantly smaller subset of diagnoses and procedures selected by our method, to show that we attained comparable accuracy using only the subset of features that are strongly relevant to the Target DXs. Furthermore, experiments showed the somewhat surprising fact that having more data for training doesn't offer much improvement in accuracy, but does affect the stability of both the achieved accuracy and the chosen features.

We evaluate the "stability" of the features selected by our method by examining how often each one appears in each of our training models. That is, after acquiring the ranked list of PageRank values for each of the features during each experiment, we selected the 20 highest-ranked features in each experiment (giving us 60 potentially overlapping sets of 20 features for each sample size), and then took the intersection of those feature sets to find features that appear in all experiments. We labeled these features as stable, since they were shown to be in the top 20 PageRank lists over the course of all experiments.

This procedure was repeated for each of the three sample sizes for each of the four target diagnoses, yielding the results summarized in Tables 2 and 3. These tables portray the features that were selected as the most influential by PageRank over each set of experiments, for each of the Target DX, as well as the sample size for which each feature was stable. In other words, features that are stable in all three sample sizes are the most frequently occurring and therefore are the best candidates for the MB approximation, while those that only appear during the experiments in the largest sample size are less frequent.

While accuracy is an important criterion for our task, interpretability and relevance of the selected features also play key roles in the utility and acceptability of the approach, and our results are very encouraging in this regard. For AMI (Table 2), the two selected diagnoses, cardiac arrest and shock, are two of the most common consequences of a heart attack. Similarly, the selected procedures all align directly with common clinical practice, representing a variety of cardiac imaging, diagnostic, and surgical procedures. CHF (Table 2) is a complex condition reflecting the intersection of cardiac and pulmonary systems. These are well represented among the selected diagnoses and procedures, as is the overlap (in terms of cardiovascular disease) with AMI. For PN (Table 3), one recent study[19] found that readmissions rates for PN could be made more accurately and stably when sepsis and respiratory failure were also included. Both of these diagnoses are selected for PN, along with a range of respiratory procedures. Finally, the expected overlap between Sepsis (Table 3) and PN emerges for diagnoses and procedures when the former is considered as the target diagnosis. Additionally, there are more serious diagnoses relating to injuries and abscesses along with various stoma.

In addition to validating the relevance of the small subset of features we selected for each diagnosis, we also tested their representative power by comparing the predictive accuracy (measured in AUC) obtained using only this small subset of features against the accuracy obtained when using all (approximately 500) features for the same task.

These results are presented in Figure 3 as boxplots for each Target DX at each sample size, with variance obtained from the 10 repetitions of each experiment. Several observations can be made from these plots, the first being that while using all 500 features offers the best accuracy, we can attain very competitive predictive accuracy using a significantly smaller number (20) of features which have the added benefit of being relevant to the Target DX in addition to being useful for discriminating between the two classes. The gap between accuracies using the subset vs using all features varies among the four Target DX, but even in the case of the most difficult to predict diagnosis, PN, the AUC we attained using only 20 features was very respectable (in the 0.68-0.75 range). In the case of the easiest to predict diagnosis, Sepsis, we were able to attain a nearly perfect level of prediction using a very small number of features, which we believe to be particularly important given the frequency and deadliness of this diagnosis. A somewhat more surprising finding was that varying the sample size did not have as big of an impact as expected. Although a general trend of AUC being higher when the sample sizes are larger can be seen, there is a noticeable variability in the results, while the mean values of AUC are quite close between sample sizes. While we expected sample size to play a bigger role in increasing accuracy, these findings suggest that the features we chose were representationally powerful enough to offer good performance even when the sample size is limited, further proving their relevance to the target DXs.

Table 2. Table showing the diagnoses and procedures that were used in all experiments for appropriate sample sizes (marked by x in appropriate raw) to form the PageRank approximated MB for AMI and CHF

	Size		
Acute Myocardial Infarction	I	II	III
Diagnoses			
Cardiac arrest and ventricular fibrillation	x	x	x
Shock	x	x	x
Procedures			
Contrast aortogram	x	x	x
Contrast arteriogram of femoral and lower extremity arteries			x
Conversion of cardiac rhythm	x	x	x
Coronary artery bypass graft (CABG)	x	x	x
Coronary thrombolysis	x	x	x
Diagnostic cardiac catheterization; coronary arteriography	x	x	x
Diagnostic ultrasound of heart (echocardiogram)			x
Extracorporeal circulation auxiliary to open heart procedures	x	x	x
Nuclear medicine imaging of pulmonary			x
Other non-OR therapeutic cardiovascular procedures	x	x	x
Other OR heart procedures	x	x	x
Other OR procedures on vessels other than head and neck	x	x	x
Other therapeutic procedures		x	x
Percutaneous transluminal coronary angioplasty (PTCA)	x	x	x
Respiratory intubation and mechanical ventilation	x	x	x
Swan-Ganz catheterization for monitoring	x	x	x
Tracheostomy; temporary and permanent	x	x	x

	Size		
Congestive Heart Failure	I	II	III
Diagnoses			
Acute and unspecified renal failure		x	x
Acute myocardial infarction	x	x	x
Cardiac arrest and ventricular fibrillation	x	x	x
Chronic kidney disease			x
Conduction disorders	x	x	x
Heart valve disorders	x	x	x
Hypertension with complications and secondary hypertension		x	x
Peri-; endo-; and myocarditis; cardiomyopathy	x	x	x
Pulmonary heart disease	x	x	x
Respiratory failure; insufficiency; arrest (adult)	x	x	x
Shock	x	x	x
Procedures			
Conversion of cardiac rhythm	x	x	
Diagnostic ultrasound of heart (echocardiogram)		x	x
Extracorporeal circulation auxiliary to open heart procedures	x		
Heart valve procedures	x	x	x
Insertion; revision; replacement; removal of cardiac pacemaker or cardio			x
Nuclear medicine imaging of pulmonary			x
Other OR heart procedures	x	x	x
Respiratory intubation and mechanical ventilation	x	x	x
Swan-Ganz catheterization for monitoring	x	x	x
Tracheostomy; temporary and permanent	x		

Table 3. . Table showing the diagnoses and procedures that were used in all experiments for appropriate sample sizes (marked by x in appropriate raw) to form the PageRank approximated MB for AMI and CHF

	Size		
Pneumonia	I	II	III
Diagnoses			
Cystic fibrosis			x
Other injuries and conditions due to external causes			x
Pleurisy; pneumothorax; pulmonary collapse		x	x
Respiratory failure; insufficiency; arrest	x	x	x
Septicemia (except in labor)		x	x
Shock	x	x	x
Procedures			
Arterial blood gases			x
CT scan chest			x
Diagnostic bronchoscopy and biopsy of bronchus	x	x	x
Enteral and parenteral nutrition		x	x
Gastrostomy; temporary and permanent	x	x	x
Incision of pleura; thoracentesis; chest drainage	x	x	x
Other diagnostic procedures of respiratory tract and mediastinum	x		x
Other non-OR therapeutic procedures on respiratory system	x	x	x
Other respiratory therapy		x	x
Respiratory intubation and mechanical ventilation	x	x	x
Swan-Ganz catheterization for monitoring	x	x	x
Tracheostomy; temporary and permanent	x	x	x

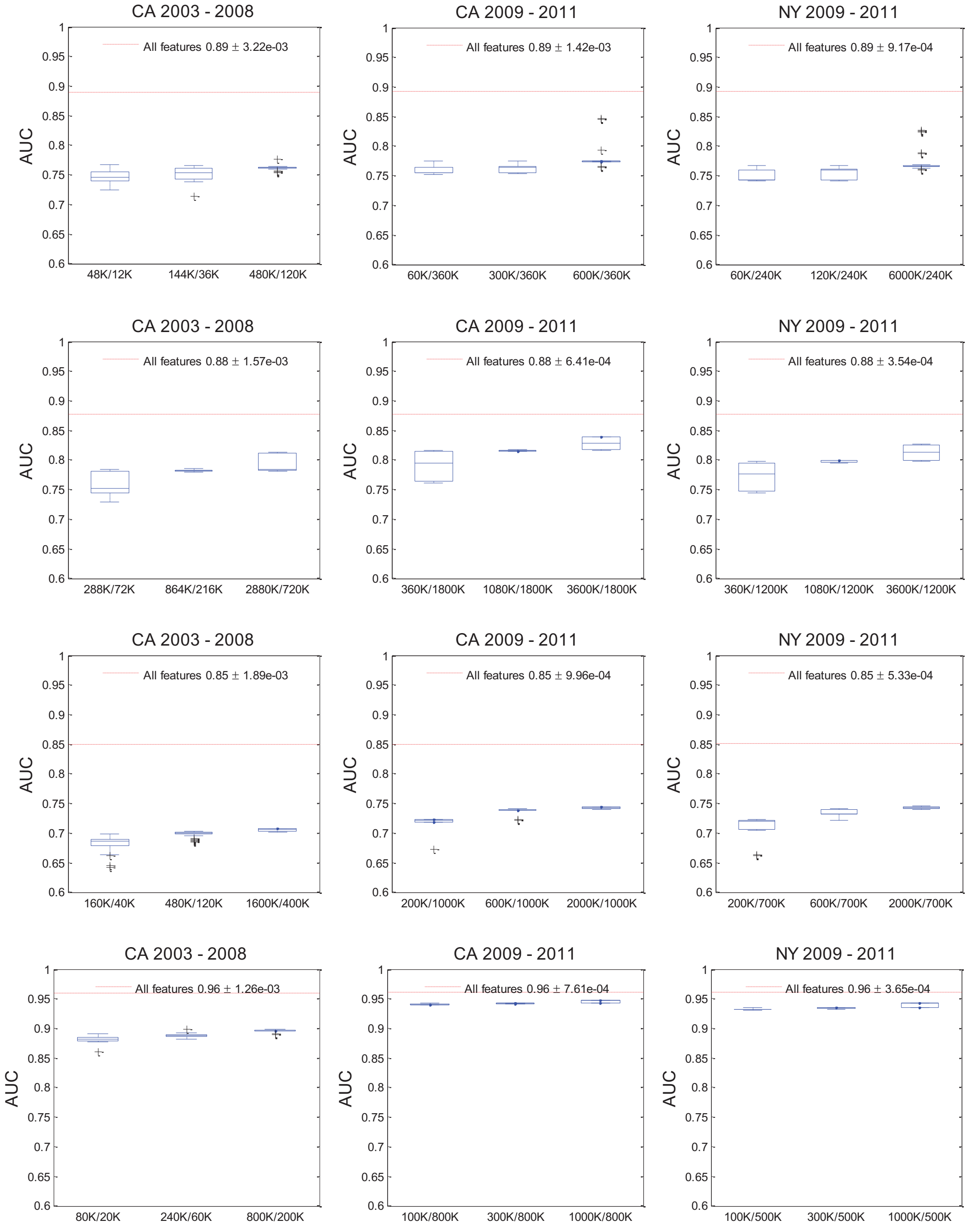
	Size		
Septicemia (except in labor)	I	II	III
Diagnoses			
Aspiration pneumonitis; food/vomitus			x
Cardiac arrest and ventricular fibrillation			x
Other injuries and conditions due to external causes	x	x	x
Peritonitis and intestinal abscess	x	x	x
Respiratory failure; insufficiency; arrest			x
Shock	x	x	x
Procedures			
Bone marrow transplant			x
Colostomy; temporary and permanent	x	x	x
Conversion of cardiac rhythm			x
Enteral and parenteral nutrition	x	x	x
Gastrostomy; temporary and permanent	x	x	x
Ileostomy and other enterostomy	x	x	x
Myringotomy			x
Other non-OR therapeutic procedures on respiratory system	x	x	x
Other non-OR therapeutic procedures on skin and breast			x
Other vascular catheterization; not heart	x	x	x
Respiratory intubation and mechanical ventilation	x	x	x
Swan-Ganz catheterization for monitoring	x	x	x
Tracheostomy; temporary and permanent	x	x	x

Conclusion

The US healthcare system is rife with opportunities for perverse incentives. The implementation of any new healthcare policy results in changes within the healthcare system in order to minimize the adverse consequences of the policy change for healthcare providers. Changes that began in 2012 under the Affordable Care Act can be expected to reduce the number of individuals receiving target diagnoses of AMI, CHF, and PN as healthcare providers move to reduce their exposure to adverse consequences of hospital readmissions. Beginning in October 2014, the HRRP was extended to three additional conditions: acute exacerbation of chronic obstructive pulmonary disease (COPD), patients admitted for elective total hip arthroplasty (THA), and total knee arthroplasty (TKA). Early models of readmissions are being presented for COPD[30], but they suffer from the same low predictive value (AUC=0.65) as earlier work with AMI, CHF, and PN, and the value of COPD as a criterion measure has received some criticism.

In this paper, we propose a novel approach to the problem of under-diagnosing, specifically, approximating Markov Blankets by PageRank. Performance using this subset of diagnoses shows performance that is generally quite high in terms of both accuracy and precision. Additionally, these diagnoses and procedures often point to clinically meaningful patterns. However, it is unclear which will ultimately prove most useful as the network of diagnoses and procedures surrounding a Target DX changes in response to policy. To some extent, this problem is likely to pose a continuously moving target and so future research should more fully develop an understanding of the temporal forces to determine whether, for example, the indicators of PN depend on month of admission.

Figure 1. AUC values of the model using the top 20 features selected by PageRank (blue boxes) and using all features (dotted red line). Each row represents one of the Target DX: AMI, CHF, PN and Sepsis. Each column represents a different subset of data that was used for testing purposes. Results obtained from the three sample sizes described above are shown for each setting.



The approach used here is likely to be useful in the analysis of healthcare data in several ways. First, it provides a set of associated diagnoses and procedures that can be used to "impute" missing or unobserved data in an effort to estimate true prevalence of various diseases. Second, it can be used to estimate the extent of "gaming" of diagnoses in response to policy changes. Several authors consider the potential for gaming[31]. Multiple potential methods of "gaming" have been suggested, including enriching the population admitted with a target diagnosis with individuals assessed to have low readmission risk, using extended “holding areas” for patients to receive hospital care without being readmitted, and selectively coding target diagnoses among patients with expected low readmission risk. Such gaming strikes us as particularly likely given the narrow range of criteria that factor into reimbursement rates under the HRRP. Enriching the pool of related diagnoses has been shown to provide more accurate and stable estimates in the case of PN[19].

Extending Rothberg et al. (2014)[19], Sjoding et al. (2015)[20] performed a Monte Carlo study using 2009 Medicare data. They found that hospitals could substantially improve their pneumonia readmission and mortality rates by converting pneumonia diagnoses to sepsis or respiratory failure. The improvements are often substantial. From a sample of 100 hospitals with pneumonia readmission rates above the 50th percentile, 66 improved their readmissions rate, and 15 dropped below the 50th percentile. Changes were even more dramatic when mortality was considered (90 and 41 hospitals, respectively). This suggests that our approach may also prove useful in order to adjust estimates for this kind of gaming and could also provide more robust methods to estimate true hospital readmission rates where intentional under-diagnosis of such diagnoses is likely. Historically, there are several precedents for this kind of under-reporting. The effects of the Omnibus Budget Reconciliation Act of 1987 (OBRA87) was observed to have considerable impact on medical practice in nursing home settings[21]. Considerable decreases in the per capita diagnosis of AMI can already be seen leading up to implementation of the HRRP (Gerhardt, et al., 2012).

Going forward, several areas of inquiry are likely to be fruitful. For example, future work should consider the conditions newly added in October 2014. Related to this, now that a reasonable amount of data are available following the implementation of the HRRP, researchers should evaluate evidence for shifting of diagnoses away from the targets. Another related area of critical importance has to do with the way in which individual hospital reimbursement rates are set and risk-adjusted. There is some evidence that building a larger network of related diagnoses may produce more stable and accurate performance estimates. Thus, a direct extension of this work is to consider the performance of a model trained on historical data, but used to predict performance of individual hospitals. Given how well our models generalize across time and state, we have reason to expect strong performance. Finally, models such as the ones we present here may have direct application to fraud detection and upcoding (Suresh et al., 2014)[28].

Acknowledgements

This research was supported by DARPA Grant FA9550-12-1-0406 negotiated by AFOSR and NSF through grant number NSF-14476570. HCUP, Agency for Healthcare Research and Quality, provided data used in this study.

References

[1] Keenan PS, Normand SLT, Lin Z, Drye EE, Bhat KR, Ross JS, et al. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. Circulation: Cardiovascular Quality and Outcomes. 2008;1(1):29–37.
[2] Krumholz HM, Lin Z, Drye EE, Desai MM, Han LF, Rapp MT, et al. An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among patients with acute myocardial infarction. Circulation: Cardiovascular Quality and Outcomes. 2011;4(2):243–252.
[3] Stiglic G, Davey A, Obradovic Z. Temporal Evaluation of Risk Factors for Acute Myocardial Infarction Readmissions. In: Healthcare Informatics (ICHI), IEEE International Conference on. IEEE; 2013. p. 557–562.
[4] Kansagara D, Englander H, Salanitro A, et al. Risk prediction models for hospital readmission: a systematic review. Jama. 2011;306(15):1688–1698.
[5] Rice TH. The impact of changing Medicare reimbursement rates on physician-induced demand. Medical care. 1983;21(8):803–815.
[6] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann; 2014.
[7] Lou Q, Obradovic Z. Feature Selection by Approximating the Markov Blanket in a Kernel-Induced Space. In: ECAI; 2010. p. 797–802.
[8] Lou Q, Obradovic Z. Predicting viral infection by selecting informative biomarkers from temporal high-dimensional gene expression data. In: Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on. IEEE; 2012. p. 1–4.
[9] Spencer FA, Moscucci M, Granger CB, Gore JM, Goldberg RJ, Steg PG, et al. Does comorbidity account for the excess mortality in patients with major bleeding in acute myocardial infarction? Circulation. 2007;116(24):2793–2801.

[10] Gerber Y, Rosen LJ, Goldbourt U, Benyamini Y, Drory Y. Smoking status and long-term survival after first acute myocardial infarction: A population-based cohort study. Journal of the American College of Cardiology. 2009;54(25):2382–2387.

[11] Braunstein JB, Anderson GF, Gerstenblith G, Weller W, Niefeld M, Herbert R, et al. Noncardiac comorbidity increases preventable hospitalizations and mortality among Medicare beneficiaries with chronic heart failure. Journal of the American College of Cardiology. 2003;42(7):1226–1233.

[12] Estes JM, Guadagnoli E, Wolf R, LoGerfo FW, Whittemore AD. The impact of cardiac comorbidity after carotid endarterectomy. Journal of vascular surgery. 1998;28(4):577–584.

[13] Vergis EN, Brennen C, Wagener M, Muder RR. Pneumonia in long-term care: a prospective case-control study of risk factors and impact on survival. Archives of internal medicine. 2001;161(19):2378–2381.

[14] Yende S, Angus DC, Ali IS, Somes G, Newman AB, Bauer D, et al. Influence of Comorbid Conditions on Long-Term Mortality After Pneumonia in Older People. Journal of the American Geriatrics Society. 2007;55(4):518–525.

[15] Weycker D, Akhras KS, Edelsberg J, Angus DC, Oster G. Long-term mortality and medical care charges in patients with severe sepsis. Critical care medicine. 2003;31(9):2316–2323.

[16] Janda S, Young A, FitzGerald JM, Etminan M, Swiston J. The effect of statins on mortality from severe infections and sepsis: a systematic review and meta-analysis. Journal of critical care. 2010;25(4):656–e7.

[17] Lou Q, Parkman HP, Jacobs MR, Krynetskiy E, Obradovic Z. Exploring genetic variability in drug therapy by selecting a minimum subset of the most informative single nucleotide polymorphisms through approximation of a markov blanket in a kernel-induced space. In: Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on. IEEE; 2012. p. 156–163.

[18] Yu L, Liu H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: ICML. vol. 3; 2003. p. 856–863.

[19] Rothberg MB, Pekow PS, Priya A, Lindenauer PK. Variation in diagnostic coding of patients with pneumonia and its association with hospital risk-standardized mortality rates: a cross-sectional analysis. Ann Intern Med. 2014;160(6):380–388.

[20] Sjoding MW, Iwashyna TJ, Dimick JB, Cooke CR. Gaming Hospital-Level Pneumonia 30-Day Mortality and Readmission Measures by Legitimate Changes to Diagnostic Coding. Crit Care Med. 2015.

[21] Elon R, Pawlson LG. The impact of OBRA on medical practice within nursing facilities. Journal of the American Geriatrics Society. 1992;40(9):958–963.

[22] Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A library for large linear classification. The Journal of Machine Learning Research. 2008;9:1871–1874.

[23] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.

[24] Ramljak D, Davey A, Uversky A, Roychoudhury S, Obradovic Z. Hospital Corners and Wrapping Patients in Markov Blankets. In: Proceedings of the 4th Workshop on Data Mining for Medicine and Healthcare, SDM, Vancouver, Canada, April 30 - May 02, 2015; 2015. .

[25] Yu HF, Huang FL, Lin CJ. Dual coordinate descent methods for logistic regression and maximum entropy models. In: Machine Learning. 2011;85(1-2):41–75.

[26] HCUP State Inpatient Databases (SID), Healthcare Cost and Utilization Project (HCUP). 2009-2011. Agency for Healthcare Research and Quality, Rockville, MD. http://www.hcup-us.ahrq.gov/sidoverview.jsp; http://www.hcup-us.ahrq.gov/db/state/siddist/sid_multivar.jsp

[27] Wei, Qiong ,Dunbrack, Jr, Roland L. The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS ONE 2013; 8(7): e67863.

[28] Suresh NC, de Traversay J, Gollamudi H, Pathria AK, Tyler MK. Detection of upcoding and code gaming fraud and abuse in prospective payment healthcare systems. 2014.https://www.google.com/patents/US8666757.

[29] Feemster LC, Au DH. Penalizing hospitals for chronic obstructive pulmonary disease readmissions. Am J Respir Crit Care Med. 2014;189(6):634–639.

[30] Shahian DM, He X, O'Brien S, et al. Development of a clinical registry-based 30-day readmission measure for coronary artery bypass grafting surgery. Circulation. 2014:CIRCULATIONAHA–113.

[31] Burgess JF, Hockenberry JM. Can all cause readmission policy improve quality or lower expenditures? A historical perspective on current initiatives. Heal Econ Policy Law. 2014;9(02):193–213.

[32] Joynt KE, Jha AK. Characteristics of hospitals receiving penalties under the Hospital Readmissions Reduction Program. Jama. 2013;309(4):342–343.