Knowledge Discovery from Massive Healthcare Claims Data

Varun Chandola
Oak Ridge National Laboratory
chandolav@ornl.gov

Sreenivas R. Sukumar
Oak Ridge National Laboratory
sukumarsr@ornl.gov

Jack Schryver
Oak Ridge National Laboratory
schryverjc@ornl.gov

ABSTRACT

The role of big data in addressing the needs of the present healthcare system in US and rest of the world has been echoed by government, private, and academic sectors. There has been a growing emphasis to explore the promise of big data analytics in tapping the potential of the massive healthcare data emanating from private and government health insurance providers. While the domain implications of such collaboration are well known, this type of data has been explored to a limited extent in the data mining community. The objective of this paper is two fold: *first*, we introduce the emerging domain of "big" healthcare claims data to the KDD community, and *second*, we describe the success and challenges that we encountered in analyzing this data using state of art analytics for massive data. Specifically, we translate the problem of analyzing healthcare data into some of the most well-known analysis problems in the data mining community, *social network analysis*, *text mining*, and *temporal analysis and higher order feature construction*, and describe how advances within each of these areas can be leveraged to understand the domain of healthcare. Each case study illustrates a unique intersection of data mining and healthcare with a common objective of improving the cost-care ratio by mining for opportunities to improve healthcare operations and reducing what seems to fall under fraud, waste, and abuse.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

Keywords

Healthcare Analytics; Fraud Detection

1. INTRODUCTION

Healthcare spending in United States is one of the key issues targeted by policy makers, owing to the fact that it is a major contributor to the high national debt levels that are projected for next two decades. In 2008, the total healthcare spending in US was 15.2% of its GDP (highest in the world) and is expected to reach as much as 19.5% by 2017 [2]. But while the healthcare costs have risen (by as much as 131% in the past decade), the quality of healthcare in the US has not seen comparable improvements (See Figure 1) [24].

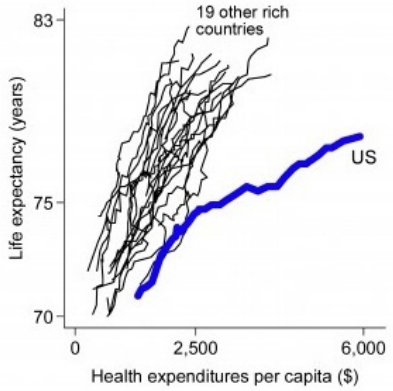


Figure 1: Life expectancy compared to healthcare spending from 1970 to 2008, in the US and the next 19 most wealthy countries by total GDP [14].

Experts agree that inefficiencies in the current healthcare system, resulting in unprecedented amounts of waste, is the primary driver for the discrepancy between the spending and the returns in the healthcare domain [11]. Recent studies estimate that close to 30% ($\sim$ $765 billion in 2009) of total healthcare spending in United States is wasted, which in turn is caused by many factors such as unnecessary services, fraud, excessive administrative costs, and inefficiencies in the healthcare delivery.

In recent years, several experts as well as the federal government[1] have stressed on the role of big data analytics in addressing the issues with healthcare. The 2011 report by *Mckinsey Global Institute* [19] estimate that the potential value that can be extracted from data in the healthcare sector in US could be more than $300 billion per year. The same report lists out several areas within the healthcare sector which can benefit from using big data analytics. These include segmentation of patients based on their health profiles to identify target groups for proactive care or lifestyle

[1] `http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf`

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

changes, development of fraud resistant payment models, creating information transparency and accessibility around healthcare data, and conducting comparative effectiveness research across providers, patients, and geographies.

Healthcare insurance claims have the potential of answering many of the questions currently faced by the healthcare sector, In fact, until shareable *electronic health records* become a reality, healthcare claims, especially from organizations with a large spatial and demographic coverage such, which is the case with many of the government run health insurance programs in the country, are the most reliable resource for understanding the current healthcare landscape, from conditions, care, and cost perspective. But the transactional format of claims data is not amenable for advance analytics that the state of art KDD methodologies have to offer. In this paper we explore transformations of the healthcare claims data which bridge this gap between the healthcare domain and modern data analytics.

We present a study of big data analytics on health insurance data collected by a large national social health insurance program. While previous efforts have used data mining methods for analyzing healthcare claims data within small organizations, we believe that this is one of the first instances where advanced data analytics and healthcare have interacted at a national scale. In all, we approximately analyzed 2 billion insurance claims for approximately 45 million beneficiaries and over 3 million healthcare providers.

1.1 Our Contributions

In this paper, we make the following contributions:

1. We introduce the emerging domain of health care claims data and identify multiple research problems that can be solved using the existing big data analytics solutions.

2. We propose three transformations of the transactional claims data which enables application of state of art KDD methodologies in this domain.

3. We present several approaches to identify and understand the fraud, waste, and abuse in the health care system. The potential of each proposed approach is demonstrated on real claims data and validated using a true set of fraudulent providers.

4. We highlight the unique nature of healthcare data when analyzed using methods such as social network analysis and text analysis.

2. RELATED WORK

The role of big data in healthcare has been well acknowledged across government and industrial sectors[2][3]. But only a few published studies have analyzed such data [12]. The primary reason is the data availability, given that the healthcare claims data has strong proprietary and privacy requirements[4]. Moreover, existing studies have considered claims data from a payment system perspective and have analyzed the data for payment errors [12]. Limited efforts exist that have analyzed the healthcare claims data to understand the inefficiencies in the healthcare system [6], but from the analytics perspective they are limited to simple summary statistics such as population means for various demographics. In this paper we explore the application of three advanced KDD technologies, viz. , text mining [18], social network analysis [29], and time series analysis [20], all of which have been successful in a variety of applications but have not been applied at a large scale to healthcare claims data.

Domains such as credit card and property insurance have long studied the issue of fraud identification [4, 10]. But healthcare fraud detection has unique characteristics given that the actual beneficiary is typically not the fraud perpetrator, which is not the case for other domains. So existing fraud detection methods cannot be directly applied to the healthcare domain. Most of the existing fraud detection solutions in the healthcare domain are not public, primarily because of the fact that the data is highly sensitive and is usually not made available for research and publishing.

Figure 2: Different Types of Healthcare Data

3. BACKGROUND

Healthcare data can be broadly categorized into four groups (See Figure 2): **Clinical data** (patient health records, medical images, lab and surgery reports, etc.) and **patient behavior data** (collected through monitors and wearable devices) provide an accurate and detailed view of the health of the population. But such data, which is increasingly being stored electronically, can be leveraged in a big data setting only when the owners (doctors, hospitals, and individuals) share, which, owing to privacy concerns is still limited to being analyzed within an organization such as a hospital or a network of hospitals. **Pharmaceutical research data** (clinical trial reports, high throughput screening results) often face privacy concerns owing to business practices. In this paper, we focus on **health insurance data**, which has been collected and stored for several years by various health insurance agencies. While the primary justification for this data is to track payments and address fraud, such data also has great potential to address some of the other aforementioned issues of the healthcare system. For United States, such data is extremely valuable, given that around 85% of Americans use some form of insurance (private or government). Moreover, insurance data is the only source of the cost associated with healthcare which is vital to address the

[2] www.eweek.com/database/emc-says-big-data-is-essential-to-improving-health-outcomes/
[3] www.intel.com/content/dam/www/public/us/en/documents/white-papers/healthcare-leveraging-big-data-paper.pdf
[4] http://www.hhs.gov/ocr/privacy/

economic challenges associated with modern healthcare system. The strong challenge presented by insurance data on the other hand is that it is not readily in the form to infer strong analytic insights into healthcare, besides the payment model. A key contribution of this paper is the transformation of the insurance data into formats that allow application of existing analytic tools for knowledge discovery.

3.1 Health Insurance Data

The typical health insurance payment model is a *Fee-for-service* (FFS) model in which the providers (doctors, hospitals, etc.) render services to the patients and are paid for each service by the payor or the insurance agency. The providers record the details of each service, including the cost and justification and submit the record to the payor. The payor decides to either pay or reject the claim based on the patient's eligibility for the particular service which are determined by the policy guidelines.

The insurance agency typically maintains three types of data for their operations:

1. *Claim information* captures the information about the service transaction including the nature of the service and the cost.
2. *Patient enrollment and eligibility data* that captures demographic information about the patients (or beneficiaries of the system) and their eligibility for different services.
3. *Provider enrollment data* that captures the information about the physicians, hospitals, and other healthcare providing organizations.

3.2 Health Insurance in United States

In the US, approximately 85% of the population has some of form of health insurance. Majority of these individuals ($\approx$ 60%) obtain insurance through their employer or employer of parent or spouse. Almost 28% of population (83 million individuals) is covered under government health insurance programs. These include programs such as *Medicare*, *Medicaid*, *Veterans Health Services*, etc.

The data managed by each of these programs is at a massive scale. Medicare alone provides health insurance to 48 million Americans and covers for hospitalization, out patient, medical equipments, and drugs. There are a few million providers enrolled with the Medicare *Provider Enrollment, Chain, and Ownership System* (PECOS). In 2011, Medicare received close to 1.2 billion claims (4.8 million claims per day) for their fee for service programs. An almost equivalent number of claims were received in the prescription drugs program (also known as *Part D*). Under Medicaid, more than 60 million individuals received benefits in 2009 (one in every five). In the state of Texas alone, there are more than 0.5 million providers within Medicaid.

4. CHALLENGES AND OPPORTUNITIES

As mentioned in Section 1, *fraud*, *waste*, and *abuse* form a significant amount of healthcare spending. Fraud ranges from single providers billing the system for services that were not provided to large scale fraud carried out by organized criminals [27]. One form of waste happens due to *improper payments*, since organizations are mandated to process payments in a short duration of time, resulting in authorization

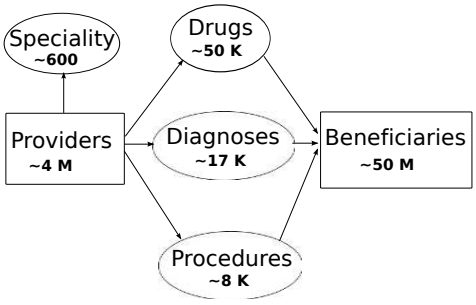


Figure 3: Entities and relationships in healthcare claims data along with approximate number of entries in each entity set.

of double payments for duplicate claims, payments using an outdated fee schedule, etc. A major cause for abuse is due to the fact that programs such as Medicare follow a *prospective payment system* for hospital care, which means that providers are paid for services at predetermined rates. Thus if the actual service costs more than the allowed cost, the provider has to cover its losses. If the actual service costs less than the allowed cost, the provider keeps the remainder. This drives the providers to charge unnecessary or more expensive services (also known as *upcoding*) by making more severe diagnosis to safeguard against any losses and to make profit. Some of the examples listed above, e.g., double payment for duplicate claims, can be identified by applying business rules to the data. For others, such as upcoding or miscoding providers, there is need for advanced algorithms that can analyze the vast amounts of claims data.

In the subsequent sections we describe three case studies that were conducted onhealthcare claims and associated data. The common thread among the studies is the analysis of the behavior and interaction between healthcare providers, which is highly important given that they are the primary drivers for the wasteful spending in the system.

5. DATA

The data used for the subsequent case studies captures three different aspects of healthcare. First is the **claims data** for close to 48 million beneficiaries for the entire US. Second is the **provider enrollment data** which can be obtained from several private organizations. The third is a list of **fraudulent providers** that have been sanctioned for fraudulent behavior in the state of Texas. The list of fraudulent providers was obtained from the Office of Inspector General's exclusion database[5]. Note that in this paper we will treat the rest of the providers as non-fraudulent for evaluation, even though it is evident that there are a significant number of fraudulent actors who have not been identified.

The claims and the provider enrollment data comes from transactional data warehouses. Each claim, consists of several data elements with information about the beneficiary, provider, the health condition (or diagnosis), the service provided (procedure or drug), and the associated costs. Figure 3 shows the different entities and their relationships that are present in the healthcare claims data. Note that the providers typically are affiliated to each other through orga-

[5]`https://oig.hhs.gov/exclusions/exclusions_list.asp`

nizations such as hospitals. This information and additional data about the providers is present in the provider data.

6. LARGE SCALE TEXT ANALYTICS

Simply stated, the two ultimate goals necessary to address rising healthcare costs are: 1). a healthy population, and 2). optimal healthcare in terms of cost and quality. To reach these goals, the first vital step is to understand the current landscape in terms of prevalent diseases and the resulting treatments and costs. Identifying the key disease profiles for patients will allow segmentation of the population into groups which can then be targeted for proactive care or lifestyle changes. For providers, typical treatment profiles used by doctors and hospitals will be instrumental in identifying the costly areas which need to addressed through policy changes or medicinal research. Moreover, such profiles can also be used to compare providers across the country and potentially across organizations to identify fraudulent (upcoding) or wasteful providers.

In the first case study, we show how such profiles can be generated from the claims data (See Figure 3) using advanced text analytic solutions [18]. Text data, especially from the web domain, has been the foremost target of the big data paradigm and a host of open-source solutions (Apache Hadoop based Mahout library [8], MADlib for parallel databases [9]) exist for deploying text mining algorithms on massive text data sets. The interaction between text mining and healthcare, for obvious reasons, has been in analyzing the text available within the patient health records (clinical data) [25]. But these solutions have never been applied in the context of healthcare claims data.

6.1 Representing Entities as Documents

To capture the behavior profiles of the providers and beneficiaries we construct several sparse matrices from a temporal aggregate of claims data, as follows:

Let the set of providers be denoted as P, set of beneficiaries be denoted as B, set of procedures as C, set of diagnoses as G, and set of drugs as D. Let the symbol $\mathbf{XY}$ denote a matrix with $X \in \{B, P\}$ representing rows and $Y \in \{G, C, D\}$ representing columns. For example, the matrix $\mathbf{PG}$ (providers vs. diagnoses) captures the nature of diagnoses that a doctor assigns to patients. Each cell PC_{ij} in the matrix $\mathbf{PC}$ denotes the number of times the provider $P_i \in P$ uses the procedure $C_i \in C$ in the given time frame.

Each of matrices thus created can be viewed as a **document-term matrix** with providers/beneficiaries as *documents* and drugs/procedures/diagnoses as *terms*. Such representation opens up the claims data to a wide spectrum of existing text analysis methods [18]. Given that multiple document-term matrices can be generated for the same entity (providers), it also allows application of methods that deal with learning from multiple views [16].

6.2 Profiling Providers

In this study we used topic modeling using *Latent Dirichlet Allocation (LDA)* [5] as our text analysis tool. LDA is a probabilistic topic model which is widely used for determining hidden topics in a set of documents. In the LDA model, each document (provider) is represented as a mixture of a fixed number of hidden topics and each topic is a probability distribution over a vocabulary of words (diagnosis codes).

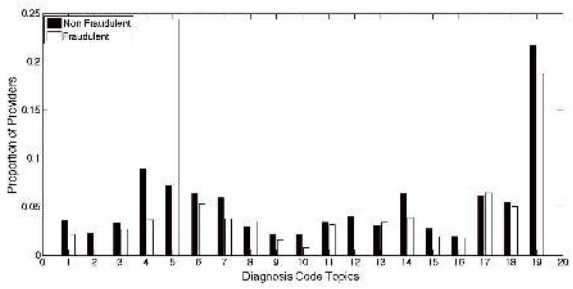


Figure 4: Relative proportion of fraudulent and non-fraudulent providers assigned to each topic of diagnosis codes.

We applied the Mahout implementation of *collapsed variational Bayesian inference* (cvb) algorithm [28] for LDA on the provider-diagnoses ($\mathbf{PG}$) matrix constructed from a year of claims data (361,117 providers, 9,378 diagnosis codes, 43,331,004 tuples). The matrix was normalized using *tfidf* normalization. Using LDA we identified 20 hidden topics from the $\mathbf{PG}$ matrix.

From healthcare perspective, each topic can be thought of as a category for providers. While one would expect that topic driven categorization should closely match the actual specializations of the providers, we found out that this was not always the case. While some topics were dominated by diagnoses that belonged to same area of medicine (e.g. , oncology, ophthalmology), there were other topics which were made up of diagnosis codes that are seemingly different from each other. For example, one topic consisted of diagnosis codes associated with *Diabetes* as well as *dermatoses* (skin condition). While this might appear surprising, further research revealed a medicinal connection between the two [17].

Another possible application of the topic modeling is to use the topic distributions as features or profiles for the providers. To validate the discriminatory potential of topic distributions we conducted the following analysis. For each provider, we chose the topic with highest probability assigned by LDA. We then compared the proportion of fraudulent and the non-fraudulent providers (using the list of fraudulent providers as described in Section 5) that fall under each topic. The relative proportions are shown in Figure 4.

The non-fraudulent providers are evenly distributed across most topics, except for the topic 19. This particular topic is composed of "generic" diagnoses such as *follow-up surgery* and *benign tumor* and hence is expected to represent a large number of providers. For the fraudulent providers, the distribution follows the non-fraudulent providers, except for topic 5, which represents close to 25% of all fraudulent providers as opposed to only 7% of non-fraudulent providers. Topic 5 is dominated by very distinct diagnosis codes.

From the domain perspective, this discovery is valuable since it identifies the diagnosis codes that are used by same providers and have historically been targets of fraud. Given the fact that the codes in topic 5 are medicinally distinct from each other, such discovery can only be made by methods that allow all diagnosis codes to be related to each other, i.e. , topic models.

The promising findings in Figure 4 show that analyzing claims data using text mining methods can reveal very in-

teresting interactions and patterns. Similar analysis can be conducted for other matrices for additional insights.

7. SOCIAL NETWORK ANALYSIS

As health insurance companies shift focus from fraud detection to fraud prevention, building a predictive model to **estimate the risk of a provider before making any claims** has been a challenging problem. Furthermore, substantial amount of healthcare fraud is expected to be hidden in the relationships among providers and between providers and beneficiaries making insurance claims. In this case study, we present results of applying social network analysis methods [21] to understand the relationships of providers in the healthcare system and visualizing features and patterns of fraudulent behaviors in such a network. We describe the construction of social-network features and the predictive model built on those features as a solution to assessing healthcare fraud risk at the time of enrollment.

7.1 Constructing a Provider Social Network

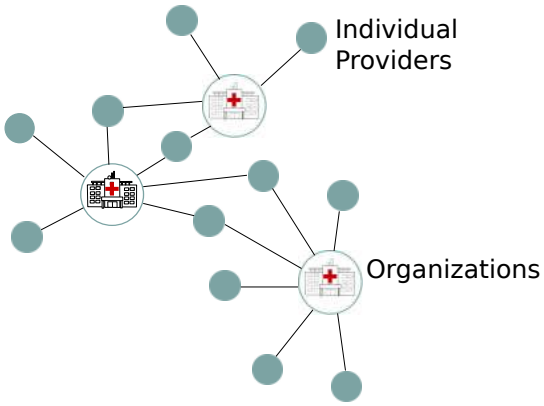


Figure 5: A Sample Provider Network

Providers in the US healthcare system are typically associated with multiple hospitals and health organizations. The information about the providers can be obtained from multiple sources. Some of such data sources are public[6] while others may be purchased[7]. We use data from such sources to construct a social network in which providers (both individual and organizations) are the nodes. The edges are between individual and organization nodes (See Figure 5). A graph when constructed for all providers in the United States is expected to have nearly 35 million nodes and more than 100 million edges. In this study we construct a graph for providers in the state of Texas with almost 1 million nodes and close to 3 million edges.

7.2 Properties of the Provider Network

A snapshot of the provider network for the state of Texas is shown in Figure 6. This provider network is different from a typical "social network" in the following ways: (i) the networks consist of both organizations and individuals. This introduces a latent hierarchy in the network because several individual physicians work for organizations and can also own group practices, (ii) the network is a collection of disconnected graphs, the largest network being a network of a few 100,000 providers and the smallest as little as 3, and, (iii) the network is constructed based on self-reported data and inferred data based on subject matter expertise and may be subject to omissions, errors and quality issues.

[6] https://nppes.cms.hhs.gov/NPPES/
[7] http://www.healthmarketscience.com/

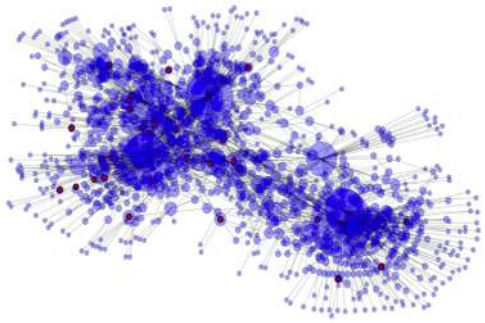


Figure 6: Snapshot of the provider network for Texas. The width of circle at each node denotes the number of affiliations. The large circles indicate organizations, such as hospitals. Nodes in red are fraudulent providers.

7.3 Extracting Features from Provider Network

For this particular study, we focus on analyzing the network properties with respect to the fraudulent providers described in Section 5. The objective is to extract multiple features for every node in the network and use them for discriminating between fraudulent and non-fraudulent providers. We investigated several network based features [21, 23] as listed in Table 1.

Centrality	Degree, Closeness, Betweenness, Load
	Current Flow Closeness, Communicability
	Current Flow Betweenness, Eigenvector
Assortativity	Average Neighbor Degree
	Average Degree Connectivity
Clustering	Triangles, Average Clustering
	Square Clustering
Communities	K-Clique
Components	Connectivity
Cores	Core number, k-Core
Distance Measures	Center, Diameter, Eccentricity
	Periphery, Radius
Flows	Network Simplex
Link Analysis	PageRank, Hits
Rich Club	Rich Club Coefficient
Shortest Paths	Shortest Path
Vitality	Closeness Vitality

Table 1: Network features studied for the provider network [21].

7.4 Relevance for Identifying Fraud

For each feature we estimated its capability to distinguish between fraudulent and non-fraudulent nodes using the *Information Complexity* (ICOMP) measure [15] which compares the distribution of the features for the fraudulent and non-fraudulent populations. The five network based features that we found to be most distinguishing were: **Node degree**, **Number of fraudulent providers in 2-hop network**, **Page rank**, **Eigenvector centrality**, and **Current-**

flow closeness centrality. Figure 7 shows the distribution of each feature with respect to the fraudulent and non-fraudulent populations.

For instance, the red line in 7(a) indicates the node degree distribution for providers previously identified as fraudulent. The blue lines are for a random sample of non-fraudulent providers. We observe that increase in degree of provider correlates to a higher risk of fraud. Similar conclusions can be drawn from analyzing the 2-hop network (See Figure 7(b)). In fact, the chance of finding other fraudulent providers within the 2-hop network of a fraudulent provider is ∼40% compared to the chance of finding a fraudulent provider within the 2-hop network of a random provider (∼2%).

Given the ability of the above features to distinguish between fraudulent and non-fraudulent providers, we plan to utilize them within either an unsupervised multivariate anomaly detection algorithm [7] for automatic detection of such providers or in a binary classification algorithm that learns from the available labeled data.

8. TEMPORAL ANALYSIS OF CLAIMS SEQUENCES

In the context of identifying healthcare fraud perpetrated by providers, two generic response mechanisms are possible: 1). identify and prosecute providers *after* claims are submitted (*pay and chase*), and 2). timely denial of payment for a submitted claim based on the associated risk. Whereas legal prosecution is expensive, time-consuming and difficult to wage, a policy of *selective denial* of payment based on statistical risk factors is much easier to implement once the risk estimation algorithms have been developed.

In this case study we use temporal analytics to address the following two questions: i). How can we identify the transition of a *good* provider into a *bad* actor in an online fashion using the temporal sequence of claims?, and ii). how can the temporal sequence be used to discriminate fraudulent providers from others?

We pose the first question as a change-point detection problem and employ a statistical process control methodology to identify the transition. The strength of this method is that it can be implemented online to examine each claim as it enters a processing queue for payment. For the second question we compare the temporal claim submittal patterns of every provider to estimated population norms for similar providers (e.g., by speciality and geographic location) and define features from these comparisons. Classifiers can subsequently be trained to learn the differences between known fraudsters and presumed normal providers.

8.1 Change-point Detection with Statistical Process Control Techniques

The statistical process control (SPC) literature [20] has evolved into a fairly mature technology for implementation of a temporal approach to processing data sequences.

A number of methods have been proposed in SPC theory to monitor processes for exceedance of control limits. One popular metric is the cumulative sum (CUSUM) statistic [22]. Here we illustrate an application of CUSUM to identify changes in the patient enrollment. A useful assumption in this approach is that a fraudulent providers often start "taking" more patients than usual [27].

Suppose we have a time-ordered sequence of claims $X = x_1, x_2, \ldots, x_n$. This sequence could represent all insurance claims submitted by a single provider over a fixed time interval (e.g., one year). One of the simplest SPC statistics is the Bernoulli CUSUM [26], where X is simply a vector of zeros and ones. For example, we can define x_i according to the following:

$$x_i = \begin{cases} 1 & \text{if the } i^{th} \text{ claim has a new beneficiary number} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

This vector tracks the introduction of new beneficiaries in the stream of claims submitted by a specific claimant, and provides a basis for estimating whether a large number of new beneficiaries were seen by a provider during a particular time interval. The Bernoulli CUSUM statistics to analyze this vector are:

$$S_t = \max(0, S_{t-1} + L_t), t = 1, 2, \ldots, \quad (2)$$

where $S_0 = 0$ and the chart signals if $S_0 > h$. The values of the log-likelihood scores are

$$L_t = \begin{cases} \ln\left(\frac{1-p_1}{1-p_0}\right) & \text{if } X_t = 0 \\ \ln\left(\frac{p_1}{p_0}\right) & \text{if } X_t = 1 \end{cases} \quad (3)$$

A more common form of fraud occurs when providers start taking patients with conditions different from their past profile. Given that the condition codes can have multiple categories, the above method needs to be generalized to a multinomial case. The multinomial CUSUM statistic [13] can be applied here as follows:

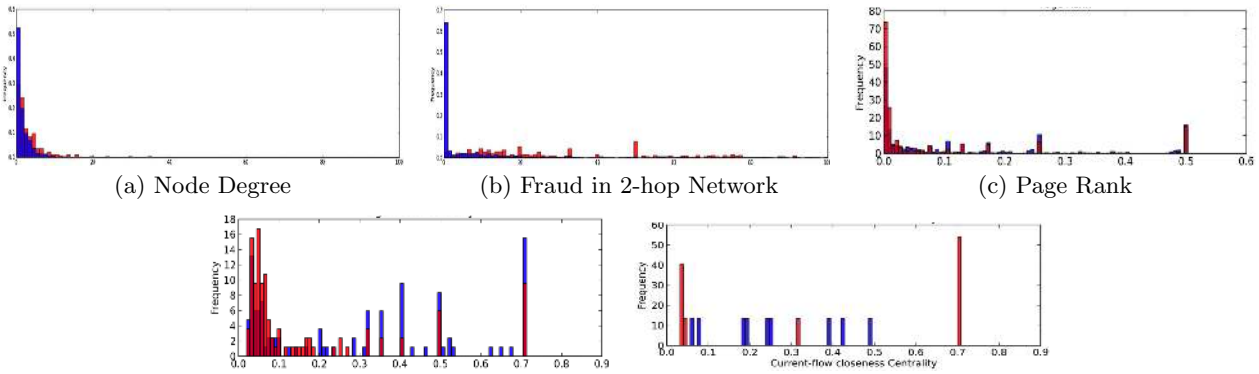
$$L_t = \ln\left(\frac{p_{i1}}{p_{i0}}\right) \text{ when } X_t = i \quad (4)$$

Where p_{i1} is the i^{th} alternative hypothesis, and p_{i0} is the i^{th} null hypothesis. A typical multinomial/categorical CUSUM for a "presumed normal" is shown in Figure 8(a). The CUSUM statistic spikes when the provider uses different condition codes than the typical, but falls back to 0 since the atypical behavior is sporadic. On the other hand, Figure 8(b) shows the CUSUM statistic for an unusual (potentially fraudulent) physician who uses many condition codes that are not typical for his speciality; the CUSUM statistic captures this unusual behavior.

A complete set of out-of-control probabilities are selected for the multinomial CUSUM. In the absence of a specific alternative hypothesis, a simple method of formulating the out-of-control probabilities is to assume that every probability reverses direction to become less extreme, i.e., probabilities migrate in the direction of the grand mean. We specify a proportional change constant to compute the exact probabilities. The alternative hypothesis $\mathbf{p_0}$ is then $\mathbf{p_1} = \mathbf{p_0} + c * (m - \mathbf{p_0})$ where c is the proportional constant and the mean m is simply the reciprocal of the number of categories.

8.2 Anomaly Detection using CUSUM Statistic

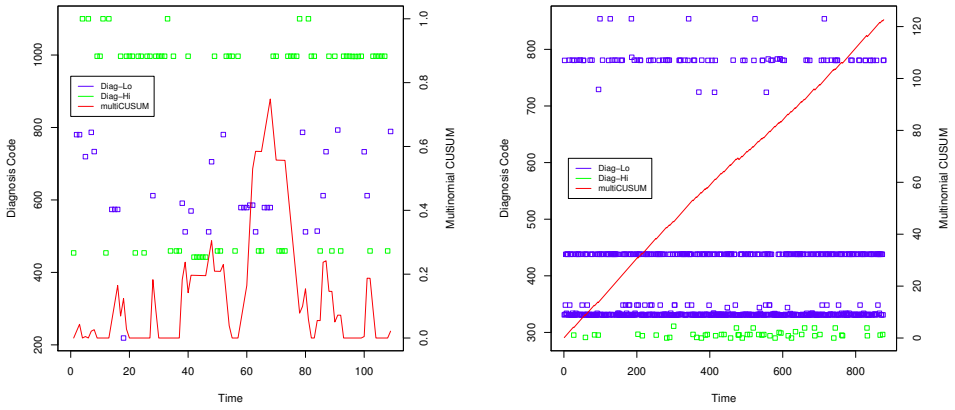
One promising metric for screening anomalies is the maximum value of a CUSUM statistic over a fixed time interval. However, this statistic is biased by the number of claims submitted by a provider during that interval. Some outlier providers exhibit continuously anomalous behavior, even



(a) Node Degree (b) Fraud in 2-hop Network (c) Page Rank

(d) Eigenvector Centrality (e) Current-flow Closeness Centrality

Figure 7: Distribution of top distinguishing features for fraudulent (red) vs. non-fraudulent providers (blue).



(a) Normal Provider (b) Potentially Fraudulent Provider

Figure 8: Multinomial CUSUM chart to track time ordered condition codes from insurance claims. Green squares indicate the typical codes and blue squares indicate the atypical codes for the given speciality.

over a large time interval, and their CUSUM statistics often resemble a linear function. This pattern suggests another metric that is not similarly biased by the length of the claims sequence - the average CUSUM rate. Figure 9 shows a scatterplot using these metrics for a typical provider population color-coded by speciality. The scatterplot displays a cornucopia-shaped pattern with the tail originating at the lowest CUSUM values, and the mouth arcing upward toward the highest CUSUM rates. Anomalies are separated from the main cluster near the top of the scatterplot. Further analysis is required to determine whether these anomalies are normal in a statistical sense, or whether they are more likely members of an anomalous cluster. A horizontal line boundary is drawn at a CUSUM rate of about 0.35 to suggest a possible division of outliers from normals.

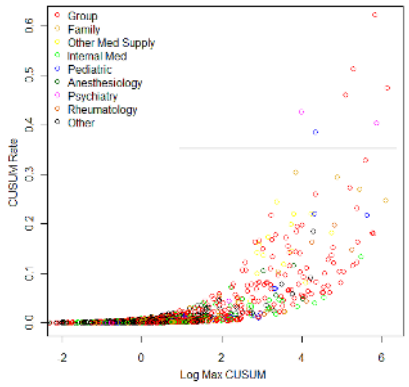


Figure 9: Distribution of CUSUM metrics for a provider population.

8.3 Temporal Feature Construction

The previous section explored the possibility that a provider presumed to be normal up to some position within a temporal claims sequence is diverted either temporarily or permanently to anomalous behavior. Additionally we expect that some providers are either fraudsters from the moment they

enroll in an insurance program, or that they revert to fraudulent activity permanently at some time before the beginning of a limited claims sequence. In this case the availability of provider ground truth affords the analyst an opportunity to go beyond anomaly detection as a method of identifying potential bad actors. In particular, if every provider can be labeled as either "bad actor" or "presumed normal", we can discriminate between temporally stable characteristics of normal and bad actors. Here we assume that providers naturally cluster into a main normal group and a main bad actor group. This assumption may only be approximately correct, especially for the class of bad actors if multiple paths exist to fraudulent behaviors.

In this section we extract 10 temporal features for each provider based on their submitted claims. **Given the fact that some of these features might "help" real providers to adapt and avoid future identification, we are not disclosing the actual features in this paper**. In general, we consider a set of temporally stable features are defined over observed fields in a claims sequence. Features may either be direct functions of elements of the claims sequence or goodness-of-fit statistics that compare empirical distributions to normative distributions. While some features are conditioned on the provider speciality (denoted as *Spec* features) other are independent (*NonSpec* features).

To assess the value of such features, we train two weighted binary logistic regression classifiers [3] for the *Spec* and *NonSpec* feature sets, respectively. We use a labeled training set with 8557 instances constructed using several million Medicaid claims. Less than 1% of the labeled set contained 1% of known fraudulent providers using the fraud data discussed in Section 5. Since the bad actor group was comparatively small, and because we had greater confidence in their labels, the logistic regression was weighted toward the bad actor group by a 10:1 ratio.

We calculated the sensitivity and specificity of both classifiers along with the area under the curve (AUC). Figure 10 shows that including speciality in the model significantly improved performance (DeLong's test; Z=7.22, p<.0001) of the *Spec* model over the *NoSpec* model, yielding an AUC of 0.814.

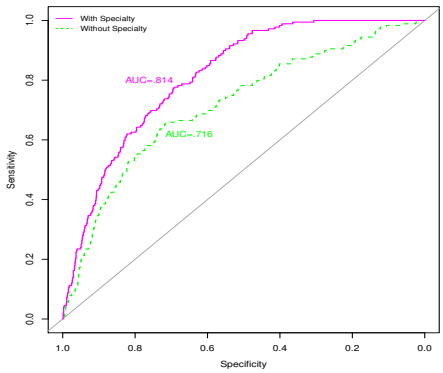


Figure 10: ROC curves for *Spec* and *NoSpec* classifiers using weighted logistic regression

Thus we show that supervised learning methods can be effective in discriminating between bad actors and those presumed normal if the features are well-designed and normative data is available. An immediate next step is to try semi-supervised methods on provider feature vectors so that providers that are not labeled "fraudulent" can be treated as unlabeled.

9. CONCLUSIONS AND FUTURE DIRECTIONS

This paper showcases the relevance of advanced big data analytics in the emerging domain of healthcare analytics using claims data. Our main contribution is the translation of some of key challenges faced by the healthcare industry as knowledge discovery tasks. The three case studies presented in this paper attack the problem of identifying fraudulent healthcare providers in three independent ways, using state of art KDD methodologies, which have never been previously used in this context. Our results on real fraud data highlight the promise that advanced data analytics hold in this important domain. The analysis for these case studies was conducted using the Hadoop/Hive data platform and used open source software such as Mahout, R, and Python networkx[8] and hence are repeatable in other contexts.

In each case study we identified the potential and challenges associated with the existing analytic solutions. Treating providers and beneficiaries as text documents opens the possibility of using the vast text mining literature and the highly sophisticated text mining tools that have been developed specifically for big text data, and can lead to valuable discoveries as shown in Section 6. Studying affiliations between providers as social networks is valuable, given that organized fraud is rampant in the healthcare system, and can be identified by analyzing the relationships between the providers using network science methods. However, we identified certain differences between the provider network and a traditional "social network" which researchers should bear in mind before applying these methods. Temporal analysis methods are also useful because they do not require trained classifiers to identify anomalies, and are sensitive enough to be employed as timely online techniques for detection of transient billing practices that are anomalous. In future, we intend to combine the features generated from each of the case studies in a multi-view learning framework to better identify fraudulent providers.

An important conclusion from these analyses is that while insurance claims data are typically considered as payment records, they contain valuable information that can be used to answer many other healthcare related questions. For instance, studying topics of diagnosis or drug codes (see Section 6) can be done in the context of beneficiaries to understand the major behavior modes of the population in terms of health indicators. Networks that capture interaction between beneficiaries and providers can be constructed from claims data and can be used in conjunction with the provider network to better understand the healthcare system.

10. ACKNOWLEDGEMENTS

Prepared by Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6285, managed by UT-Battelle, LLC for the U. S. Department of Energy under contract no. DEAC05-00OR22725.

[8] `http://networkx.github.com/`

11. REFERENCES

[1] National health expenditure projections 2009–2019. Center for Medicare and Medicaid Services, 2010.

[2] World health statistics. WHO Library Cataloguing-in-Publication Data, 2011.

[3] A. Agresti. *Categorical Data Analysis*, chapter 5. Wiley-Interscience, Hoboken, New Jersey, second edition, 2002.

[4] E. Aleskerov, B. Freisleben, and B. Rao. Cardwatch: A neural network based database mining system for credit card fraud detection. In *Proceedings of IEEE Computational Intelligence for Financial Engineering*, pages 220–226, 1997.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[6] Y. M. Chae, et al. Data mining approach to policy analysis in a health insurance domain. *International Journal of Medical Informatics*, 62(2-3):103–111, 2001.

[7] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection – a survey. *ACM Computing Surveys*, 41(3), July 2009.

[8] C. T. Chu, et al. Map-Reduce for Machine Learning on Multicore. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *NIPS*. MIT Press, 2006.

[9] J. Cohen, et al. Mad skills: new analysis practices for big data. *Proceeding of VLDB Endowment*, 2(2):1481–1492, Aug. 2009.

[10] T. Fawcett and F. Provost. Activity monitoring: noticing interesting changes in behavior. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 53–62. ACM Press, 1999.

[11] A. M. Garber and J. Skinner. Is american health care uniquely inefficient? Working Paper 14257, National Bureau of Economic Research, August 2008.

[12] R. Ghani and M. Kumar. Interactive learning for efficiently detecting errors in insurance claims. In *Proceedings of the 17th ACM SIGKDD*, pages 325–333. ACM, 2011.

[13] M. HÃűhle. Online change-point detection in categorical time series. In T. Kneib and G. Tutz, editors, *Statistical Modelling and Regression Structures*, pages 377–397. Physica-Verlag HD, 2010.

[14] L. Kenworthy. AmericaâĂŹs inefficient health-care system: another look. http://lanekenworthy.net/2011/07/10/americas-inefficient-health-care-system-another-look/, 2011.

[15] S. Konishi and G. Kitagawa. *Information Criteria and Statistical Modeling (Springer Series in Statistics)*. Springer, 2007.

[16] B. Long, P. S. Yu, and Z. Zhang. A General Model for Multiple View Unsupervised Learning. In *Proceedings of the SDM*, 2008.

[17] C. B. Lynde and M. D. Pratt. Acquired perforating dermatosis: association with diabetes and renal failure. *Canadian Med 'Association Journal*, 181(9), 2009.

[18] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[19] J. Manyika, et al. Big data: The next frontier for innovation, competition, and productivity, May 2011.

[20] D. C. Montgomery. *Introduction to Statistical Quality Control*. John Wiley and Sons, 4th. edition, 2001.

[21] M. Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010.

[22] E. S. Page. On problems in which a change can occur at an unknown time. *Biometrika*, 44(1-2):248–252, 1957.

[23] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999.

[24] S. H. Preston and J. Ho. Low life expectancy in the United States: Is the health care system at fault? In *National Research Council (US) Panel on Understanding Divergent Trends in Longevity in High-Income Countries*. 2010.

[25] U. Raja, T. Mitchell, T. Day, and J. M. Hardin. Text mining in healthcare. Applications and opportunities. *Journal of healthcare information management : JHIM*, 22(3):52–56, 2008.

[26] M. Reynolds and Z. Stoumbos. A cusum chart for monitoring a proportion when inspecting continuously. *Journal of Quality Technology*, 1999.

[27] M. Sparrow. *License to steal: how fraud bleeds America's health care system*. Westview Press, 2000.

[28] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *NIPS*, 2006.

[29] S. Wasserman and K. Faust. *Social Network Analysis. Methods and Applications*. Cambridge University Press, 1994.