Journal of Innovative Technology, Vol. 1, No. 1, pp. 51-61, March 2019
http://doi.org/10.29424/JIT.201903_1(1).0007

CASCADED PROPENSITY MATCHED FRAUD MINER: DETECTING ANOMALIES IN MEDICARE BIG DATA

Saad Sadiq [1] and Mei-Ling Shyu [2*]

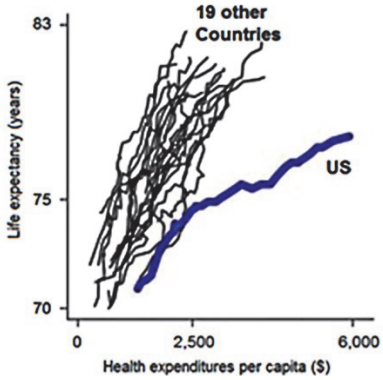
ABSTRACT

Healthcare fraud is a serious cause of concern due to its unrestrained growth in funded medical aid plans around the globe. Apart from the monetary deficiencies caused by fraudulent practices, a greater challenge is the shortage of leftover funding that translates into unavailability of medical services for the ones who need it the most. Organizations such as the Center for Medicare Services (CMS) in the U.S. have started providing access to comprehensive medical big data to face the onslaught of healthcare frauds. Through the use of statistical machine learning and the ability to process medical big data, we are starting to see promising developments for the analysis of fraud in these expansive medical databases. This paper builds upon our previous work in fraud type classifications and the multidimensional Medicare data to provide a multivariate data model that aids in predicting the likelihood of healthcare fraud instances. A novel Cascaded Propensity Matching (CPM) Fraud Miner is proposed to identify fraudulent outliers in the CMS Medicare dataset. The proposed CPM Fraud Miner targets the most widespread known types of malpractices and should be helpful in exploring new and evolved fraud practices. This paper also performs a comprehensive review of current state-of-the-art models in healthcare fraud and functionality evaluation against leading methods with known fraud cases.

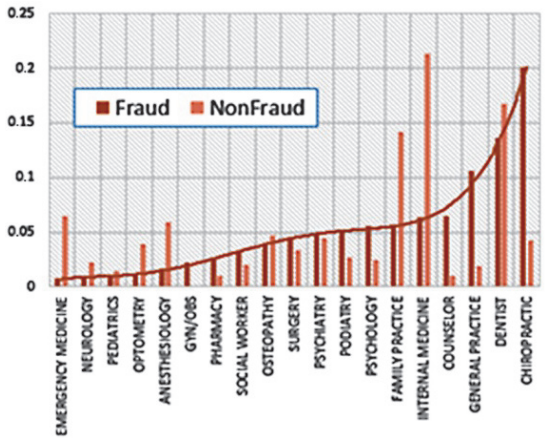
Keywords: Fraud detection, medicare, cascaded propensity matching, fraud miner, propensity score matching, concept drift learning.

1. INTRODUCTION AND BACKGROUND

In a system ravaged by fraud, the waste and abuse of the current state of federally funded healthcare programs such as Medicare and Medicaid are becoming a fraud fest for criminals. An estimated $700 billion from the annual US healthcare budget of $2.7 trillion can be ascribed to fraud which is expected to expand considerably until 2013 (Kelley 2009). This marks US as the highest healthcare spender in the world, as compared to an average of 9.5% among other developed countries (World Health Statistics 2011). While the costs are mounting, the value in treatments and other quantitative measures is declining. This is illustrated in Fig. 1(a) as the measure of life expectancy versus the healthcare spending from 1970 to 2015 in the US and comparable 19 richest nations by GDP (Department of Health and Human Services 2017). In this paper, we present a novel fraud detection method in medical insurance claims, based on unsupervised statistical methods such as propensity matching, mixed variable cosine distances, and inertia-based clustering. We obtain abnormal behaviors of the physicians and evaluate our findings by several verification methods. Our main contributions are shown as follows.



(a) Life expectancy compared to healthcare spending among first world countries



(b) Relative proportions of known fraud cases vs non-fraud cases

Fig. 1 Key indicators of the healthcare industry in the US

Manuscript received June 7, 2018; revised August 23, 2018; accepted September 30, 2018.

[1] Ph.D. student, Department of Electrical and Computer Engineering University of Miami, Coral Gables, Florida, U.S.A.
[2*] Professor (corresponding author), Department of Electrical and Computer Engineering University of Miami, Coral Gables, Florida, U.S.A. (e-mail: shyu@miami.edu).

1) We present an unsupervised fraud miner to detect fraudulent healthcare malpractices and also fraud in general. To the best of our knowledge, this kind of unsupervised approach has not been used in fraud detection.
2) Our novel feature engineering approach extracts unique and highly predictive features from the Center for Medicare & Medicaid Service (CMS) dataset that completely characterizes the feature space.
3) Experimental results indicate that the proposed framework handles medical big data significantly better than other unsupervised methods and gives better performance.

The rest of the paper is organized as follows. Section 2 introduces a comprehensive overview of the previous work done in healthcare fraud detection. Section 3 introduces the CMS. Section 4 presents the details of the proposed framework and section 5 illustrates the experiment setup and the results of our framework. Section 6 discusses and concludes the article with a brief view into the topics of further enhancements and future contributions.

2. EXISTING WORK ON HEALTHCARE FRAUD

The identification of fraud can be handled in several ways using machine learning and data mining (Chen *et al*. 1999; 2003; 2006; 2009; 2013; Chen *et al*. 1999; Huang *et al*. 2002; Lin *et al*. 2007; 2009; Shyu *et al*. 2001; 2003; 2005; 2007; 2008; Zhu *et al*. 2011) where an anomaly detection method can generate an outlier subset from the general data so that the physicians who behave differently can be identified. Other methods may include the following: (1) Density estimation problem with joint probability densities; (2) Clustering problem; (3) Pattern recognition problem; and (4) Anomaly detection problem.

Although there are severe limitations of the supervised learning methods, we present the work from both supervised and unsupervised approaches.

2.1 Supervised Fraud Detection

Supervised fraud mining and abuse discovery in the health-care domain require the methods to be trained on ground truths from former fraud convictions. However, the sheer lack of true positive fraud cases in healthcare makes data a significant weakness in supervised methods. Some of the foremost subdivisions of supervised methods in health care fraud and abuse detection include decision trees (Shin *et al*. 2012; Williams and Huang 1997), neural networks (Liou *et al*. 2008; Ortega *et al*. 2006), genetic algorithms (He and Graco 1998), and Support Vector Machine (SVM) (Kirlidog and Asuk 2012; Kumar and Ghani 2010).

Some notable mentions include Ormerod *et al*. (2003) who proposed a Bayesian Network, where they proposed a Suspicion Building Tool based rule estimator. He *et al*. (1998) used a k-nearest neighbor technique where the distances were estimated by genetic algorithms. However, their method was limited in detecting only two sub-types of fraud, *i.e*., upcoding and doctor shopping. Further details of the different fraud types are provided in section 3. Cooper (2003) used neural networks to predict fraud claims processed by a Chilean health-insurance firm. One of the leading concerns with neural networks is the mandatory requirement of thousands of training samples which are rarely available in any kind of fraud detection application. This scarcity ends up in overfitting the models, producing a relatively big error when new data are tested on the network (Sadiq *et al*. 2017a; 2017b; Sadiq *et al*. 2016; Yan *et al*. 2017).

Yang and Hwang (2006) used the decision tree C4.5 framework to train a model for service provider fraud for the National Health Insurance Administration in Taiwan. A similar approach using the C5.0 framework was used by Williams and Huang (1997) in predicting insurance fraud for the Health Insurance Commission of Australia. Even with a marginally big dataset of 40,000 patients, they reported difficulties concerning using an overly complex tree with several thousand rules, which is unfavorable for interpreting the data.

At present, the majority of healthcare fraud detection methods work on static data and lack real-time prediction. Francis *et al*. (2011) used SVM to develop a real-time method for fraud detection systems. Tsai *et al*. (2014) also lowered the large overhead cost and computational complexity of insurance fraud detection by proposing a knowledge model built on domain knowledge schema and rules.

2.2 Unsupervised Fraud Detection

Fraud miscreants mimic the behavior of dynamic systems by adapting and mixing their strategies to keep a high ratio of their legitimate to fraudulent cases. Unsupervised methods typically assess one's claims in relation to other claims and determine the variable correlations without requiring any ground truth.

The literature is appreciative of the unsupervised methods being used in healthcare fraud and abuse. Some prominent methods include: clustering (Ekina *et al*. 2013; Liu and Vasarhelyi 2013), outlier detection (Batchelor 2015; Jones 2015) and association rules (Kennedy 2016; Mangan 2016). Notable mentions include the work by Yang and Su (2014) who proposed an unsupervised data mining method to evaluate if the physicians follow pre-defined clinical practices. They assert anomalies and outliers based on the distances when providers deviate from the standard clinical practices. Joudaki *et al*. (2016) used unsupervised clustering techniques on physician data. They clustered and ordered critical masses using rules developed by domain experts that affect health expenses. A Korean empirical trial targeted abuse in 3705 internal medicine claims (Shin *et al*. 2012). They estimated a risk value to identify the likelihood of abuse by the physicians and classified physicians using decision trees.

Shan *et al*. (2008) worked on association rule mining to verify insurance claims of specialist providers. Such as, if a provider prescribes treatment A and drug B, then the treatment will consequently follow on to drug C with a certain likelihood. The cases breaking these rules were treated as outliers and were assigned an escalated risk of fraud.

A web-based unsupervised system called SmartSifter was developed by the Australian Health Insurance Commission (Yamanishi *et al*. 2004). They used finite mixtures coupled into discounting learning algorithms that process non-stationary datasets. Meanwhile, a Bayesian co-clustering algorithm specifically for 'conspiracy fraud' was proposed (Ekina *et al*. 2013).

3. CMS DATASET

The CMS data are part of the Dept. of Health and Human Services dataset, released in 2012 to engage the public research community in fighting healthcare fraud. Since the CMS dataset does not provide the label of the fraudulent healthcare providers, the anomaly detection is processed in an unsupervised way. However, we do utilize the fraudulent provider label available from the Office of Inspector General's exclusion database (Department of Health and Human Services 2017) and use a bias adjustment procedure. Our study takes into consideration the medical charges, procedures, prescriptions, drugs and equipment, possible anomalies, and geographical analysis combined with the nationwide procedure charges, and payment distributions. There are more than 20 million insurance claims, distributed over several prescription and drug subcategories and recorded since the year 2012. Table 1 lists some of the most commonly performed procedures in the CMS dataset to illustrate some sense of the size of the data.

Table 1 Count of unique procedures by provider type

Provider type	# of time procedures performed (FL,CA,TX,NY)
Internal Medicine	8200865
Family Practice	6251855
Nurse Practitioner	1540376
Cardiology	1412213
Radiology	850456
Psychiatry	652865
General Practice	589963
Ophthalmology	503518
Neurology	484380
Pulmonary Disease	452059
Nephrology	443789

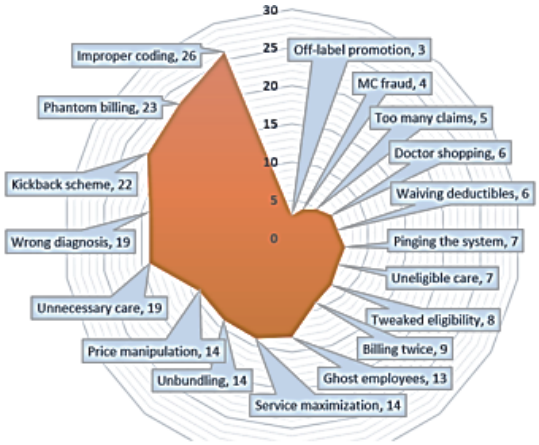
3.1 Profiling Providers

We compared the relevant distributions of fraudulent and the non-fraudulent providers that fall under each type of service and each state as illustrated in Fig. 1(b). Their combined distributions are matched across most areas, except for the 'Family Medicine', 'Internal Medicine' and 'Dentist', where there are far less convicted frauds. For the fraudulent providers, their distribution follows the non-fraudulent providers, except for 'Chiropractic', 'Counselor', 'General practice', 'Psychology' and 'Podiatry' which represent up to 300% markup in fraudulent providers as opposed to the non-fraudulent providers.

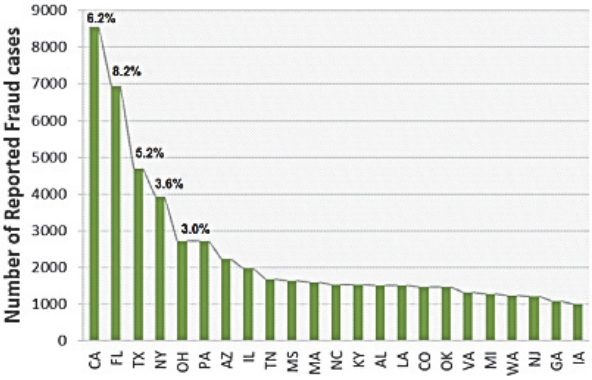
3.2 Types of Fraud

Figure 2(a) illustrates different types of fraud schemes and how many times they were discussed in the literature, thus signifying their importance. Figure 2(b) complements this finding with the number of convicted ground truth cases with the top 5 provider to fraudster ratios. We will now describe the eighteen types of fraud discussed in the literature and specifically target the more damaging types of fraud in our algorithm.

1. Improper coding and upcoding: sometimes called upcoding, is a practice of billing for more costly procedures than the ones actually prescribed.
2. Phantom billing: is a practice of submitting claims for procedures that have never been performed.
3. Kickback schemes: is when physicians and pharmacists write prescriptions from particular drug companies that promise them opulent gifts in return.
4. Wrong diagnosis: is a type of fraud where patients are given false diagnosis to submit claims that yield higher profits.



(a) Major types of frauds in current US healthcare market based on incidents in literature



(b) The number of convicted fraud cases in each state

Fig. 2 Growing healthcare fraud cases and their demographics

5. Unnecessary care: is when providers submit false claims for unnecessary services such as submitting a claim for a patient that has passed away.
6. Price Manipulation: is when providers of medical equipment of health services raise the cost directly or travel to nearby zip codes of higher mean average income.
7. Unbundling: is a practice where a provider breaks down a claim into multiple micro services to obtain a higher profit value.
8. Service Maximization: is providing more service than what is required to treat the patient. This includes unnecessary tests, follow-ups, and consultations.
9. Ghost employees: submit large number of claims for a small relatively small facility.
10. Billing Twice: is when the provider submits the same claim multiple times with slight modifications for the same procedures.
11. Tweaked eligibility: is when patients lie about their situation to claim insurance coverage
12. Uneligible care: is the care provided by people with no credentials or license to perform the procedures
13. Pinging the system: is referring the patients within the same financial organization to elude significant audit scrutiny.
14. Waiving deductibles: are to waive the co-payments for patients to receive a direct benefit from them
15. Doctor shopping: is when patients shop and bribe a doctor to write their desired drug prescriptions.
16. Too many claims: is a malpractice of sending claims for a group of patients or combined treatment sessions, although only a single patient is being serviced.

17. Managed care fraud: is when providers pass the payment risk from the insurance firm to a transitional insurer
18. Off label promotions: is when pharmaceutical companies market their products that are not evaluated and backed by the FDA.

4. CASCADED PROPENSITY MATCHING (CPM) FRAUD MINER

In this paper, we introduce a novel CPM Fraud Miner for the identification of fraudulent outliers in the CMS Medicare dataset. Figure 3 illustrates the basic workflow diagram of the proposed CPM Fraud Miner. The significant contributions of our proposed fraud miner are as follows.

1. Highly accurate healthcare fraud outlier estimator;
2. Achieving an average improvement of 22.3% over the other clustering methods and 35.4% over the other outlier detection methods;
3. Continuous learning on complete dataset and cascaded in time;
4. Applicable to multiple transactional and identity fraud applications; and
5. Unsupervised learning gaining the significant advantage of missing ground truth.

The proposed CPM Fraud Miner first creates a data store profile from all the different CMS sources and performs pre-defined pre-processing and cleaning. Then, several statistics (derived features) are estimated from the raw data, as detailed in Section 5. The data are broken by CMS in sub-domains, states, years, and treatment categories to enhance reusability. We perform schematization and normalization to join all data stores in one schema as shown in Fig. 3. This is necessary for the CPM Fraud Miner to perform unsupervised techniques on the data. The various modules of the proposed CPM Fraud Miner are detailed below in the sequence that they are applied, as also shown in Fig. 3.

4.1 Propensity Score Matching

Our goal is to know if a deliberate fraudulent action causes any perturbation in our data, but all we have in the CMS dataset is observational data. As an example, we want to know if a cataract is always treated by retinoblastoma removal surgery, which is a highly rare plan of action. However, there are cases where years of neglect and other medical conditions might lead to an expensive retinoblastoma surgery. Propensity score matching (Rosenbaum and Rubin 1983) is a way to overcome this problem by making sure that there is no other reason for the outcome to be an outlier and is purely due to the covariates in the causal pathway.

The propensity score is defined as the conditional probability of receiving the intervention given = x , denoted here by $e(x)$ = $P(T = 1 \mid X = x)$. Here $T = 1$ indicates patients who were treated by other physicians; while $T = 0$ for the case in question. The propensity score possesses a balancing property that T and X are conditionally independent given $e(X)$. Thus, variables X are balanced between the two treatment groups after conditioning on the propensity score. This models the feature space into strata of connections between different cases. As an example, consider if several cataract patients are treated with laser surgery while few are treated with retinoblastoma surgery. Then, for all the cases with propensity score estimates that match each other, the

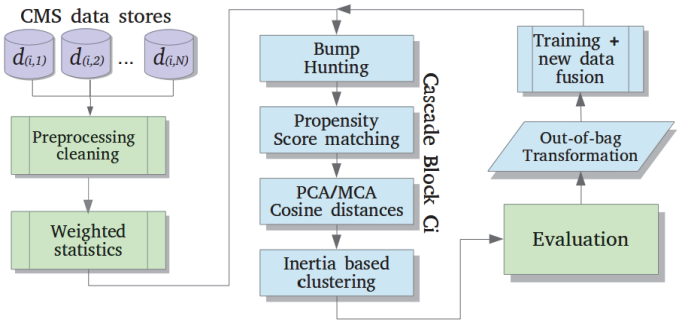


Fig. 3 The proposed cascaded propensity matching fraud miner

expensive treatment was an outlier. If the propensity score is bounded $0 < e(X) < 1$, then the treatment assignment is conditionally independent of the potential outcomes given the propensity score, *i.e.*, $T \perp Y(0), Y(1) \mid e(X)$. Such a result is the foundation for the estimators based on stratification or matching of the data on propensity scores. The derived weighted estimators are shown in Eq. (1), that become the basis for the weighted propensity score estimations. For more details, please see Lunceford and Davidian (2004).

$$E\left[\frac{T\,Y}{e(X)}\middle|X=x\right]=E[Y|T=1,X=x],$$
$$E\left[\frac{(1-T)\,Y}{1-e(X)}\middle|X=x\right]=E[Y|T=1,X=x] \quad (1)$$

4.2 Potential Outcomes Model

After finding the cases with patients treated by other physicians (T = 1), the average of their outcome values is calculated. Next, we subtract the candidate outcome vs. propensity score matched Y from the other treatment condition and obtain the potential outcome (PO). As if the person had both treatments, one from the original physicians and from all other physicians averaged. This counterfactual perspective (Rubin 1990) gives us the comparative effectiveness of treatment options between supposedly different providers. A higher difference means bigger outliers in later stages.

Formally, let (X_1, T_1, Y_1), ..., (X_I, T_I, Y_I) denote the data where X_i is the covariate vector for individual i and Y_i is the observed outcome. Here, T_i denotes the fraud class of case i. For concreteness, let us say $T_i = 0$ represents the non-fraud cases, and $T_i = 1$ represents the fraud cases. Our goal is to estimate the difference in treatments, defined as the difference in the mean outcome for an individual under both fraud classes, conditional on the given features. More formally, let $Y_i(0)$ and $Y_i(1)$ denote the potential outcome for case i under treatments $T_i = 0$ and $T_i = 1$, respectively. Given $X_i = x$, the difference in treatments for i is defined as the conditional mean difference in potential outcomes (as shown in Eq. (2)).

$$\tau(x)=E[Y_i(1)|X_i=x]-E[Y_i(0)|X_i=x] \quad (2)$$

The measure τ indicates the difference in treatments if the

same patient is hypothetically treated by two competing physicians at the same time. However, the difficulty with estimating Eq. (2) is that although potential outcomes $Y_i(0)$ and $Y_i(1)$ are hypothesized to exist, only the outcome Y_i from only one actual provider is available. We assume that the treatment assignment is conditionally independent of the potential outcomes given the variables, *i.e.*, $T \perp Y(0), Y(1) \mid X$. Thus, we have

$$\begin{aligned}\tau(x) &= E[Y(1) \mid T=1, X=x] - E[Y(0) \mid T=0, X=x] \\ &= E[Y \mid T=1, X=x] - E[Y \mid T=0, X=x] \end{aligned} \quad (3)$$

Thus, under the conditional independence of Eq. (3), $\tau(x)$ becomes estimable as each instance x_i can now be expressed in terms of other observable cases from X.

4.3 Normalizing the Mixed Feature Types

The proposed CPM Fraud Miner uses a fusion of Multiple Correspondence Analysis (MCA) (Abdi and Valentin 2007) and Principal Component Analysis (PCA) (Jollife 2002) for multivariate mixed data. Consider a multivariate dataset of features X and the univariate outcome Y with samples I. The outcome Y is not the true outcome, and it is a derived variable from the feature space X. Then, let X_q be the number of quantitative features and X_c be the categorical variables such that $X = X_q + X_c$. Moreover, let x_i be an arbitrary sample and X_k be any arbitrary feature variable. Then at the crossing of row x_i and column X_k, belonging to set j, we have:

1. the value x_{ikj} of the variable X_k for the sample x_i, when j is a quantitative set;
2. the value 1 if x_i belongs to the category X_k and 0 if it doesn't, when j is a categorical set.

First, we evaluate the principal components of the subset j, and then normalize this subset by dividing the weights of the features by the first eigenvalue λ_1^j. These standardized and unit variance principal components make up the basis function for our feature space X. Second, all categorical features are converted into a disjunctive data table having the binary indication values (0 or 1). Next, we perform MCA on the resultant data in order to scale the features and get the eigenvalues. Then the normalized subsets are merged to construct a distinct matrix. This equalizes the effect of continuous and categorical features so that both types of features equally influence the variability. The equivalence between subsets X_q and X_c is obtained as follows.

1. Apply Global PCA to the table with the general term $(z_{ikj} - w_{kj})/w_{kj}$;
2. Assign the weight $w_{kj} \cdot Q_j$ to column X_k of feature subset j; and
3. Assign the weight p_i to row x_i.

Here, $z_{ikj} = 1$ if i belongs to the category k and 0 if otherwise. $W_{kj} = \sum_{x_i \in I} p_i \cdot z_{ikj}$ with p_i being the uniformly distributed weight allocated to each sample x_i, with a default value of 1, and Q_j are the distinct categories in subset j. A distance is calculated among each variable in the form of a weighted sum. This weighted sum helps obtain the final square distances between individual features. This distance metric is defined as:

$$\begin{aligned} d^2(x_i, l) = &\sum_{j \in jq} \frac{1}{\lambda_1^j} \sum_{X_k \in K} \left[\frac{x_{ikj} - x_{lkj}}{s_{kj}} \right] \\ &+ \sum_{j \in jc} \frac{1}{Q_j w_{kj}} [z_{ijk} - z_{lkj}]^2 \end{aligned} \quad (4)$$

where K refers to the total number of features from both quantitative X_q and categorical X_c subsets. Equation (4) indicates the importance of variables in estimating the global principal components as follows.

1. Quantitative variables j ($j \in J_q$) evaluate the distance between units x_i and l when PCA is applied; and
2. Categorical variables j ($j \in J_c$) evaluate the distance between units x_i and l when MCA is applied.

We can then perform inertia-based clustering on the principal components by evaluating their Euclidean distances.

4.4 Inertia-Based Clustering

Inertia or within cluster sum-of-squares is a common clustering technique used to detect the splits. We use inertia to perform the Ward's criterion clustering (Murtagh and Legendre 2014) on the principal components evaluated in the PCA/MCA step as shown in Eq. (5). Ward's criterion breaks the total cluster inertia to two separate parts.

$$\begin{aligned} \sum_{k=1}^{K} \sum_{c=1}^{C} \sum_{i=1}^{I} (x_{ick} - x_k)^2 &= \sum_{k=1}^{K} \sum_{c=1}^{C} I_c (\bar{x}_{ck} - \bar{x}_k)^2 \\ &+ \sum_{k=1}^{K} \sum_{c=1}^{C} \sum_{i=1}^{I_c} (x_{ick} - \bar{x}_{ck})^2 \end{aligned} \quad (5)$$

Here, the left hand side of Eq. (5) refers to the total inertia, and the two parts on the right hand side refer to the between inertia and within inertia. Let x_{ick} be the value of variable X_k for sample x_i of cluster c, $\bar{x}_{ck}$ be the mean of variable X_k for cluster c, where K refers to the total number of columns in the dataset. Let $\bar{x}_k$ be the overall mean of variable k, and Ic be the number of samples in cluster c. The hierarchical divisions are then defined in terms of the correlation between each categorical variable and

1. each quantitative feature by the square correlation ratio η^2 ; and
2. each categorical feature by the Cramer's coefficient V (as given in Eq. (6)). Cramer's coefficient standardizes the χ^2 statistic by the maximum value of the χ^2 statistic. This helps normalize the categorical to categorical feature correlations.

$$V = \sqrt{\frac{x^2}{G * \min(I-1, K-1)}}, \quad (6)$$

where χ^2 is the chi-square statistic, G is the grand total of the table, and I and K refer to the total numbers of samples and features of the table, respectively.

5. EXPERIMENT SETUP AND RESULTS

Empirical data indicate that the states of Florida, California and Texas are the hotbeds of healthcare fraud. Our experiment uses data from the years 2012-2015 and contains close to 20 million insurance claim records. The sheer size of the data made it impossible to use conventional data modeling tools. Thus, we use a 48 core Intel Xeon processor with 192 GBytes of memory to load the data in the R environment. The feature processing and method computations are performed using a Titan Xp GPU with 3840 CUDA cores.

Further analyses reveal that the population density and the number of medical service providers are important considerations in evaluating high fraud conditions. An important empirical observation was obtained through the CMS big data, as indicated in

Fig. 2(b), that the state of Florida has the highest provider to fraud ratio in the nation. This was backed by numerous recent high profile fraud convictions caught in South Florida and listed in the national database (Department of Health and Human Services 2017). Therefore, our focus is on Florida, California and Texas to ensure a higher probability of finding fraudulent outliers and medical payment scammers in the data.

Our selection of three critical states (FL, TX and CA) results in a subset of dataset composed of 7,983 distinct types of procedures done by 509,606 physicians practicing in 158 medical specialties. Table 2 records the statistics of our subset composed of treatment diagnosis (Provider utilization), medical equipment (Prescriber DME-prosthetics, orthotics and supplies), and drug prescription (Drug Public Use File data, respectively. Each physician is denoted by his or her National Provider Identifier and each procedure is labeled by its Healthcare Common Procedure Coding System (HCPCS) code. Figure 4 shows the resulting multidimensional schema that we have developed from the existing data stores.

Table 2 Vectors alignment for propensity score matching

Category	Medical Areas	Provider Procedures	National Provider Identifier
Provider utilization	91	5924	307151
Provider DME-POS	161	999	102115
Drug PUF	192	2954	305414

After the individual variable derivations, we go for the second and third degree variable interactions. These variable interactions capture mixed fraud practices that model the mixing of fraud techniques. For the outcome variable, we formulate a conditional probability variable for the drug, equipment, and services given by a provider pertaining to a specific area of medicine (as shown in Eq. (7)). This calculates, for example, if a physician prescribes muscle relaxer to a patient, how likely it is that the physician is an Orthopedic. A lower value of the conditional probability indicates a higher likelihood that the medical treatment is a fraudulent flag or an improper prescription.

$$P(Area_x \mid Prescription_y) = \frac{P(Area_x \cap Prescription_y)}{P(Prescription_y)} \quad (7)$$

We also calculate the mean of that specialist's treatment area and evaluate ANOVA to prove that the mean value is consistently biased. The ANOVA test is performed to prove the statistical significance against the hypothesis that if a medical specialist has consistent occurrences of charging aberrant prices to the patients, then this may be a plausible indication towards fraud. The F-score in the ANOVA test is calculated using Eqs. (8a) and (8b).

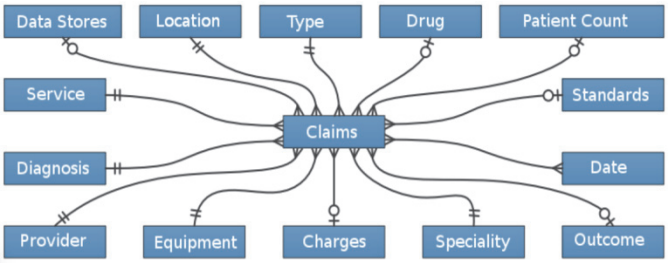


Fig. 4 Medicare multidimensional schema

$$F_1 = \frac{var\,iance\ between\ treatments}{var\,iance\ within\ treatments} \quad (8a)$$

$$F_2 = \frac{MS_{Treatments}}{MS_{Error}} = \frac{SS_{Treatments}/(I-1)}{SS_{Error}/(n_T-1)} \quad (8b)$$

where MS is the mean square, SS is the sum of the square, I is the number of treatments, and n_T is the total number of cases. The derived variables make our proposed fraud miner perform considerably better as the effectiveness of these variables are shown in Section 5.

5.1 Incremental learning

Healthcare systems work in a dynamic environment where behaviors of the providers and patients are continuously changing. There are three sources of dynamic drifts: (1) Streaming new data; (2) Lack of true positives; and (3) Continuously evolving fraud methods.

Thus we use a continuous temporal training model where each future sample is predicted and recruited in the training process using the out-of-bag techniques. We use Random Forests (RF) at the propensity matching stage because it gives us an unprecedented advantage of predicting over full samples without the fear of overfitting.

Let us define an incremental learning process at time instance t, where the historical training data are learned into the object model L_t. Then if a target instance $x_i(t + 1)$ arrives, our goal is to first predict the fraud classification F_{t+1} for this new instance. Second, we want to include this new data in the object model L_t, but as an out-of-bag instance coupled with all or selected historical data $X_{historical} = (X_1, X_t)$. This is illustrated in Fig. 5. By an incremental learning process, the label F_{t+1} becomes available with $x(t + 1)$ as a part of the training to predict $x(t + 2)$.

We use the RF objects in propensity score estimation by modeling the prediction based on the regression of conditional probability outcome variable Y and a propensity score $PrpS$. The historical data are considered to be independently drawn from the

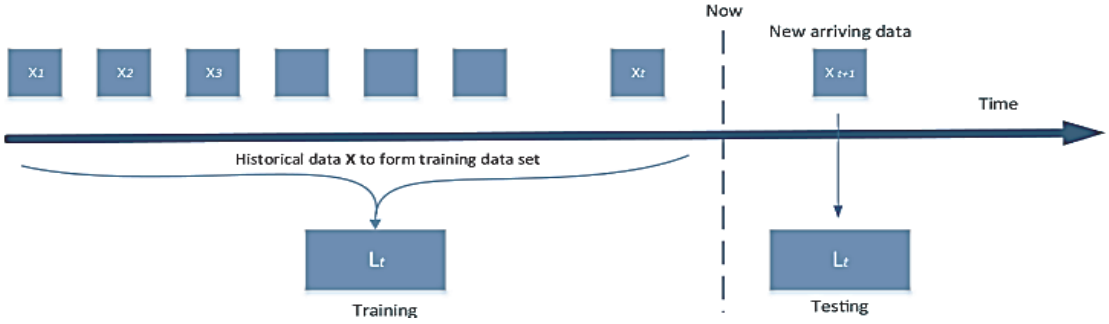


Fig. 5 Incremental learning in time *t*

joint distribution of feature set X and Y and comprises of I samples, namely (x_1, y_1), ..., (x_I, y_I). X is an I by K matrix indicating the total number of insurance claims and their conditional probability Y, where $X = [x_1, ..., x_I]^t$, $Y = [y_1, ..., y_1]^t$. x_i is the subsampled vector (of size 1 by K) from X for the ith sample, K is the total number of features (or dimensions), and Y indicates the vector of outcome variables (y_i, i = 1 to I) that are to be regressed using the RF.

The RF is built by growing the trees based on a random vector θ_t such that each tree predictor $h(\boldsymbol{x}, \theta_t)$ takes on numerical values. The vector θ_t represents the regressed propensity score probabilities based on the conditional probability Y for the tree t. Then, the regression-based RF prediction is defined as the unweighted average over the collection of the predictor trees as shown in Eq. (9), where $h(\boldsymbol{x}, \theta_t)$, t = 1, …, *ntree* is the collection of the tree predictors and x represents the observed input variable vector of length *mtry* with the associated i.i.d random vector θ_t.

$$\bar{h}(x) = \left(\frac{1}{ntree}\right)\sum_{t-1}^{ntree} h(\boldsymbol{x}; \theta_t). \tag{9}$$

As $t \to \infty$, the Law of Large Numbers ensures:

$$E_{X,Y}(Y - \bar{h}(X))^2 \to E_{X,Y}(Y - E_{\theta}(X; \theta))^2 \tag{10}$$

where θ represents the regressed propensity score averaged over ntree trees. The common element in all of these procedures is that for the t^{th} tree, a random vector θ_t is generated, independent of the past random vectors θ_1,… $\theta_{(t-1)}$, but with the same distribution, and a tree is grown using the training dataset resulting in a classifier $h(\boldsymbol{x}, \theta_t)$, where $\boldsymbol{x}$ is an input vector.

5.2 Transformation and Normalization

Since the regression basis is the squared distances in the attribute space, the outlier distances become large when they are squared. For example, the slope m of a regression fit line, as described in Eq. (11), is inversely proportional to the variance of X. Thus, the outliers change the variance of X much higher, causing the fit to rotate down and taking it away from the truth, where X ($x_i \in X$) and Y ($y_i \in Y$) are the attribute space and outcome, respectively and i = 1 to I. We can apply the log rank transformation or the square root transformation to make the slope line straighter. These transformations will also pull in the curve to make the distribution become more Gaussian.

$$m = \frac{\sum_{i=1}^{I}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{I}(x_i - \bar{x})^2} = \frac{Cov(X,Y)}{Var(X)} \tag{11}$$

Moreover, variable interactions cause the feature distributions to become highly skewed.

To rectify this problem, we apply z-score normalization to our dataset L_1 in order to bring all the features to more coherent ranges. The formula for Z-score normalization is given in Eq. (12).

$$Z = \frac{x_j - \mu_{xj}}{\sigma_{xj}}, \tag{12}$$

where Z represents the normalized matrix and attributes x_j for $j \in \{1, ..., K\}$. The resultant changes in the ranges are also shown in Table 3, where U refers to the case when the charged amount is higher than other specialists, S refers to the case when the charged amount is higher than the Medicare standard, and A refers to ANOVA in charged prices. The prime symbols of each of the aforementioned variables represent the normalized ranges. The results indicate that only 2.4% of the learning data are linked to the low conditional probability. We adopt the confusion matrix illustrated in Table 4.

Table 3 Conversion ranges after *z*-score normalization

	Min	1st	Median	Mean	3rd	Max
U	–99.580	–21.880	–4.463	0.0	12.680	4517
U’	–1.887	–0.045	–0.085	0.0	0.24	85.6
S	–97.94	79.31	147.7	241.8	265.4	71650
S’	–0.583	–0.28	–0.16	0.0	0.04	122.6
A	0.0	4800	2753	3.9e27	1.4e4	2.3e32
A’	–0.005	–0.005	–0.005	0.0	0.0057	314.1

Table 4 Confusion matrix of observations

		Classified	
		Observations as Fraud	Observations as Non-Fraud
Actual Data	Observations as Fraud	*a*: the number of observations correctly identified as outliers	*b*: the number of observations wrongfully identified as outliers
	Observations as Non-Fraud	*c*: the number of observations wrongfully identified as non-outliers	*d*: the number of observations correctly identified as non-outliers

After calculating the numbers of a, b, c, and d in Table 4, we calculate the sensitivity, recall, and F1 measure using the following equations.

$$Sensitivity = \frac{a}{a+b} \tag{13a}$$

$$Recall = \frac{a}{a+b} \tag{13b}$$

$$F1 = 2 * \frac{Sensitivity * Recall}{Sensitivity + Recall}. \tag{13c}$$

$$Accuracy = \frac{(a+b)}{a+b+c+d}. \tag{14}$$

The $F1$ measure is defined as the F-score and it has the values between zero and one. A value close to one implies that most of the observations are classified correctly. Another important measure is to evaluate the accuracy defined in Eq. (14).

5.3 Performance Comparison

This section compares the proposed CPM Fraud Miner with other leading algorithms in the domain of fraud detection. We begin by measuring the reproducibility of finding the fraudulent region of the conditional probability. This is done by treating the prediction as a classification problem and identifying if the same bump is found repeatedly in the low conditional probability region. The second step is to characterize the attribute space using variable importance and/or variable correlation metrics. The comparison is to see if some other leading frameworks are able to persistently identify and characterize the attribute space that causes the low conditional probability. To conduct the comparison, our CPM Fraud Miner is evaluated

against several popular classifiers, including the SVM (Suykens and Vandewalle 1999), the Naive Bayes (NB) (Murphy 2006), RF (Breiman 2001), the discriminant analysis classifier (DAC) (Lin and Ravitz 2008; Lin and Shyu 2010; Meng and Shyu 2012), and the Logistic Regression (LR). As shown in Table 5, it is clear that other comparative classifiers struggle to classify the uniqueness of the health care data. Thus, potential frauds with a low conditional probability (*i.e.*, indicating malpractice) may go unidentified. The proposed CPM Fraud Miner, however, achieves a better F-score and better accuracy values because it characterizes the input space and then classifies new instances.

Table 5 F-score and accuracy comparison in a supervised fashion

Classifier	F-score	Accuracy	Recall	Precision
SVM	0.637	0.525	0.73	0.60
NB	0.534	0.519	0.54	0.69
RF	0.601	0.588	0.65	0.52
DAC	0.659	0.564	0.55	0.60
LR	0.578	0.521	0.71	0.75
CPM Fraud Miner	0.812	0.687	0.87	0.85

The average fraud score for Florida, Texas, and California is 6.53% on average. However, we keep the experiment very conservative and only subset the extreme lower end of 2% conditional probability as a thresholds cutoff. We consider these lowest 2% as the extreme outliers and estimate these measurements against the most popular predictors. Table 5 lists the F-score of the methods we compared and depicts that the proposed CPM Fraud Miner outperforms all the major estimating methods by almost 35.4% improvement in direct regression. These marked improvements are due to the ability of the proposed CPM Fraud Miner to shrink the data and handle the data in smaller subsets, thus effectively overcoming the challenge of big data.

We also perform an unsupervised clustering on the feature space by completely ignoring the outcome variables and only using the highly predictive derived features. We fit a hierarchical clustering model on X and then compare the outliers against the low conditional probability ground truth of lower 2%. The results are compared to other unsupervised methods such as unsupervised RF, KNN, C-Means, and Expectation Maximization. Table 6 shows that the other clustering methods are not able to characterize the X space accurately. The results indicate that the other comparing methods had, on average, 22.3% lower F-score than our proposed CPM Fraud Miner.

Table 6 Comparative evaluation of unsupervised models

	F-score	Accuracy	Recall	Precision
K-Means	0.68	0.61	0.73	0.60
UnSupRF	0.71	0.58	0.79	0.56
ExpMax	0.73	0.58	0.65	0.52
C-Means	0.75	0.61	0.55	0.60
CPM Fraud Miner	0.88	0.84	0.87	0.85

5.4 Outlier Significance Test

In this section, we perform statistical evaluation on the measured outliers to see whether they are indeed statistically significant. The reason why an outlier is a disjoint member of the covariate space is due to the following three reasons.

1. There exists a large variability in the measurement;
2. The data point seems like an outlier but actually is just noise;
3. The data point is actually a biased observation because of a biased measurement.

To verify this, we develop a diagnostic statistic that separates out the variance from the bias using the Leverage point outlier test. We break down the treatment codes to calculate the confidence interval for a population mean and evaluate the outliers using Eq. (15).

$$\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + \frac{\sigma}{\sqrt{n}} \tag{15}$$

where $\bar{x}$ is the mean for each treatment/prescription code (HCPCS code), σ is the standard deviatin of the HCPCs codes, and z^* represents the point on the standard normal density curve such that the probability of observing a value greater than z^* is equal to p. p is the statistical significance threshold for the critical value test. For example, if $p = 0.025$, the value of z^* is 1.96 such that $P(Z > z^*) = 0.025$ or $P(Z < z^*) = 0.975$. Table 7 presents a list of z^* values against a range of confidence intervals.

Table 7 Probability of a standard normal variable z for different confidence intervals

Confidence Interval	Z
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

Based on the confidence intervals using the Leverage point outlier estimations (Rousseeuw and Zomeren 1990), the calculated distances are simply the Mahalanobis distance with robust scale estimates. Equations (16a) and (16b) calculate both of the aforementioned distances.

$$MD(x_i) = [(x_i - \bar{\boldsymbol{x}})' \, \bar{\boldsymbol{C}}(\boldsymbol{A})^{-1} (x_i - \bar{\boldsymbol{x}})]^{1/2}, \tag{16a}$$

$$RD(x_i) = [(x_i - \boldsymbol{T}(\boldsymbol{A})^{-1} \, \bar{\boldsymbol{C}}(\boldsymbol{A})^{-1} (x_i - \boldsymbol{T}(\boldsymbol{A}))]^{1/2}, \tag{16b}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\bar{\boldsymbol{C}}(\boldsymbol{A}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})' (x_i - \bar{x})$ are the empirical multivariate location and scale, and T(A) and C(A) are the robust multivariate location and scale estimates. These distances are used to detect the leverage points. Two new variables, Leverage and Outlier, are defined as given in Eqs. (17) and (18).

$$LEVERAGE = \begin{bmatrix} 0 & if\ RD(x_i) \le C(p) \\ 1 & otherwise \end{bmatrix}, \tag{17}$$

where $C(p) = \sqrt{\chi^2_{p;1-\alpha}}$ is the cutoff value evaluated using the significance statistics p and α based off the *Z* score. The final binary outlier decision is obtained by estimating

$$OUTLIER = \begin{bmatrix} 0 & if\ |r_i| \le k\lambda \\ 1 & otherwise \end{bmatrix} \quad (18)$$

where r_i are the residuals, $i = 1, \ldots, n$ based on the cluster estimates, and k and λ are the scaling and tuning parameters respectively with the default values $k = 3$ and $\lambda = 0.75$. The model summary results after removing the low conditional probability instances and with the low conditional probability instances included are given as follows.

Model Summary:
After removing the low conditional probability instances

S	R-sq	R-sq(adj)	R-sq(pred)
2.59199	97.74%	97.17%	96.63%

Model Summary:
With the low probability instances included

S	R-sq	R-sq(adj)	R-sq(pred)
4.71075	96.32%	96.01%	89.63%

The R^2 value is slightly reduced after the insertion of outliers from 97.74% to 96.32%, but the association among Y and X is still considerably strong. This is due to the miniscule number of the estimated fraud outliers in our thresholds. The standard error used to calculate the confidence interval is greater when the outliers are involved, thus increasing the size of our confidence intervals. However, in our hypothesis that including outliers in the dataset would have a significant impact on the X and Y relation doesn't hold well. The 0.05 threshold p-value in each case was 0.001, indicating a high relationship in both cases. The outliers are not highly dominant in the massive data set and does not produce high enough coefficients to impact the X-Y relationship, but the increased error indicates the high leverage in the leverage point test.

6. CONCLUSIONS

Medicare and Medicaid are a group of government funded healthcare assistance programs that serve low-income families and individuals across the U.S. These programs consume one of the highest medical funding to GDP ratios around the world. However, they are fraught with exploitations and fraudulent malpractices costing the system billions of dollars in waste. We proposed a medical fraud miner that analyzed and engineered the medical big data obtained from the Center of Medicare/Medicaid Services. Our fraud miner successfully reduces the 20 million insurance cases to a subset of instances that have high outlier significance. Due to their divergence from their practice and pricing, we assign flags of potential involvement in fraudulent and wasteful use of Medicare insurance. The states of Florida, California, and Texas were isolated for this study because of the high and persistent evidence of health care fraud in these states. The experimental results show that our proposed CPM Fraud Miner can infer a possible subset of healthcare providers who implicate irregular claims and are probably capable of fraud. Several unique and highly predictable features were engineered from the feature correlations to characterize the low conditional probability regions. The consequential model delivers a profound understanding of how certain key predictors act in identifying outliers. The identified cases were validated and compared with other classification methods that indicated significant improvement in F-score readouts of the proposed fraud characterization.

REFERENCES

Abdi, H. and Valentin, D. (2007). "Multiple correspondence analysis." *Encyclopedia of Measurement and Statistics*, 651-657.

Batchelor, E.R.A. (2016). *12 Arrested in Miami-Based Medical Fraud Scheme*. Newspaper, WPLG, August 2016.

Breiman, L. (2001). "Random forests." *Machine Learning*, **45**(1), 5-32.

Chen, S.C., Sista, S., Shyu, M.L., and Kashyap, R. (1999). "Augmented transition networks as video browsing models for multimedia databases and multimedia information systems." *Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence*, 175-182.

Chen, S.C., Shyu, M.L., Peeta, S., and Zhang, C. (2003). "Learning-based spatio-temporal vehicle tracking and indexing for transportation multimedia database systems." *IEEE Transactions on Intelligent Transportation Systems*, **4**(3), 154-167.

Chen, S.C., Shyu, M.L., Zhang, C., and Chen, M. (2006). "A multimodal data mining framework for soccer goal detection based on decision tree logic." *International Journal of Computer Applications in Technology*, **27**, 312-323.

Chen, X., Zhang, C., Chen, S.C., and Rubin, S. (2009). "A human-centered multiple instance learning framework for semantic video retrieval." *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, **39**(2), 228-233.

Chen, C., Zhu, Q., Lin, L., and Shyu, M.L. (2013). "Web media semantic concept retrieval via tag removal and model fusion." *ACM Transactions on Intelligent Systems and Technology*, **4**(4), 61:1-61:22.

Cooper, C. (2003). *Turning Information into Action. Computer Associates, the Software that Manages eBusiness*, Report, available at http://www.ca.com

Department of Health and Human Services (2017). *LEIE Downloadable Databases*. Retrieved from https://oig.hhs.gov/exclusions/exclusions_list.asp

Ekina, T., Leva, F., Ruggeri, F., and Soyer, R. (2013). "Application of Bayesian methods in detection of healthcare fraud." *Chemical Engineering Transaction*, **33**, 151-156.

Francis, C., Pepper, N., and Strong, H. (2011). "Using support vector machines to detect medical fraud and abuse." *Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC2011*, pp. 8291-8294.

He, H., Graco, W., and Yao, X. (1998). "Application of genetic algorithm and k-nearest neighbor method in medical fraud detection." *Asia-Pacific Conference on Simulated Evolution and Learning*, Springer, 1998, pp. 74-81.

Huang, X., Chen, S.C., Shyu, M.L., and Zhang, C. (2002). "User concept pattern discovery using relevance feedback and multiple instance learning for content-based image retrieval." *Proceedings of the Third International Workshop on Multimedia Data Mining, In Conjunction with the 8th ACM International Conference on Knowledge Discovery & Data Mining*, 100-108.

Jones, T. (2015). *Therapist Arrested in Miami Medical Fraud Investigation,* Newspaper, CBS Miami, Dec. 20, 2015.

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., and Arab, M. (2016). "Improving fraud and abuse detection in general physician claims: A data mining study". *International Journal of Health Policy and Management*, **5**(3), 165-172.

Kelley, R. (2009). *Where Can $700 Billion in Waste Be Cut Annually from the US Healthcare System,* Ann Arbor, MI., Thomson Reuters.

Kennedy, K. (2016). *3 Charged in $1 Billion Health Care Fraud That Took Advantage of Medicare in Miami*. Newspaper, NBC 6 South Florida, Jul. 22, 2016.

Kirlidog, M. and Asuk, C. (2012). "A fraud detection approach with data mining in health insurance." *Procedia-Social and Behavioral Sciences*, **62**, 989-994.

Kumar, M., Ghani, R., and Mei, Z.S. (2010). "Data mining to predict and prevent errors in health insurance claims processing." *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 65-74.

Lin, L., Ravitz, G., Shyu, M.L., and Chen, S.C. (2007). "Video semantic concept discovery using multimodal-based association classification." *Proceedings of the IEEE International Conference on Multimedia & Expo*, 859-862.

Lin, L., Shyu, M.L., Ravitz, G., and Chen, S.C. (2009). "Video semantic concept detection via associative classification." *Proceedings of the IEEE International Conference on Multimedia and Expo ICME*, 418-421.

Lin, L. and Shyu, M.L. (2010). "Weighted association rule mining for video semantic detection." *International Journal of Multimedia Data Engineering and Management*, **1**(1), 37-54.

Liou, F.M., Tang, Y.C., and Chen, J.Y. (2008). "Detecting hospital fraud and claim abuse through diabetic outpatient services." *Health Care Management Science*, **11**(4), 353-358.

Liu, Q. and Vasarhelyi, M. (2013). "Healthcare fraud detection: Asurvey and a clustering model incorporating geo-location information." available at http://raw.rutgers.edu/docs/wcars/29wcars

Lunceford, J.K. and Davidian, M. (2004). "Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study." *Statistics in Medicine*, **23**(19), 2937-2960.

Mangan, D. (2016). *$1 Billion Medicare Bust in Florida Is Biggest Criminal Health-Care Fraud Case Ever*. Newspaper, CNBC, July 22, 2016.

Meng, T. and Shyu, M.L. (2012). "Leveraging concept association network for multimedia rare concept mining and retrieval." *Proceedings of the IEEE International Conference on Multimedia and Expo*, Melbourne, Australia, 860-865.

Murphy, K.P. (2006). *Naive Bayes Classifiers*, University of British Columbia.

Murtagh, F. and Legendre, P. (2014). "Wards hierarchical agglomerative clustering method: which algorithms implement wards criterion?" *Journal of Classification*, **31**(3), 274-295.

Ormerod, T., Morley, N., Ball, L., Langley, C., and Spenser, C. (2003). "Using ethnography to design a mass detection tool (mdt) for the early discovery of insurance fraud." *ACM CHI'03 Extended Abstracts on Human Factors in Computing Systems*, 650-651.

Ortega, P.A., Figueroa, C.J., and Ruz, G.A. (2006). "A medical claim fraud/abuse detection system based on data mining: A case study in Chile." *DMIN'06 International Conference on Data Mining,* 26-29.

Rosenbaum, P.R. and Rubin, D.B. (1983). "The central role of the propensity score in observational studies for causal effects." *Biometrika*, **70**(1), 41-55.

Rousseeuw, P.J. and van Zomeren, B.C. (1990). "Unmasking multivariate outliers and leverage points." *Journal of the American Statistical Association*, **85**(411), 633-639.

Rubin, D.B. (1990). "On the application of probability theory to agricultural experiments. Essay on principles. Section 9." *Statistical Science*, **5**(4), 472-480.

Sadiq, S., Yan, Y., Shyu, M.L., Chen, S.C., and Ishwaran, H. (2016). "Enhancing multimedia imbalanced concept detection using VIMP in random forests." *Proceedings of the IEEE 17th International Conference on Information Reuse and Integration (IRI)*, 601-608.

Sadiq, S., Tao, Y., Yan, Y., and Shyu, M.L. (2017a). "Mining anomalies in Medicare big data using patient rule induction method." *Proceedings of the IEEE Third International Conference on Multimedia Big Data*, 185-192.

Sadiq, S., Yan, Y., Taylor, A., Shyu, M.L., Chen, S.C., and Feaster, D. (2017b). "AAFA: Associative affinity factor analysis for bot detection and stance classification in twitter." *Proceedings of the IEEE International Conference on Information Reuse and Integration*, 356-365.

Shan, Y., Jeacocke, D., Murray, D.W., and Sutinen, A. (2008). "Mining medical specialist billing patterns for health service management." *Proceedings of the 7th Australasian Data Mining Conference*, **87**, 105-110.

Shin, H., Park, H., Lee, J., and Jhee, W.C. (2012). "A scoring model to detect abusive billing patterns in health insurance claims." *Expert Systems with Applications*, **39**(8), 7441–7450.

Shyu, M.L., Chen, S.C., and Kashyap, R.L. (2001). "Generalized affinity-based association rule mining for multimedia database queries." *Knowledge and Information Systems*, **3**(3), 319-337.

Shyu, M.L., Haruechaiyasak, C., and Chen, S.C. (2003). "Category cluster discovery from distributed www directories." *Information Sciences*, **155**(3), 181-197.

Shyu, M.L., Sarinnapakorn, K., Kuruppu-Appuhamilage, I., Chen, S.C., Chang, L., and Goldring, T. (2005). "Handling nominal features in anomaly intrusion detection problems." *Proceedings of the 15th International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications*, 55-62.

Shyu, M.L., Quirino, T., Xie, Z., Chen, S.C., and Chang, L. (2007). "Network intrusion detection through adaptive sub-eigenspace modeling in multiagent systems." *ACM Transactions on Autonomous and Adaptive Systems*, **2**, 9:1-9:37.

Shyu, M.L., Xie, Z., Chen, M., and Chen, S.C. (2008). "Video semantic event/concept detection using a subspace-based multimedia data mining framework." *IEEE Transactions on Multimedia*, **10**(2), 252-259.

Suykens, J.A. and Vandewalle, J. (1999). "Least squares support vector machine classifiers." *Neural Processing Letters*, **9**(3), 293-300.

Tang, M., Mendis, B.S.U., Murray, D.W., Hu, Y., and Sutinen, A., (2011). "Unsupervised fraud detection in medicare australia." *Proceedings of the Ninth Australasian Data Mining Conference*, **121**, 103-110.

Tsai, Y.H., Ko, C.H., and Lin, K.C. (2014). "Using common KADS method to build prototype system in medical insurance fraud detection." *Journal of Networks*, **9**(7), 1798-1802.

U.S. H.H.S. (OB) (2017). *FY 2018 Budget & Performance*. Available: http://www.hhs.gov/budget/

U.S. Government Accountability Office (2012). *Medicare Fraud Prevention: CMS Has Implemented a Predictive Analytics System, But Needs to Define Measures to Determine Its Effectiveness*.

Williams, G.J. and Huang, Z. (1997). "Mining the knowledge mine." in: Sattar A., Ed., *Advanced Topics in Artificial Intelligence (Lecture Notes in Artificial Intelligence)*, Springer, Berlin, Heidelberg, 340-348.

World Health Statistics (2011). *WHO Library Cataloging-in-Publication Data*.

Yamanishi, K., Takeuchi, J.I., Williams, G., and Milne, P. (2004). "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms." *Data Mining and Knowledge Discovery*, **8**(3), 275-300.

Yan, Y., Chen, M., Sadiq, S., and Shyu, M.L. (2017). "Efficient imbalanced multimedia concept retrieval by deep learning on spark clusters." *International Journal of Multimedia Data*

Engineering and Management, **8**(1), 1-20.

Yang, W.S. and Hwang, S.Y. (2006). "A process-mining framework for the detection of healthcare fraud and abuse." *Expert Systems with Applications*, **31**(1), 56-68.

Yang, W. and Su, Q. (2014). "Process mining for clinical pathway: literature review and future directions." *Proceedings of the IEEE 11th International Conference on Service Systems and Service Management*, 1-5.

Zhu, Q., Lin, L., Shyu, M.L., and Chen, S.C. (2011). "Effective supervised discretization for classification based on correlation maximization." *Proceedings of the IEEE International Conference on Information Reuse and Integration*, 390-395.