Health Care Manage Sci
DOI 10.1007/s10729-007-9045-4

A survey on statistical methods for health care fraud detection

Jing Li · Kuei-Ying Huang · Jionghua Jin · Jianjun Shi

Received: 29 May 2007 / Accepted: 11 December 2007

Abstract Fraud and abuse have led to significant additional expense in the health care system of the United States. This paper aims to provide a comprehensive survey of the statistical methods applied to health care fraud detection, with focuses on classifying fraudulent behaviors, identifying the major sources and characteristics of the data based on which fraud detection has been conducted, discussing the key steps in data preprocessing, as well as summarizing, categorizing, and comparing statistical fraud detection methods. Based on this survey, some discussion is provided about what has been lacking or under-addressed in the existing research, with the purpose of pinpointing some future research directions.

Keywords Fraud detection · Health care · Statistical methods

1 Introduction

Health care has become a major expenditure in the US since 1980. According to a report by the GAO (General Accounting Office) to Congress in 2004, annual health care expenditures were approaching two trillion dollars, which accounted for 15.3% of the GDP (Gross Domestic Product). The size of the health care sector and the enormous volume of money involved make it an attractive fraud target. Health care fraud, based on the definition of the NHCAA (National Health Care Anti-fraud Association), is an intentional deception or misrepresentation made by a person or an entity, with the knowledge that the deception could result in some kinds of unauthorized benefits to that person or entity. The NHCAA estimated conservatively that at least 3%, or more than $60 billion, of the US's annual health care expenditure was lost due to outright fraud. Other estimates by government and law enforcement agencies placed this loss as high as 10% or $170 billion [28]. Not only is the financial loss a great concern, fraud also severely hinders the US health care system from providing quality and safe care to legitimate patients. Therefore, effective fraud detection is important for improving the quality and reducing the cost of health care services.

In recent years, systems for processing electronic claims have been increasingly implemented to automatically perform audits and reviews of claims data. These systems are designed for identifying areas requiring special attention such as erroneous or incomplete data input, duplicate claims, and medically noncovered services. Although these systems may be used to detect certain types of fraud, their fraud detection capabilities are usually limited since the detection mainly relies on pre-defined simple rules specified by domain experts.

More sophisticated antifraud systems incorporating a wide array of statistical methods are needed and are being developed for effective fraud detection. The major advantages of these systems include: (1) automatic learning of fraud patterns from data; (2) specification of "fraud likelihood" for each case, so that efforts for investigating suspicious cases can be prioritized; and (3) identification of new types of fraud which were not previously documented.

J. Li (✉)
Department of Industrial Engineering, Arizona State University,
P.O. Box 875906, Tempe, AZ 85287-5906, USA
e-mail: jinglz@asu.edu

K.-Y. Huang · J. Jin · J. Shi
Department of Industrial and Operations Engineering,
University of Michigan,
1205 Beal Avenue,
Ann Arbor, MI 48109-2117, USA

Statistical fraud detection methods can be divided into two categories: supervised and unsupervised methods. Supervised methods require all cases in the training dataset to be labeled by domain experts. Unsupervised methods do not have this requirement and their objective is to find outliers in the cases. Examples of the statistical methods that have been applied to health care fraud detection include neural networks [7, 15, 17, 31, 34], decision trees [4, 41], association rules [38], Bayesian networks [30], and genetic algorithms [16, 40]. As a result of applying these methods, some fraud behaviors can now be detected, including home/hospital stay conflict, hospital stay with no associated physician inpatient visit, excessive lab/radiology services per client per day, X-ray duplicate billing, fragmented lab and X-ray procedures, lab/X-ray interpretation with no associated technical portion, and ambulance trips with no associated medical service [35].

It is known that the effectiveness of a statistical fraud detection method is affected by the extent to which the unique characteristics of health care data conform to the inherent assumptions of this method. Therefore, it is important to understand the strengths and limitations of each method when applied to health care data, and to identify what has been missing or under-addressed in the existing research. This requires a global view across individual research efforts. However, most existing papers focus on how a single method or a few methods combined is/are applied to a specific dataset for detecting a particular type of fraud. There lacks a comprehensive survey that summarizes the characteristics of the health care data, classifies fraudulent behaviors, discusses key steps in data preprocessing, and categorizes and compares various statistical methods. This paper aims to address these issues.

2 Classification of fraudulent behaviors

Three parties may be involved in the commission of health care fraud. They are (a) *service providers*, including doctors, hospitals, ambulance companies, and laboratories; (b) *insurance subscribers*, including patients and patients' employers; and (c) *insurance carriers*, who receive regular premiums from their subscribers and pay health care costs on behalf of their subscribers, including governmental health departments and private insurance companies. According to which party commits the fraud, fraud behaviors can be classified as follows [28, 45].

(a) Service providers' fraud:

- Billing services that are not actually performed;
- *Unbundling*, i.e., billing each stage of a procedure as if it were a separate treatment;
- *Upcoding*, i.e., billing more costly services than the one actually performed; for example, "DRG creep" is a popular type of upcoding fraud, which classifies patients' illness into the highest possible treatment category in order to claim more reimbursement;
- Performing medically unnecessary services solely for the purpose of generating insurance payments;
- Misrepresenting non-covered treatments as medically necessary covered treatments for the purpose of obtaining insurance payments; and
- Falsifying patients' diagnosis and/or treatment histories to justify tests, surgeries, or other procedures that are not medically necessary.

(b) Insurance subscribers' fraud:

- Falsifying records of employment/eligibility for obtaining a lower premium rate;
- Filing claims for medical services which are not actually received; and
- Using other persons' coverage or insurance card to illegally claim the insurance benefits.

(c) Insurance carriers' fraud:

- Falsifying reimbursements; and
- Falsifying benefit/service statements.

Among these three types of fraud, the one committed by service providers accounts for the greatest proportion of the total health care fraud and abuse. Although the vast majority of service providers are honest and ethical, the few dishonest ones may have various possible ways to commit fraud on a very broad scale, thus posing great damage to the health care system [28]. Some service providers' fraud, such as that involving medical transportation, surgeries, invasive testing, and certain drug therapies, even places patients at a high physical risk. Therefore, detection of service providers' fraud is the most urgent problem for improving the quality and safety of a health care system, and has attracted many researchers. In Fig. 1, we developed a bar chart showing the percentage of papers in the literature on each type of fraud. It can be seen that a

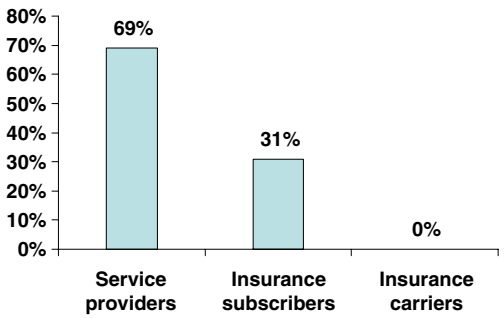

Fig. 1 Percentages of papers on detecting three types of fraud

majority of research efforts have been devoted to detecting service providers' fraud, while the efforts on the other two types of fraud are limited.

In addition to the aforementioned types of fraud, which are committed by a single party, there is another newly emerging type of fraud called conspiracy, which involves more than one party. A typical conspiratorial fraud scenario is that a patient colludes with his physician, fabricating medical service and transition records to deceive the insurance company to whom he subscribes. Not much research has been found on detecting conspiratorial fraud. However, it is reasonable to believe that such research can be very rewarding due to the complexity, increasing popularity, and severe consequences of conspiratorial fraud.

3 Health care data: sources, characteristics, and preprocessing

Raw data for health care fraud detection come mostly from insurance carriers (this also partly explains why little research exists to detect insurance carriers' fraud), including governmental health departments and private insurance companies. Major governmental health departments that have been reported in the literature include the US Health Care Financing Administration (HCFA) [34, 13], the Bureau of National Health Insurance (NHI) in Taiwan [6, 21, 39, 43, 45], and the Health Insurance Commission (HIC) in Australia [16, 17, 20, 41, 42]. The data from private insurance companies have also been used by several researchers [27, 31]. Figure 2 shows the percentages of papers that use the data from each major source.

The raw data, no matter which source they come from, are mostly insurance claims. An insurance claim involves the participation of an insurance subscriber and a service provider. The claim data have two characteristics. First, they contain a rich amount of attributes to describe the behaviors of the involved service providers and insurance subscribers, allowing for detection of the types of fraud committed by these two parties. Second, each claim usually contains unique identifiers for the involved service provider and insurance subscriber, respectively. By using the unique identifiers to link different claims, it is possible to obtain a global view of a service provider's behaviors over time and across different insurance subscribers, and also a global view of an insurance subscriber's behaviors over time and across different service providers. The global views help significantly in identifying the fraud committed by service providers and by insurance subscribers.

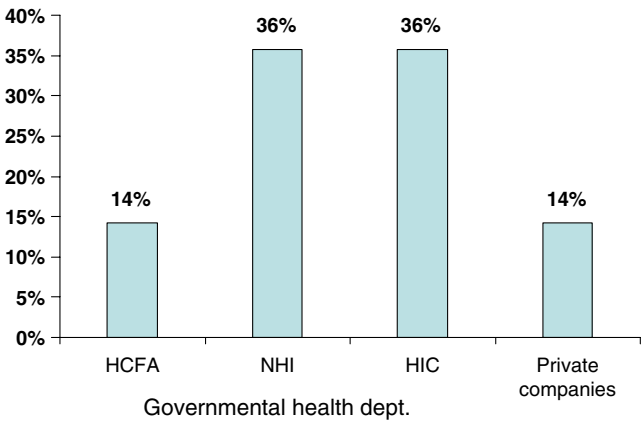

Fig. 2 Percentages of papers using the data from four major sources

Raw data must be processed into a form that is acceptable by the statistical method, which is called preprocessing. Preprocessing performs a large amount of work, taking roughly 80% of the total time in fraud detection. While it has been well accepted that preprocessing is a very challenging and time-consuming task, only two papers [25, 35] have been found to dedicatedly address this issue. By combining the important considerations in these two papers, and also extracting useful information from other papers in which preprocessing is not the focus but just briefly mentioned, we developed a flow chart of the key steps in data preprocessing, as shown in Fig. 3. Details for each step are discussed as follows.

3.1 Goal setting

The objective of this step is to identify and prioritize the types of fraud on which detection should be focused. This step needs the input and knowledge from both domain experts and statistical analysts. Domain experts have a good sense about which types of fraud happen the most frequently or incur the most financial loss, and thus set these types of fraud as the top priorities for detection. Statistical analysts provide expert opinions from the data analysis point of view to judge if the data used for fraud detection can satisfy the constraints for effective statistical analysis such as sample size requirements and data quality. While it is reasonable to believe that "goal setting" must have been carried out in almost every documented research project, a very limited amount of literature details the process and outcomes of this step. An exception is the paper by Sokol et al. (2006), in which they reported their goal setting discussions with the representatives from the HCFA and the Office of the Inspector General in a project for detecting service providers' fraud from the HCFA claim data. The discussions led to six focused types of fraud to be identified, including Ambulance Services, Skilled Nursing Facilities, Laboratory Services, Psychiatric Services, Home Health Services, and new or expanded benefits under the Balanced Budget Act of 1997.

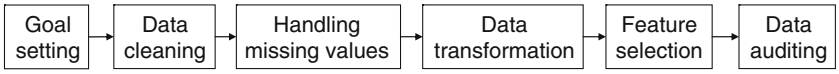

Fig. 3 Data preprocessing steps

3.2 Data cleaning

One prominent issue in health care data is inconsistent representations of the same concept. For example, Sokol et al. (2006) reported that different digit formats were used to represent the same physician's unique identification number in the HCFA data, making automatic computer algorithms fail to link the claims involving the same physician. Similarly, Lin and Haug [25] reported that for the same disease, such as abdominal rebound pain, a number of representations may exist in the dataset including text description ("abdominal rebound pain"), a diagnostic code, or simply a "YES" for indicating the pain's existence. Other critical data quality issues include outliers due to entry errors. Although data cleaning is believed to be an extremely important step in preprocessing, it is not discussed in detail in the literature. Most work on data cleaning is done manually through collaborative efforts between data analysts and domain experts. It is difficult to develop universal automatic computer algorithms because each dataset may have a unique problem in data quality.

3.3 Handling missing values

Missing values are common in health care data. Some data elements are not collected due to omission, irrelevance, excess risk, or inapplicability in a specific clinical context. Most statistical methods require a complete set of data elements. Even for some methods that accept missing values, such as decision trees and Bayesian networks, the performance of these methods will improve if we are able to appropriately handle the missing values rather than blindly remove all cases with missing values from the dataset, especially when the data elements are not missing in a random manner [26]. One effective method that copes with missing values is called imputation, which has been widely used in clinical data analysis, although not found in the literature of health care fraud detection. The concept of "imputation" is briefly introduced as follows. More details can be found in the book by Little and Rubin [26].

Imputation is the substitution of some value for a missing data element. The two most commonly used imputation methods are *hot-deck imputation* and *regression imputation*. Hot-deck imputation fills in the missing values in an incomplete case using values from the most similar (in terms of other variables without missing values) but complete cases of the same dataset. In regression imputation, a regression model is fitted for each variable with missing values, with other variables without missing values as covariates. Then, the missing values are substituted by the predicted values from the regression. In addition, instead of filling in a single value for each missing value, *multiple imputation* replaces each missing value with a set of plausible values, with each plausible value obtained from a single imputation method such as hot-deck imputation or regression imputation. Then, the multiply imputed datasets are analyzed by the same statistical method, and the results are combined. Multiple imputation is superior to single imputations because the results from multiple imputation can better reflect the uncertainty due to missing values.

3.4 Data transformation

This step concerns transformation of the raw claim data into new and more relevant views. Because the database structure of the raw claim data is usually designed to support an online transaction system, the database, containing multiple rows and columns, has each row represent a transaction and each column represent an attribute of the transaction. Given a focused type of fraud to be detected, a new dataset in the format of a "flattened table" must be created from the raw data. A flattened table format of data is required by most statistical methods, in which each row represents a case for training and/or testing a statistical model, and each column contains the values for an attribute across cases. For example, if one is focusing on detecting the fraud committed by insurance subscribers such as patients, each row in the new dataset should correspond to a different patient, and each column should correspond to an attribute describing either a demographic characteristic (e.g., sex and date of birth) or one aspect of a patient's usage of the health care system (e.g., the total number of top service codes and the average charge per service, for a certain time period). Figure 4 gives examples of the raw data and the new transformed dataset for detecting patients' fraud.

Note that in the new dataset, the attributes describing a patient's usage of the health care system contain the information for a specific aggregated time period such as a week, a month, or a year. In other words, a new dataset can only be created after a specific aggregated time period of interest has been defined. Considering that effective fraud detection may need to analyze the data from aggregated time periods with various lengths, a shorter aggregated time period (e.g., a week) is commonly used. The new datasets defined on a short time period (e.g.,

Raw claim data (transaction-based)

Time of claim	Patient ID	Patient sex	…	Physician ID	Physician location	Specialty	…	Service performed (code)	Charge of service	…
t1	a001	F	…	b005	AZ	GI	…	7081	$$$	…
t2	a002	M	…	b002	AZ	OB/GYN	…	4925	$$	…
t3	a003	M	…	b007	MI	OB/GYN	…	2164	$$$	…
t4	a001	F	…	b005	AZ	GI	…	3380	$$$	…
…	…	…	…	…	…		…	…	…	…
tn	a002	M	…	b007	MI	GP	…	9975	$$	…

New dataset (patient-based flattened table)

Demographic info. Aggregated usage of health care system for a certain time period

Patient ID	Sex	...	Total number of top service codes	Total number of physicians seen	Average charge per service	…
a001	F	...	1	2	$$$	...
a002	M	...	1	2	$$	...
...	...	...	...	...	...	...

Fig. 4 Examples of the raw claim data and a new transformed dataset for detecting fraud committed by patients

weekly datasets) can be further aggregated to create a dataset with any desired longer time period (e.g., monthly or yearly datasets), if needed.

Also, the process of transforming raw data into new and more relevant views is a continuous process. After statistical methods are applied to the new dataset, the validity and properness of the transformation can be tested, leading to refinement or revision of the original transformation. Through some interactions, we may obtain a better understanding of which time period is the best for fraud detection.

3.5 Feature selection

This step aims to define new features out of the original attributes, to maximize the discrimination power of the statistical method in separating fraudulent and legitimate cases. In this sense, the previous step, data transformation, can also be considered as a feature selection step or can more precisely be called a "coarse" feature selection step because it focuses on how to transform the raw data into a new and more relevant dataset, which, however, usually includes an excessive number of features. The number of these features needs to be further reduced, i.e., only those with certain discrimination power will be kept for statistical analysis, which will be discussed in this section as a "fine" feature selection step.

Discriminating features can be identified manually by domain experts or automatically by some statistical machine learning algorithms. In the majority of papers in the literature, the features are identified by domain experts. These papers usually only report the outcomes of the feature selection (e.g., how many features are selected and some examples of what they are) without detailing much the feature selection process and the associated methods. Manual feature selection can also benefit from statistical checking of each selected feature's relevance, leading to further refinement of the feature set. In the statistics literature of classification, feature selection is an important topic and various generic algorithms have been developed. Interested readers can refer to a review paper on this subject by Dash and Liu [10]. However, little of the literature in health care fraud detection has detailed if/how these generic feature selection algorithms were implemented. An exception is the work by Ortega et al. [31], in which they reported their feature selection procedure. First, domain experts defined a preliminary set of features; second, correlation checks were performed to delete redundant features; and third, the discriminating power of each feature was tested and only those features with discriminating power above a certain threshold were kept.

The number of features selected for fraud detection ranges from several to fifty. Figure 5 provides a bar chart that shows the percentages of papers in which specific numbers of features are used. It can be seen that the most commonly used numbers of features are between 10 and 30. Use of fewer than 10 features is rare, with papers found on detecting the fraud committed by laboratories. Use of more than 30 features is also rare, with papers found on unsupervised fraud detection.

Although we understand that it would be of great interest to summarize what features have been used in the literature, very few papers have released such details due to legal issues or privacy protection. One exception is the paper by Yamanishi et al. [42] in which they listed features for detecting the fraud committed by medical test laboratories.

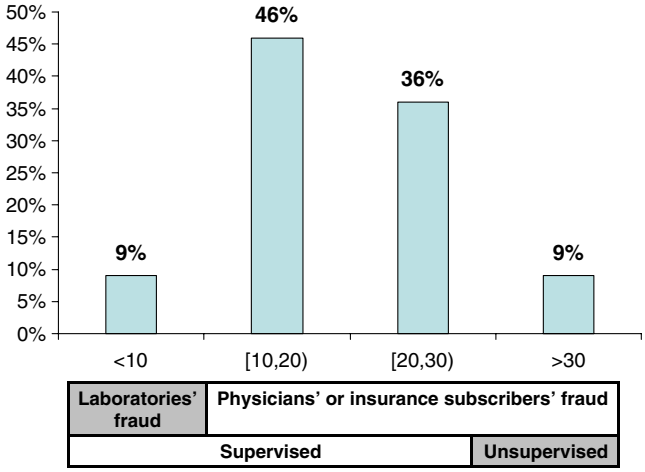

Fig. 5 Percentages of papers in which certain numbers of features are used

These features include the percentage distributions over several test categories (chemical, microbiology, and immunology), the number of different patients dealt with, and the frequency of tests performed. In addition, the paper by Major and Riedinger [27] classified potentially useful features into five categories, capturing the major aspects of service providers' profiles. These categories of features are shown in Table 1.

Automatic computer algorithms for feature selection were developed by a group of researchers focusing on the detection of service providers' fraud from the claim data in the Bureau of National Health Insurance (NHI) in Taiwan. Their algorithms can only be applied to the service providers whose practices, if legitimate, are supposed to follow well-defined standard clinical pathways. The concept of clinical pathways was initially developed in the early 1990s. A clinical pathway is a flow chart that sequences the necessary medical care activities (e.g. activities involved in diagnosis and treatment) given to a patient or a patient group with a certain disease. For example, the clinical pathway of cholecystectomy [33] starts with preadmission testing and anesthesia consultation, followed by several assessments, surgeries, and physician orders, and ends with a follow-up visit at the surgeon's office. The purpose of implementing clinical pathways in medical practice is to have medical professionals perform health care services in the right sequence, leading to the best practice (i.e., without rework and resource waste). The concept of clinical pathways shows great potential in detecting service providers' fraud. That is, if an honest service provider always prescribes the standard clinical pathway to his patients with a certain disease, then a service provider prescribing an irregular pathway, which might include different-from-standard care activities or a different-from-standard execution sequence of the standard care activities, is suspicious. For example, although a single ambulant visit is normal, repetitive visits are problematic; also, non-invasive tests are normally performed before a more complex invasive test, and a reverse order is problematic.

In clinical pathway based research, not only is the feature selection step different from other research, but also the data transformation step. An integrated view of these two steps will be briefly discussed as follows.

In the new dataset created from the raw claim data, each case corresponds to a patient, and includes the medical care activities, as well as their execution sequence, in the diagnosis and treatment of this patient's disease. Each case is converted to a temporal graph, as shown in Fig. 6, in which an arc with an arrow links two temporally successive activities. Furthermore, an algorithm, rooted from association rule and sequential pattern induction, is used to break down the temporal graph into a set of subgraphs, each preserving part of the relationship in the original graph. By pooling the subgraphs of all available cases together, the algorithm further identifies "frequent" subgraphs which are those shared by at least s% (a predefined threshold) of all the cases. In practice, the number of frequent subgraphs is huge considering that disease diagnosis and treatment is usually a complex process. Thus, a feature selection algorithm was developed to exclude irrelevant or redundant frequent subgraphs, and the remaining subgraphs are used as features in the subsequent supervised learning. In a case study with pelvic inflammatory disease, a significant

Table 1 Categories of features commonly used in detecting service providers' fraud [27]

Categories of Features	Description
Financial	The flow of dollars
Medical logic	Whether a medical situation would normally happen
Abuse	Frequency of treatments
Logistics	The place, time, and sequences of activities
Identification	How providers present them

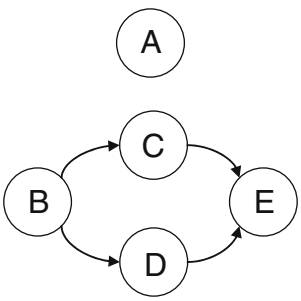

Fig. 6 Temporal graph of a case with five care activities

dimension reduction is achieved by applying the feature selection algorithm, which downsizes the original 30,701 frequent subgraphs (s%=2%) into 3,120 features.

3.6 Data auditing

This step performs basic statistical analysis and visualization to become familiar with the data, including, but not limited to, checking probability distributions and understanding feature relationships (i.e., linear or non-linear relationships, and clustering patterns).

4 Statistical modeling for fraud detection

Statistical methods in health care fraud detection are generally divided into two classes, supervised and unsupervised methods, depending on the availability of labels in the training dataset. This section will first review the popularly used methods in each class, and then discuss the integration of the two classes.

4.1 Supervised methods

Several supervised methods have been used in health care fraud detection, including neural networks (NNs), decision trees, fuzzy logic, and Bayesian networks. A software survey of these methods was given by Abbott (1998). The two most popular methods, i.e., NNs and decision trees, as evidenced in Fig. 7, will be discussed in detail in this section.

It is also noteworthy that in applying supervised methods to health care fraud detection, there an increasing trend to combine several supervised methods in order to improve classification performance. For example, Ormerod et al. [30] proposed to detect fraud by a Bayesian network (BN), whose weights were refined by a rule generator called Suspicion Building Tool (SBT). He et al. [16] proposed the use of a k-nearest neighbor algorithm whose distance metric was optimized by a genetic algorithm in detecting

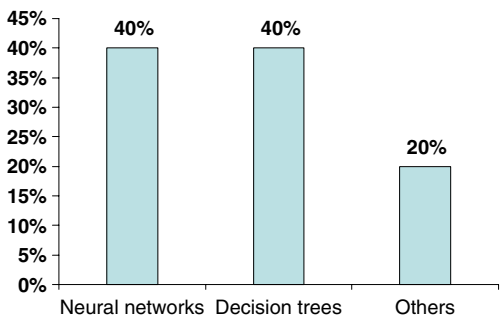

Fig. 7 Percentages of papers on different supervised methods

two types of fraud: inappropriate practice of service providers and "doctor-shoppers." A model that combined fuzzy sets theory and a Bayesian classifier was designed to detect suspicious claims in the NHI of Taiwan [6]. Viveros et al. [38] recommended association rules and a neural segmentation algorithm for fraud detection.

4.1.1 Neural networks

NNs have been used extensively in detecting health care fraud due to their capability of handling complex data structures and non-linear variable relationships [7, 15, 17, 31]. Hall [15] used an NN to detect service providers' fraud based on discriminating features identified by domain experts. Cooper [7] used an NN to identify fraudulent medical claims submitted to a Chilean health insurance company. One of the common concerns with NNs is overfitting, which produces a relatively small error on the training dataset but a much larger error when new data is presented to the network. Overfitting is especially prominent with skewed data [32] such as health care claims, which have many more legitimate than fraudulent cases.

To address the overfitting problem, He et al. [17] proposed, in their work on detecting service providers' fraud, to add a weight delay term, $\alpha \sum w_i^2$, to the error function for training an NN with 28 input features, 15 neurons in the hidden layer, and 4 output classes (i.e., different likelihoods of fraud), in which α is pre-defined and w_i are all the weights trained in the NN. This strategy intended to achieve a tradeoff between the training error and complexity of the NN. It was shown in the paper that inclusion of the weight decay (α=0.01) improved generalization of the NN from the training to the test dataset with a better agreement rate on the test dataset.

Overfitting was also addressed by Ortega et al. [31], who developed a fraud detection system for a Chilean health insurance company. Their system contains sub-systems to detect the fraud committed by service providers, insurance subscribers, and the subscribers' employers. This system implemented a technique known as "early stopping" to avoid overfitting. This technique uses two different datasets in training an NN: one is used to update the weights and biases, and the other is used to stop training when the network begins to overfit the data.

In addition to overfitting, another challenge Ortega et al. [31] faced is small training sample size with a large number of features (e.g., 424 insurance subscribers with 25 features for fraud detection). This led to a high variance (>8%) between different runs of a standard two-layer NN. In order to decrease the variance, they applied a committee of 10 independently trained NNs to the data instead of just a single NN, and averaged the outputs, thus successfully decreasing the variance to 1.8%.

Table 2 Comparison of neural networks and decision trees

Parameter	Neural network	Decision tree
Strengths	Ability to handle complex data structure especially non-linear relationship	Good interpretability of results Ability to generate rules from tree
	High tolerance to noisy data	Ability to handle missing values
Limitations	"Black box" method, hard to understand how it works	*Too many rules generated for large dimensional datasets, decreasing interpretability
	*Few effective approaches to handle overfitting (i.e., the approaches addressing overfitting require large sample size or long training time)	Few adjustable parameters available
	Too many parameters to tune, no generic guideline on how to tune (case by case, based on trial and error)	
	*Large prediction variance among different runs of a single NN when sample size is small	

4.1.2 Decision trees

Decision trees have also been widely used for health care fraud detection [4, 41, 45]. Yang [45] used the C4.5 algorithm to train decision trees for identifying service providers' fraud for the NHI in Taiwan. As an extension to C4.5, the C5.0 algorithm offers several advanced mechanisms, thus having been adopted by more researchers. Bonchi et al. [4] used the C5.0 algorithm to train decision tree classifiers for planning audit strategies in fraud detection. They constructed several classifiers, each of which was driven by a different policy in audit planning. These policies included minimizing false positives (i.e., minimizing wasteful costs for unnecessary fraud investigations), minimizing false negatives (i.e., maximizing detectability), and achieving prescribed tradeoffs between false positives and false negatives. They also implemented adaptive boosting in training the classifiers in order to reduce misclassification errors. Like all ensemble methods, adaptive boosting works by producing a set of classifiers and voting on them to classify cases. Its uniqueness is that the various classifiers are constructed sequentially, with each focused on those training cases that have been misclassified by previous classifiers.

The C5.0 algorithm was also used by William and Huang [41] in detecting insurance subscribers' fraud for the Health Insurance Commission (HIC) of Australia. Due to the huge amount of data (40,000 insurance subscribers), they faced the challenge that an overly complex decision tree with thousands of rules was generated, which made it difficult to interpret. To address this challenge, they proposed a three-step so-called "divide and conquer" procedure. First, a clustering algorithm was applied to divide all insurance subscribers' profiles into groups. Second, a decision tree was built for each group and then converted into a set of rules. An example of a rule is: "if *Age* is between 18 and 25 and *Weeks*>=10, then cluster=4." Finally, each rule was evaluated by establishing a mapping from the rule to a measure of its significance using simple summary statistics such as the average number of claims made in that cluster and the average size of the claims; then extremes (either defined by experts or compared with the overall average) would be signaled for further investigation. Through this procedure, the number of rules was significantly reduced to 280 summed over ten clusters.

In Table 2, we summarize the strengths and limitations of neural networks and decision trees. In the two cells corresponding to "Limitations," the bullets starting with an asterisk (*) are those limitations that have been addressed in the health care fraud detection literature. Note that a majority of limitations have not been addressed, leaving plenty of room for further research.

4.2 Unsupervised methods

Most unsupervised methods have been combined with supervised methods in health care fraud detection, which will be reviewed in Section 4.3. Research on employing unsupervised methods alone is limited to the following two papers.

Electronic Fraud Detection (EFD) [27] is an expert system assisting in detecting service providers' fraud, and includes several steps. First, discriminating features (called "behavioral heuristics") are defined by experts. Then, the information gain for a provider is computed as $\int \log\left(f_p(\mathbf{x})/f_A(\mathbf{x})\right) f_p(\mathbf{x})\mathrm{d}\mathbf{x}$, where $f_p(x)$ and $f_A(x)$ are probability density functions of the features for the provider and the aggregate peer group, respectively. The information gain measures how different the distribution of the provider is from that of all the peers taken together. Computation of the information gain can be greatly simplified under the assumption of independently normally distributed features. Finally, the providers are plotted as points on a 2-D space with one axis representing the information gain and the other representing total dollars paid to the providers. The frontier points on the plot are highlighted and the corresponding providers are

considered to be suspicious. Note that the advantage of using the frontier points, over maximizing the product of the information gain and total dollars paid, is that a multivariate maximization can be achieved in a situation where no weights exist *a priori* to equate dollars and information bits in a linear or log-linear function.

Another unsupervised method is called SmartSifter [42], used to detect outliers in the pathology dataset provided by the HIC of Australia. SmartSifter uses a probabilistic model to represent the underlying data-generating mechanism. In the probabilistic model, a histogram is used to represent the probability distribution of categorical variables; for each bin of the histogram, a finite mixture model is used to represent the probability distribution of continuous variables. When a new case is coming, SmartSifter updates the probabilistic model by employing an SDLE (Sequentially Discounting Laplace Estimation) and an SDEM (Sequentially Discounting Expectation and Maximizing) algorithm to learn the probability distributions of the categorical and continuous variables, respectively. A score is given to this new case, measuring how much the probabilistic model has changed since the last update. A high score indicates that this new case may be an outlier.

4.3 Hybrid methods of combining unsupervised and supervised methods

Hybrid methods, combining supervised and unsupervised methods, have been developed by a number of researchers. When an unsupervised method is followed by a supervised method, the objective is usually to discover knowledge in a hierarchical way. Williams and Huang [41] integrated clustering algorithms and decision trees to detect insurance subscribers' fraud. Their three-step "divide and conquer" procedure was detailed in Section 4.1 under the subtitle "Decision trees." This procedure has been followed in some other unsupervised-supervised methods, but different algorithms were used in each step. For example, Williams [40] recommended the use of a genetic algorithm to generate rules such that extra freedom can be achieved, e.g., specification/revision of the rules by domain experts and the evolving of rules by statistical learning.

Unsupervised methods have been used to guide the selection of the number of classes in supervised methods. He et al. [17] reported the use of an NN to classify service providers' profiles. The training data were originally divided into four classes representing different likelihoods of fraud. Because of unsatisfactory classification results, SOM (Self-Organizing Map) was adopted to find inherent clusters embedded in the data. SOM suggested that combining the original four classes into two was a more appropriate way to represent the data.

4.4 Performance evaluation of supervised methods

This section summarizes and compares different ways of evaluating the performance of binary classifiers, in which cases are labeled as either fraudulent or legitimate. Although efforts to evaluate k-class classifiers ($k>2$) and unsupervised learning have been seen [17, 42], they only occupy a very small portion of the papers in health care fraud detection, and the performance evaluation procedures in these papers are not detailed.

In the performance evaluation of a binary classifier, an important initial step is to construct a confusion matrix based on the testing dataset, as shown in Fig. 8. The cell labeled with "TP" records the number of actual fraudulent cases that are correctly predicted by the classifier; other cells, FP, FN, and TN, are defined in similar ways.

There are in general two categories of methods for evaluating the performance of a classifier based on the confusion matrix. If the costs of FP and FN can be explicitly specified, the performance should be evaluated based on misclassification costs; otherwise, it is generally evaluated based on misclassification errors. These two categories of methods are called error-based and cost-based methods, respectively.

Among error-based methods, an ROC curve is commonly used [6, 31, 43], which plots the true positive against the false positive rates at different decision making thresholds of a classifier. The curve can be used to select a decision threshold that leads to a satisfactory sensitivity (i.e., true positive rate) and specificity (i.e., 1–false positive rate). In addition, the area under the curve, called AUC (Area Under Curve), indicates the discriminating power of the classifier. Specially, if a classifier performs perfectly, AUC=1; if it performs randomly, AUC=0.5; the closer the AUC to 1, the higher the discriminating power of the classifier. There are other popular measures associated with ROC curves, such as CXE (Cross Entropy) and AMOC (Activity Monitoring Operating Characteristics), the uses of which, however, have not been seen in health care fraud detection. Interested readers can refer to the papers by Viaene et al. [37] and Fawcett and Provost [11]. There are

		Actual classes	
		Fraudulent (+)	Legitimate (-)
Predicted classes by classifier	Fraudulent	True positive (TP)	False positive (FP)
	Legitimate	False negative (FN)	True negative (TN)

Fig. 8 Confusion matrix

also some error-based methods which are not based on ROC curves. For example, He et al. [17] used the agreement rate (i.e., $(TP + TN)/(TP + TN + FP + FN)$) to evaluate classifiers.

To apply cost-based performance evaluation methods, the availability of a cost model is a prerequisite. A cost model usually has two sets of parameters, the costs of cases, $C_C(i)$, and the costs of the efforts (e.g., manpower) involved in investigating the cases, $C_I(i)$, where i indexes the cases in the dataset. There are two ways to specify the values for $C_C(i)$ and $C_I(i)$. If the $C_C(i)$ and $C_I(i)$ for different i's are similar, the corresponding average costs can be used for all cases, i.e., $C_C(i) = \overline{C}_C$ and $C_I(i) = \overline{C}_I$. This specification is adopted by Phua et al. [32]. For a particular classifier A, after a confusion matrix is constructed based on a testing dataset, the costs for TP, TN, FP, FN can be computed as functions of $\overline{C}_C$ and $\overline{C}_I$, i.e., $C_{TP} = TP \times \overline{C}_I$, $C_{TN} = TN \times \overline{C}_C$, $C_{FP} = FP \times (\overline{C}_I + \overline{C}_C)$, and $C_{FN} = FN \times \overline{C}_C$. Then, the cost saved by using classifier A, compared with taking no actions, can be computed, i.e., $C_{saved} = (TP + TN + FP + FN) \times \overline{C}_C - (C_{TP} + C_{TN} + C_{FP} + C_{FN})$; and the cost saved by using a perfect classifier which correctly predicts all cases, compared with taking no actions, can also be computed, i.e., $C^0_{saved} = (TP + FN) \times \overline{C}_I + (FP + TN) \times \overline{C}_C - (C_{TP} + C_{TN} + C_{FP} + C_{FN})$. Finally, an index, $p_{saved} = (C_{saved}/C^0_{saved})$, can be obtained, which measures how close the performance of classifier A is to that of the perfect classifier.

The other way to specify the values for $C_C(i)$ and $C_I(i)$ is to use different values for different cases [4]. This happens when $C_C(i)$ and $C_I(i)$ vary significantly across different cases. p_{saved} can be calculated in a similar way except that individual costs are used for each case rather than the same cost (i.e., the average cost) for all cases. As a result, the classifier's performance will not be sensitive to the cases associated with low costs, that is, even misclassification of these cases will not significantly degrade the classifier's performance.

Aside from misclassification errors and costs, there are other criteria to evaluate classifiers, including speed, scalability, and robustness, which, however, have not been found in the health care fraud detection literature. Interested readers can refer to the paper by Ghosh and Reilly [14] for a discussion on how to address these criteria in evaluating classifiers.

4.5 Discussion on the challenges and needs in future research

Case labeling Supervised methods require a training dataset in which all cases have been labeled by domain experts. The labels can be "noisy" due to human subjectiveness or bias. This issue, however, has not been addressed much in the literature. He et al. [17] suggested the use of two additional intermediate classes between strictly fraudulent and legitimate to label cases. However, this created a problem when it came to determining a case's class using its probabilistic likelihood score from the NN output. Specifically, when the case has high scores for more than one class, it is highly likely that this case will be misclassified. To address this problem, they proposed to use the difference between the highest and second highest score as the confidence level for the classification. Cases with low confidence levels are returned to experts for re-examination.

It is doubtless that the most effective way of handling labeling noise is to remove all the cases with unconfident labels from the training dataset. However, this will reduce the sample size, creating other problems in classification. Research has been conducted on how to utilize unlabeled cases to improve classification performance, including methods based on the EM (Expectation-Maximization; [29]), SVM (Support Vector Machine; [2]), and co-training algorithms [3]. There are challenges in directly applying these generic methods to health care fraud detection, because some assumptions required by the methods are not satisfied by health care data, such as the assumption of a mixture model structure required by the EM-based method, and the availability of two independent and sufficiently redundant views of the data required by the co-training-based method. Yang [45] made efforts to implement the co-training-based method in health care fraud detection. While the results based on their test dataset seem promising, the theoretical basis needs to be studied. Work aside from this effort is rarely seen, but research along this line is highly beneficial, as it will not only provide an effective way of handling labeling noise, but also expedite detection of fraud and reduce investigation and labeling costs.

Adaptive fraud detection Health care data are dynamic in nature because both fraudulent and legitimate patterns may change over time. Once a type of fraud has become well-understood and thus can be effectively detected, criminals will adjust their strategies to make the existing fraud detection methods inadequate. Legitimate patterns also shift as insurance companies constantly tune their plans in order to attract subscribers. Therefore, a fraud detection system should have self-learning and self-evolving capabilities to adapt to the ever-changing fraudulent and legitimate patterns. To achieve this requires model parameters to be online re-trained or progressively tuned as new data become available. SmartSifter, an unsupervised outlier detection system by Yamanishi et al. [42], implemented two algorithms (for categorical and continuous variables, respectively) to learn the probability distribution of multivariate data. This distribution was updated with each new

datum, using a decay factor to discount the out-of-date data. A score was computed by comparing the distributions before and after the update, and the datum was identified as an outlier if a high score was generated. Compared with unsupervised methods, there are more challenges in developing adaptive algorithms for supervised methods, as legitimate and fraudulent patterns may drift in different ways and the algorithms must take both driftings into consideration in an integrated manner. Several other issues in adaptive fraud detection are also worthy of further study. For example, sophisticated algorithms need to be improved to reduce training time without sacrificing fraud detection effectiveness. Additionally, algorithms that can discover reliable and robust knowledge/relationships with a small sample size are always desirable, as these algorithms can help capture the change in fraud patterns as soon as it takes place.

Integration of individual methods While most research papers that have been reviewed focus on a single method tailored to a fraud detection problem, more recent research suggests a trend of integrating these methods [6, 16, 30]. The integration has the potential to overcome the limitations of individual methods and take advantage of their strengths. Shapiro [34] provided an overview of merging NNs, FL (fuzzy logic) or FISs (fuzzy inference systems), and GAs (genetic algorithms) in insurance-related applications. This paper first highlights the advantages and disadvantages of each method, as shown in Table 3, and then discusses several potential merging options.

In general, it is reasonable to believe that the integration of different methods can provide a better solution than a single method. Challenges still remain for future research on how to integrate different methods in a mutually complementary fashion, such as developing effective ways to handle conflicting assumptions, achieving high compatibility, and facilitating interface designs.

Causation vs. discrimination In a broad sense, discrimination of fraudulent and legitimate cases, which is the objective of most existing methods, is only the intermediate goal of fraud detection. The ultimate goal is to identify and eliminate the causes of fraud so that fraud can be prevented in the future. Although techniques of causal modeling and inference on large datasets have been developed [18, 22, 36] and applied to several domains including genetics [12], ecology [5], social science [9], and process control [23, 24], few have explored the causal relationship between various behavioral measures and the commission of fraud. The lack of research in this area results in limited success in developing antifraud strategies. In order to develop effective detection systems to cope with increasingly sophisticated fraud, the efforts of applying causal modeling and inference techniques to health care fraud detection and prevention will be highly rewarding in practice.

5 Conclusion

The development of a safe, high-quality, and cost-effective health care system requires effective ways to detect fraud. The applications of statistical methods to health care fraud detection are rewarding but highly challenging. This paper is the first to provide a comprehensive survey of published research results in health care fraud detection. Efforts were made to classify the fraudulent behaviors, identify the sources and characteristics of health care data, provide key steps in data preprocessing, and summarize and compare existing statistical methods.

The overall objective of the research in this area is to develop fraud detection methods and algorithms that are *scalable*, *accurate,* and *fast*. *Scalable* refers to the capability of handling the immense volume of health care data. *Accurate* refers to low errors/costs induced by false alarms and misdetection of fraud. *Fast* refers to the capability of catching frauds in real time before they incur severe loss and damage. While the existing research has provided various effective tools to achieve this objective to some degree, there still remain many research challenges worthy of further study. Examples of important and open issues include data-driven (rather than human-centered) feature selection, robust classification algorithms using noise-containing labeled data, integration of individual methods in a mutually complementary fashion for better fraud detection, causal discovery for fraud prevention, and adaptive fraud detection for identifying newly emerging fraud.

Table 3 Advantages and disadvantages of NNs, FL, and GAs [34]

Parameter	Advantage	Disadvantage
NNs	Adaptation, learning, approximation	Slow convergence speed, “black box” data processing structure
FL	Approximate reasoning	Difficult to tune, lacks effective learning capability
Gas	Systematic random search, derivative-free optimization	Difficult to tune, no convergence criterion

Acknowledgement We would like to thank the reviewers for the valuable comments that helped us significantly improve the paper.

References

1. Abbott DW, Matkovsky IP, Elder JF (1998) An evaluation of high-end data mining tools for fraud detection. In Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA
2. Bennett K, Demiriz A (1998) Semi-supervised support vector machines. Adv Neural Inf Process Syst 12:368–374
3. Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training. In Proceedings of the 11th Annual Conference on Computational Learning Theory
4. Bonchi F, Giannotti F, Mainetto G, Pedreschi D (1999) A classification-based methodology for planning auditing strategies in fraud detection. In Proceedings of SIGKDD99, 175–184
5. Borsuk ME, Stow CA, Reckhow KH (2004) A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. Ecol Model 173:219–239
6. Chan CL, Lan CH (2001) A data mining technique combining fuzzy sets theory and Bayesian classifier—an application of auditing the health insurance fee. In Proceedings of the International Conference on Artificial Intelligence, 402–408
7. Cooper C (2003) Turning information into action. Computer Associates: The Software That Manages eBusiness, Report, available at http://www.ca.com
8. Cox E (1995) A fuzzy system for detecting anomalous behaviors in healthcare provider claims. In: Goonatilake S, Treleaven P (eds) Intelligent systems for finance and business. Wiley, New York, pp 111–134
9. Dai H, Korb KB, Wallace CS, Wu X (1997) A study of casual discovery with weak links and small samples. In Proceeding of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI), San Francisco, CA, pp 1304–1309
10. Dash M, Liu H (1997) Feature selection for classification. IDA 1:131–156
11. Fawcett T, Provost F (1999) Activity monitoring: noticing interesting changes in behavior. In Proceedings of SIGKDD99, 53–62
12. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. J Comput Biol 7:601–620
13. GAO (1996) Health Care Fraud: Information-Sharing Proposals to Improve Enforcement Effects. Report of United States General Accounting Office
14. Ghosh S, Reilly D (1994) Credit card fraud detection with a neural network. Proceedings of 27th Hawaii International Conference on Systems Science 3:621–630
15. Hall C (1996) Intelligent data mining at IBM: new products and applications. Intell Softw Strateg 7(5):1–11
16. He H, Hawkins S, Graco W, Yao X (2000) Application of Genetic Algorithms and k-Nearest Neighbour method in real world medical fraud detection problem. Journal of Advanced Computational Intelligence and Intelligent Informatics 4(2):130–137
17. He H, Wang J, Graco W, Hawkins S (1997) Application of neural networks to detection of medical fraud. Expert Syst Appl 13:329–336
18. Heckerman D (1998) A tutorial on learning with Bayesian networks. In Learning in Graphical Models. Kluwer Academic, Boston, pp 301–354
19. Herb W, Tom M (1995) A scientific approach for detecting fraud. Best's Review 95(4):78–81
20. Hubick KT (1992) Artificial neural networks in Australia. Department of Industry, Technology and Commerce, CPN Publications, Canberra
21. Hwang SY, Wei CP, Yang WS (2003) Discovery of temporal patterns from process instances. Comp Ind 53:345–364
22. Lam W, Bacchus F (1993) Learning Bayesian belief networks: an approach based on the MDL principle. Comput Intell 10:269–293
23. Li J, Jin J, Shi J (2008) Causation-based T^2 Decomposition for Multivariate Process Monitoring and Diagnosis. Journal of Quality Technology, to appear in January 2008.
24. Li J, Shi J (2007) Knowledge Discovery from Observational Data for Process Control using Causal Bayesian Networks. IIE Transactions 39(6):681–690
25. Lin J-H, Haug PJ (2006) Data preparation framework for preprocessing clinical data in data mining, AMIA Symposium Proceedings 489–493
26. Little RJA, Rubin DB (2002) Statistical analysis with missing data. Wiley, New York
27. Major JA, Riedinger DR (2002) EFD: A hybrid knowledge/statistical-based system for the detection of fraud. The Journal of Risk and Insurance 69(3):309–324
28. NHCAA (2005) The Problem of Health Care Fraud: A serious and costly reality for all Americans, report of National Health Care Anti-Fraud Association (NHCAA)
29. Nigam K, McCalum A, Thrun S, Mitchell T (2000) Text classification from labeled and unlabeled documents using EM. Machine Learning 39:103–134
30. Ormerod T, Morley N, Ball L, Langley C, Spenser C (2003) Using ethnography to design a Mass Detection Tool (MDT) for the early discovery of insurance fraud. In Proceedings of the ACM CHI Conference
31. Ortega PA, Figueroa CJ, Ruz GA (2006) A medical claim fraud/abuse detection system based on data mining: a case study in Chile. In Proceedings of International Conference on Data Mining, Las Vegas, Nevada, USA
32. Phua C, Alahakoon D, Lee V (2004) Minority report in fraud detection: classification of skewed data. SIGKEE Explorations 6 (1):50–59
33. Ireson CL (1997) Critical pathways: effectiveness in achieving patient outcomes. J Nurs Adm 27(6):16–23
34. Shapiro AF (2002) The merging of neural networks, fuzzy logic, and genetic algorithms. Insurance: Mathematics and Economics 31:115–131
35. Sokol L, Garcia B, West M, Rodriguez J, Johnson K (2001) Precursory steps to mining HCFA health care claims. In Proceedings of the 34th Hawaii International Conference on System Sciences
36. Spirtes P, Glymour C, Scheines R (1993) Causation, Prediction and Search. Springer, New York
37. Viaene S, Derrig R, Dedene G (2004) A case study of applying boosting Naive Bayes to claim fraud diagnosis. IEEE Trans Knowl Data Eng 16(5):612–620
38. Viveros MS, Nearhos JP, Rothman MJ (1996) Applying data mining techniques to a health insurance information system. In Proceedings of the 22nd VLDB Conference, Mumbai, India, 286–294
39. Wei CP, Hwang SY, Yang WS (2000) Mining frequent temporal patterns in process databases. Proceedings of international workshop on information technologies and systems, Australia, 175–180
40. Williams G (1999) Evolutionary Hot Spots data mining: an architecture for exploring for interesting discoveries. Lect Notes Comput Sci 1574:184–193
41. Williams G, Huang Z (1997) Mining the knowledge mine: The Hot Spots methodology for mining large real world databases. Lect Notes Comput Sci 1342:340–348

42. Yamanishi K, Takeuchi J, Williams G, Milne P (2004) On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. Data Mining and Knowledge Discovery 8:275–300
43. Yang WS, Hwang SY (2006) A process-mining framework for the detection of healthcare fraud and abuse. Expert Syst Appl 31:56–68
44. Yang WS (2002) Process analyzer and its application on medical care. In Proceedings of 23rd International Conference on Information Systems (ICIS02), Spain
45. Yang WS (2003) A Process Pattern Mining Framework for the Detection of Health Care Fraud and Abuse, Ph.D. thesis, National Sun Yat-Sen University, Taiwan