The Institute for Integrating Statistics in Decision Sciences

Technical Report TR-2013-11
May 10, 2013

Statistical Issues in Medical Fraud Assessment

Tahir Ekin
Department of Computer Information Systems and Quantitative Methods, Texas State University-San Marcos, USA

Francesca Ieva
Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano, Milano, Italy.

Fabrizio Ruggeri
Consiglio Nazionale delle Ricerche, Istituto di Matematica Applicata e Tecnologie Informatiche, Milano, Italy.

Refik Soyer
Department of Decision Sciences
The George Washington University, USA

Noname manuscript No.
(will be inserted by the editor)

Statistical Issues in Medical Fraud Assessment

Tahir Ekin · Francesca Ieva · Fabrizio Ruggeri · Refik Soyer

Received: date / Accepted: date

Abstract In this paper we provide a survey of the statistical issues in medical fraud assessment. We discuss different types of medical fraud and the type of fraud data that arise in different situations and give a review of the statistical methods that use such data to assess fraud. We also discuss "conspiracy fraud" and the associated dyadic data and introduce Co-clustering methods which have not been previously considered in the medical fraud literature. In so doing, we present some recent work on Bayesian co-clustering for fraud assessment and its extensions. Furthermore, we discuss potential use of decision theoretic methods in fraud detection and demonstrate an example for evaluating fraud detection tools.

Keywords Fraud Detection · Bayesian co-clustering · Data Mining · Decision Analysis

Tahir Ekin
Department of Computer Information Systems and Quantitative Methods, Texas State University - San Marcos, USA
E-mail: t_e18@txstate.edu

Francesca Ieva
MOX – Modellistica e Calcolo Scientifico, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy
E-mail: francesca.ieva@polimi.it

Fabrizio Ruggeri
Consiglio Nazionale delle Ricerche, Istituto di Matematica Applicata e Tecnologie Informatiche, Milano, Italy
E-mail: fabrizio@mi.imati.cnr.it

Refik Soyer
Department of Decision Sciences, School of Business, The George Washington University, Washington D.C., USA
E-mail: soyer@gwu.edu

1 Introduction

The Concise Oxford Dictionary defines fraud as "the use of false representations to gain an unjust advantage". Fraud can take many different forms. With recent development of new technologies traditional forms of fraudulent behaviour such as money laundering, phishing or identity theft have become easier to commit and have been joined by other kinds of fraud.

As noted by Li et al. [26], size of the healthcare sector and the enormous volume of money involved make it an attractive fraud target. According to the National Healthcare Anti-fraud Association (NHCAA), health care fraud 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." [32]

In most developed countries, demographic changes such as the increase of the median age have resulted in increased health care spending [3]. For instance, total U.S. health care spending reached 17.9 percent of the GDP, 2.6 trillion USD, in 2010 [10]. According to U.S. federal agencies, every year three to ten percent of this spending is lost to abuse, fraud and waste ([9],[31],[39]). Other estimates by government and law enforcement agencies placed this loss as high as 10% or 130 billion of dollars (see [31] and [32]). Asymmetry of information between providers and patients, inelastic demand for services, enormous money involved, the presence of third party fees for service payments and unconditional public trust in providers can be given as the main reasons of this huge financial loss. In addition to the financial loss, fraud also severely hinders the US health care system from providing quality care to legitimate beneficiaries. Therefore, effective fraud detection is im-

portant for improving the quality as well as reducing the cost of health care services [31].

Abuse and waste only differ from fraud by the degree of the legal intent. Activities that are inconsistent with established practices and result in unnecessary costs to the health care programs can be classified as medical abuse. Failure to document medical records adequately, providing unnecessary services and charging the insurers higher rates are among these activities. Since it is difficult to know the intent for an activity, distinguishing fraud from waste and abuse is a challenging task.

When speaking about fraud, a distinction has to be made between fraud prevention and fraud detection. Fraud prevention describes measures to stop fraud from occurring in the first place [6]. In contrast, fraud detection involves identifying fraud as quickly as possible once it has been perpetrated. In general, fraud detection comes into play once fraud prevention has failed. In what follows, we will focus on statistical methods for fraud assessment. Our focus will be on health care fraud. The statistical issues in health care fraud are valid for the assessment of medical waste and abuse as well.

There is a dearth of literature dealing with healthcare fraud assessment. The systematic use of statistical approaches in medical fraud assessment in the U.S. has gained some momentum with the "Health Care Fraud and Abuse Control Program" in 1996. However, it was not until recently that the efforts and resources put into health care fraud have increased significantly. Increasing budget deficits in the recent years have put more spotlight on health care expenditures and increased the efforts to decrease the health care spending by cutting off the unnecessary payments using medical fraud assessment tools.

Defining fraudulent behavior, detecting fraudulent cases and measuring fraud losses in health care industry is a difficult task [41]. Medical assessment efforts aim to minimize the percentage of fraud with the available resources. There is a trade-off between the costs of the maximization of the correct identification of fraudulent activities (true positives) and minimizing wasteful costs for unnecessary fraud investigations (false positives).

Also medical fraud assessment can be studied distinguishing the main categories of prevention and detection. As we said before, prevention methods aim to deter potential fraudsters. In fact, creating an anti-fraud culture and improving internal compliance systems can have long term effects against fraud. As pointed out by [36], there are not many studies for measuring the effectiveness of fraud prevention methods. While fraud prevention describes the measures to stop fraud from occurring in the first place, fraud detection involves identifying fraud as quickly as it has occurred. In [26] a comprehensive survey about the many statistical approaches that have been deployed against fraud detection is provided.

This paper is an attempt to review statistical methods used in medical fraud literature and present new tools such as co-clustering, Bayesian approach and decision theory which have not gained much attention in this literature. After providing a discussion of the medical fraud data in Section 2, we focus on statistical methods for detecting fraudulent behavior in Section 3. In particular, in §3.1 general techniques for mining high dimensional data are presented. In §3.2 we concentrate on co-clustering techniques applied to dyadic data and we describe a Bayesian approach in §3.3. Application of decision analysis tools in medical fraud detection is presented in Section 4. This is followed by an overall discussion on potential research areas and final remarks in Section 5.

2 Medical Fraud Data

Unlike e-commerce, credit card and telecommunications fraud, where statistical methods have been commonly used (see [8]) for fraud detection, there is limited research in medical fraud assessment as a result of data limitations. Until recently, it was very difficult for researchers to access data sources especially because of the privacy concerns. Data sets were not made publicly available due to legal and competitive reasons [34]. Nowadays, governmental health organizations and private insurance companies provide more opportunities for researchers to access the medical data they possess.

The Research Data Assistance Center [37] is a CMS (Center for Medicare & Medicaid Services) contractor that provides assistance in using Medicare and Medicaid data for research purposes under certain conditions. On another side project, CMS aims to combine all Medicare and Medicaid data in one suite with the "Integrated Data Repository" and the "One Program Integrity" initiatives in order to increase the accessibility and accuracy of the governmental medical data. This source is only open to contractors as of now. It should be noted that Medicare and Medicaid programs are limited to certain population groups such as people who are over 65 or people who are below a certain income level. Therefore, only one third of U.S residents have access to these governmental programs. With these limitations in mind, Health Care Cost Institute was founded by researchers and some private insurers to understand the drivers of health care costs and utilization using

private insurance data (see [1]). This institute publishes annual reports using private insurance data [22]. Another institution which may provide researchers information about disease specific fraud patterns is the CDC, i.e., the Center for Disease Control and Prevention (`www.cdc.gov`). Despite all these improvements in the transparency of the data, medical data sources are still limited and hard to retrieve. In order to overcome data availability issues, researchers can choose to work with synthetic (simulated) data [4]. Recently, [51] evaluated the impact of using simulated data by comparing real and artificial deception models.

It is important that syntethic data resemble insurance claims since most raw medical data is available in the form of insurance claims. A claim involves the participating beneficiary (patient) and a service provider (hospitals, physicians) and generally contains the attributes of patients, providers and the claim itself. Attributes of a patient can be gender, age, medical history whereas the type and the location of facility are among the attributes of a provider; see for example [33] who considered data from a Chilean insurance company about work incapacity. There are also identifiers associated with each provider and patient. In the U.S., not all identifiers are uniquely associated with a provider. One physician can submit a claim using his own account or his hospital's one. That is how fraudsters can submit the same claim more than once by using different identifier codes. Therefore, new identifiers may be required in analyzing the medical data [29]. In general, fraud in health care context is classified into three categories based on who conducts the fraud: provider (hospitals, physicians) fraud, consumer (patients) fraud and insurer fraud. U.S. law identifies the submission of false claims, the payment or receipt of kickbacks and self-referrals as provider fraud (see [23]). In addition to these, up-coding (charging for a more expensive service), unbundling charges (charging separately for procedures which are initially part of one procedure) can also given examples of provider fraudulent activities [26]. Consumer fraud are the cases that patients are involved in fraudulent activities such as falsifying documents to obtain extra prescription or misusing their insurance cards. Insurer fraud happens when insurers falsify statements or they simply do not provide the insurance they have collected premiums for.

In most industry practices, pre-processing of the data takes most of the time of fraud detection (see [27] and [40] for a relevant discussion). Main data issues include choosing and transforming the attributes (features) that are important for your statistical analysis and handling missing values. Attributes used in fraud studies can be numerical, categorical or binary type of variables. In the context we are mainly concerned in what follows, we refer to data originated in forms of medical claims. We observe if a particular provider (hospital, physician) provides service to a certain beneficiary (patient). We transform this information to a *visitation matrix* (VM) using binary values. Numerical attributes of a claim such as the payment amount is deemed to be very crucial since it is one of the determinants of the cost trade-offs within a fraud investigation. Categorical attributes such as patient characteristics (age, medical condition) and provider background can provide information in revealing the heterogeneous nature of claims data and these can be used in clustering.

Overall, modelling fraud based on available medical data is a challenging task. Fraud is a rare event as there are always more legitimate claims than fraudulent claims. More than 80 percent of the papers reviewed in [34] has skewed data with less than 30 percent fraud. The class distributions of fraudulent cases are dynamic due to changing legal characteristics. In addition, because of the multiple styles of fraud happening around the same time, the fraudulent cases are not homogeneous [18]. Legitimate claims have also changing patterns due to heavy competition in health care industry. Another issue is the presence of missing values in medical data and absence of any systematic guidelines to handle missing data. One of the widely used approaches in industry is to remove the claim lines with missing information which decreases the statistical power of an analysis. Imputation, substitution of a missing value using an estimate retrieved by a statistical analysis such as regression is another way of dealing with missing values (for a deeper discussion see [26]).

3 Statistical Methods for Detecting Fraudulent Behaviour

In this section, we discuss the statistical issues in medical fraud assessment and present a review of fraud assessment methods. Statistical methods have been used against fraud in many different fields starting early 1990s. However, fraud detection ideas have not been discussed publicly to prevent the fraudsters from adjusting to the detection methods accordingly. In [8], a comprehensive survey about statistical fraud detection methods used against money laundering, e-commerce fraud, credit card fraud and telecommunications fraud is provided. As noted by Phua et al. [34] these include approaches such as artificial intelligence, auditing, distributed and parallel computing, econometrics, expert systems, fuzzy logic, genetic algorithms, machine learning, neural networks, pattern recognition and visualiza-

tion. The review of [30] concentrates on the application of data mining methods for financial fraud detection; [14] provides a review of credit card fraud and detection techniques.

As we mentioned before, not enough academic attention has been given to medical and healthcare fraud, as confirmed by Phua et al. in [34]. One of the reasons for this is the lack of publicly available data due to confidentiality and privacy issues. Others include dynamic nature of fraud and changes in legislation over time. Statistical tools used for medical fraud issues generally involve application of data mining methods as they are beneficial when the data is complex and voluminous to be processed. Li et al. [26] provides a review of the application of data mining methods in medical fraud assessment.

In general, the use of statistical methods against fraud brings many challenges for several different reasons. As we explained before, due to the high number of beneficiaries involved and many types of services being provided, data size is huge, usually in terabytes. Beneficiaries and providers are not homogeneous since there is a great variety in the services being provided and the money amount involved. Legal systems and health care procedures change frequently which lead to changes in fraudulent and legitimate patterns.

There are many subjective decisions in health care processes (see [35] for a discussion about the role of subjectivity in medical context) that makes medical fraud detection and statistical decision making difficult. Despite all these challenges, governmental and private organizations are more willing to share the data resources nowadays and this leads to an increase in medical fraud assessment research. However, the use of sophisticated statistical methods in health care fraud detection has been relatively new. The following sections provide a methodological overview of some of the more sophisticated techniques to be adopted for approaching this problem, and propose a new approach based on Bayesian modelling and co-clustering of dyadic data and decision theory.

3.1 Data Mining Methods

Capabilities of both generating, collecting and storing data have been increasing dramatically in the last two decades. In particular, databases are increasing in size in two ways: (1) the number of records or objects in the database and (2) the number of fields or attributes to an object. Fraud detection usually has to deal with the first source of dimensional increase. This explosive growth in stored or transient data has generated an urgent need for new techniques and automated tools that can intelligently assist us in transforming the vast amounts of data into useful information and knowledge.

Data mining, a step in the process of Knowledge Discovery in Databases (KDD), is a method of unearthing information from large data sets. Built upon statistical analysis, it can analyze massive amounts of data and provide useful and interesting information about patterns and relationships that exist within the data that might otherwise be missed. Data mining applications have become more popular in health care industry, particularly medical fraud detection, because of the increasing availability of big data (see, for example, [24] and [50]). Moreover, the shift toward evidence-based practice and outcomes research presents significant opportunities and challenges to extract meaningful information from massive amounts of administrative data to transform it into the best available knowledge to guide clinical practice. As we saw in the previous Section, healthcare has been no exception: modern medicine generates a great deal of information stored in the medical database. Extracting useful knowledge and providing scientific decision-making for the diagnosis and treatment of disease from the database becomes increasingly necessary.

As mentioned before, fraud detection in medical context deals with the first source of dimensional increase, i.e., the increasing number of records/objects in the database. In fact, one of the basic problems addressed by the KDD process within medical context is one of mapping low-level data into other forms that might be more compact, more abstract, or more useful. High dimensional data creates problems in terms of increasing the size of the search space for model induction in a combinatorially explosive manner. In addition, it increases the chances that a data-mining algorithm will find spurious patterns that are not valid in general. Approaches to this problem include methods to reduce the effective dimensionality of the problem and the use of prior knowledge to identify irrelevant variables as well as using expert opinion.

The primary goals of data mining in practice are description and prediction. Description focuses on finding human-interpretable patterns describing the data, whereas prediction involves using some variables to predict unknown or future values of the variables of interest. Although the boundaries between prediction and description are not sharp, the distinction is useful for understanding the overall discovery goal. The relative importance of prediction and description for particular data-mining applications can vary considerably. The goals of prediction and description can be achieved using a variety of particular data-mining methods and modeling techniques.

Unsupervised classification methods are used to detect potential deviations from the frequent patterns in the absence of that information. The objective of unsupervised methods in medical fraud detection is to find claims which vary from the existing normal observations. As they do not require pre-labeled data, unsupervised methods may serve as initial screening to list the potentially fraudulent claims before domain experts are brought in the investigation phase. As less transactions are reviewed, the personnel cost decreases [25]. They are not dependent on a particular classified data set, therefore they can detect changing fraud patterns. Despite these potential advantages, there is a limited number of attempts in medical fraud literature and these methods are relatively untested in the literature [8]. However, even basic unsupervised approaches may prove to be beneficial when combined with the expertise regarding discriminating features (see [12] and the references therein). A cooperation between physicians, statisticians and people involved in decision making is essential both when the model has to be defined and tuned as well as when results are to be analyzed and interpreted.

3.2 Dyadic Data and Co-clustering

The emphasis of previous work in health care has been on types of fraud committed by a single party. Li et al. [26] point out that there is a newly emerging type of fraud called "conspiracy fraud" which involves more than one party. An important characteristic of conspiracy fraud is the need to deal with dyadic data connecting the involved parties. The important feature of dyadic data is that it can be organized into a matrix where rows and columns represent a symmetric relationship. In health care fraud detection the typical relationship of interest is the one between a provider and a beneficiary.

In statistics, clustering techniques are based on decomposing a data set into groups so that the points in one group are similar to each other and are different as possible from the points in other groups. In [29] it can be found how the use of clustering procedures allows for geographical analysis of potential fraud as input to his regression model to identify statistically significant regions in terms of an independent variable. In recent years, co-clustering has emerged as a powerful data mining tool that can analyze dyadic data connecting two entities. Such dyadic data are represented as a matrix with rows and columns representing each entity respectively. An important data mining task pertinent to dyadic data is to get a clustering of each entity. Traditional clustering algorithms do not perform well on such problems because they are unable to utilize the relationship between the two entities. In comparison, co-clustering [21], i.e., simultaneous clustering of rows and columns of a data matrix, can achieve a much better performance in terms of discovering the structure of data [11] and predicting the missing values [2] by taking advantage of relationships between two entities. Medical data consisting of providers' claims for treatments and/or services provided to beneficiaries, organized in visitation matrices, represent a straightforward example of it.

Another important issue in data mining aimed at healthcare fraud detection is the identification of the number of groups and co-groups within the population. To the best of our knowledge, there are few attempts in such direction, mainly inspired to social network applications [20].

3.3 Bayesian Co-clustering

One of the first work on Bayesian cluster analysis is due to Binder [5]. An application of Bayesian clustering to longitudinal data is considered by [19]. Bayesian co-clustering methods arise mainly in applications in data mining and machine learning. For example, [7] presents an example of latent Dirichlet allocation, and [38] proposes a general framework where Bayesian co-clustering models are seen as generative mixture modeling problems where estimation is carried out using a variational algorithm rather than Markov chain Monte Carlo (MCMC) methods. An MCMC based solution is presented in [45] and [46]. More recently, a Bayesian nonparametric approach to the co-clustering based on Dirichlet Process (DP) is proposed in [47], [48] and [49].

In what follows, we consider use of Bayesian co-clustering methods for detection of conspiracy fraud. In so doing, we present the Bayesian model proposed by Ekin et al. [16] for describing and capturing the dyadic dynamic that connects providers and beneficiaries. Co-clustering enables us to group providers and beneficiaries simultaneously, that is, the clustering is interdependent. On the other hand, the Bayesian approach comes out to be very suitable to fraud detection, since it allows for the medical knowledge required in order to decide whether a claim is fraudulent or not to be taken into account as a prior knowledge. The objective of the proposed approach is to identify potentially fraudulent associations among the two parties for further investigation.

Following the Bayesian co-clustering framework in [38], we assume that each row and column to have a mixed membership respectively, from which row and column clusters are generated. Each entry of the data

matrix is then generated given that row-column cluster, i.e., the co-cluster. Moreover, assume that we have I health-care providers and J health-care service users or beneficiaries. Let X_{ij} be a binary random variable representing if the provider i serves user j. In other words, X_{ij} is a Bernoulli random variable

$$X_{ij} = \begin{cases} 1 & \text{if provider } i \text{ serves beneficiary } j \\ 0 & \text{otherwise} \end{cases},$$

We have $\mathbf{X} = \{X_{ij};\, i = 1, \ldots, I,\, j = 1, \ldots, J\}$, a data matrix of size $I \times J$.
Assume that there are K clusters of providers and L clusters of users. Membership probabilities are denoted by $\pi_{1k}; k = 1, \ldots, K$ for row clusters and by $\pi_{2l}; l = 1, \ldots, L$ for column clusters such that

$$\sum_{k=1}^{K} \pi_{1k} = \sum_{l=1}^{L} \pi_{2l} = 1.$$

The latent variables Z_{1i} and Z_{2j}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, denote membership to the row (provider) and column (user) clusters such that $Z_{1i} \in \{1, \ldots, K\}$ and $Z_{2j} \in \{1, \ldots, L\}$. Given $\pi_{\mathbf{1}} = (\pi_{1k}; k = 1, \ldots, K)$ and $\pi_2 = (\pi_{2l}; l = 1, \ldots, L)$, Z_{1i} and Z_{2j} are independent discrete random variables.

Furthermore, given the latent variables Z_{1i} and Z_{2j} , X_{ij}'s are Bernoulli random variables with parameter $\theta_{Z_{1i}Z_{2j}}$, that is,

$$X_{ij}|Z_{1i} = k, Z_{2j} = l, \theta_{kl} \sim Ber(\theta_{kl}) \tag{1}$$

and X_{ij}'s are conditionally independent. The co-clustering problem involves assignment of each X_{ij} to a co-cluster defined by the latent pair $(Z_{1i},\ Z_{2j})$. The Bayesian model involves specification of priors for the unknown parameters π_1, π_2 and $\theta = (\theta_{kl}; k = 1, \ldots, K, l = 1, \ldots, L)$. We can assume independent Dirichlet priors for π_1 and π_2 and independent beta priors for elements of θ. More specifically, we have

$$\pi_1 \sim Dir(\alpha_{1k}; k = 1, \ldots, K), \pi_2 \sim Dir(\alpha_{2l}; l = 1, \ldots, L)$$
$$\theta_{kl} \sim B(a_{kl}, b_{kl}),\ k = 1, \ldots, K,\ l = 1, \ldots, L.$$

Given data matrix $\mathbf{X} = \{X_{ij}; i = 1, \ldots, I, j = 1, \ldots, J\}$, the posterior analysis can be developed by using a standard Gibbs sampler (although this may not be computationally efficient in some problems). The full conditionals for θ_{kl}'s, $k = 1, \ldots, K$, $l = 1, \ldots, L$, can be obtained as (conditionally) independent beta densities

$$\theta_{kl}|\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{X} \sim B\Big(\ a_{kl} + \sum_{i,j} X_{ij}\mathbf{I}(Z_{1i} = k, Z_{2j} = l),\ b_{kl} + \sum_{i,j}(1 - X_{ij})\mathbf{I}(Z_{1i} = k, Z_{2j} = l)\Big)$$

where $\mathbf{Z}_1 = \{Z_{1i}; i = 1, \ldots, I\}$, $\mathbf{Z}_2 = \{Z_{2j}; j = 1, \ldots, J\}$ and $\mathbf{I}(\bullet)$ is the indicator function. The full conditionals of π_1 and π_2 are (conditionally) independent Dirichlet distributions given by

$$\pi_1|\mathbf{Z}_1 \sim Dir\Big(\alpha_{1k} + \sum_{i,j} \mathbf{I}(Z_{1i} = k); k = 1, \ldots, K\Big),$$
$$\pi_2|\mathbf{Z}_2 \sim Dir\Big(\alpha_{2l} + \sum_{i,j} \mathbf{I}(Z_{2j} = l); l = 1, \ldots, L\Big).$$

Finally, the full conditionals of (Z_{1i}, Z_{2j}) can be obtained as

$$p(Z_{1i} = k, Z_{2j} = l|\pi_1, \pi_2, \theta, X_{ij}) = \frac{\theta_{kl}^{X_{ij}}\,(1 - \theta_{kl})^{1-X_{ij}}\,\pi_{1k}\,\pi_{2l}}{\sum_{r=1}^{K}\sum_{c=1}^{L}\theta_{rc}^{X_{ij}}\,(1 - \theta_{rc})^{1-X_{ij}}\,\pi_{1r}\,\pi_{2c}}. \tag{2}$$

The model presented above provides a straightforward framework for medical fraud detection and investigation. In fact, the inference carried out observing the posterior conditional distribution of Θ enables the user to flag potential fraudulent association between providers and beneficiaries, to be further investigated by decision makers and people in charge with healthcare monitoring and governance. In fact, observing the posterior distribution of Θ and coming back to the original data, it is possible to see if any anomalous association among providers and beneficiaries emerged. Moreover, posterior inference provided by (2) highlights possibly latent stratification among beneficiaries.

Further refinements and generalizations may be easily handled starting from the previously presented framework. For example, count data instead of binary data can be considered as output in the $\mathbf{X}$ matrix. In this case the entities of the matrix may include the number of times a beneficiary uses the services of a particular provider over a period of time. Such information may be more relevant in detecting fraud than just knowing that if the beneficiary has used the particular provider's service. Also covariates may be included in the model for predicting θ_{kl}s and π_i, $i = 1, \ldots, I$. Moreover, a modified algorithm allowing for dynamic clustering of patients and/or providers is given by [17]. Such extensions can provide a tool for dynamic monitoring of the system over time, which detects "possibly fraudulent" associations and behaviors to be further investigated in a decision analysis.

4 Decision Analysis for Healthcare Fraud

The focus of the statistical literature on health care fraud has been mostly limited to developing methods and algorithms to identify potential fraudulent claims. As a result, other important areas such as fraud prevention, fraud intervention and evaluation of fraud detection algorithms have not been given the consideration that they deserved. As pointed out by [36], there is a lack of evidence on the effectiveness of the health care fraud intervention strategies.

Decision theoretic approaches and Bayesian ideas have been considered in evaluation of fraud detection algorithms in the machine learning and financial and auto insurance fraud literatures. For example, [34] provide a discussion of performance measures used for different fraud detection algorithms using examples from auto insurance fraud. In addition to the evaluation of accuracy with error based methods, cost based metrics such as Receiver Operating Characteristic (ROC) analysis are also considered in performance evaluation. ROC analysis plots the costs of different true positive (correct identification of fraud) and false positive (incorrect identification or fraud) rates. In a review of financial fraud detection methods, [30] point out that the cost of false negatives (misclassifying a fraudulent case as normal) is higher than the cost of false positives. These false negatives result in opportunity costs associated with the medical education of the fraudulent providers and construction of more complex policies against fraud. Decision analysis tools have been considered in [44] for evaluating computer intrusion detection systems. The authors present an integration of ROC analysis and cost analysis to develop an expected cost metric. In so doing, they also demonstrate how decision trees can be used to combine these two tools.

A decision theoretic approach is considered in [43] to address the question of which cases to inspect (or audit) first for potential fraud given limited resources. A utility based fraud detection model is proposed, providing rankings ordered by decreasing expected outcome of inspecting the potentially fraudulent cases. Their outcome is affected by the likelihood of fraud, inspection costs and expected payoff. A more formal expected utility based approach is introduced by [15] for optimal auditing in auto insurance fraud cases.

The above decision-theoretic approaches or their extensions are applicable to healthcare fraud in addressing issues such as fraud prevention, fraud intervention and evaluation of fraud detection algorithms. Although Bayesian decision analysis have been successfully implemented in supporting management decisions in healthcare organizations, in evaluation of healthcare providers and in helping physicians in identifying effective treatments (see, for example, [42]), these approaches have not been considered in healthcare fraud literature.

4.1 A Decision Model for Evaluation of Fraud Detection

Fraud detection and the following investigation activities are typically costly and they need to be done in an optimal manner by using limited resources. Thus, evaluation of fraud detection tools (algorithms) is a crucial task. Such evaluations can be performed by using a decision analysis setup where the problem is represented by a decision tree and the solution is obtained via computation of the expected utility of the detection tool.

A typical fraud detection tool (or the algorithm) provides probability of fraud in a given case. To introduce some notation let P_D denote the probability of the event that the detection tool predicts fraud. Let us define this event by D_F and its complement by $\overline{D}_F$, that is, $\overline{D}_F$ denotes the event that the tool predicts no fraud and has probability $(1 - P_D)$. Given the probabilities provided by the tool for the specific case (such as submitted insurance claims), the decision maker has to decide whether to perform a comprehensive audit of the particular party involved in the case. We denote the actions of audit and no audit by A and $\overline{A}$ respectively.

A particular choice is made and depending on if the case is fraudulent or not a consequence is realized. The decision maker's probabilities of the case being fraudulent (F) or not $(\overline{F})$ depend on the detection tool's prediction. In other words, these are conditional probabilities. More specifically, we define $P_1 = Prob(F|D_F)$ and $P_2 = Prob(F|\overline{D}_F)$ which are referred to as the posterior probabilities of F given the prediction. In other words, we implicitly assume that prior to the prediction by the detection tool, the decision maker had prior probability $p = Prob(F)$ for the case being fraudulent. Once prediction is provided by the tool, this probability is revised accordingly to the posterior probabilities via the Bayes' rule

$$Prob(F|D_F) \ = \frac{Prob(D_F|F)Prob(F)}{Prob(D_F)}.$$

Note that in the above $Prob(D_F)$ is simply P_D which can be written as

$$P_D = Prob(D_F|F)Prob(F) + Prob(D_F|\overline{F})Prob(\overline{F})$$

where $Prob(D_F|F)$ is referred to as the *sensitivity* of the tool and $Prob(D_F|\overline{F})$ is referred to as the probabil-

ity of *false positive*. Similarly in evaluating $(1 - P_D) = Prob(\overline{D}_F)$ we have

$$Prob(\overline{D}_F) = Prob(\overline{D}_F|F)Prob(F)+Prob(\overline{D}_F|\overline{F})Prob(\overline{F})$$

where $Prob(\overline{D}_F|\overline{F})$ is called the *specificity* of the tool and $Prob(\overline{D}_F|F)$ is the probability of *false negative*. The error probabilities $Prob(D_F|\overline{F})$ and $Prob(\overline{D}_F|F)$ are also used in ROC analysis.

Given the prediction by the detection tool, the action chosen by the decision maker and outcome of the case collectively define the consequence. For example, if the detection tool prediction is fraud and the decision maker chooses to audit then the consequence depends on whether the case is fraudelent or not. We denote these consequences with the utility terms $u(D_F, A, F)$ and $u(D_F, A, \overline{F})$. The first term $u(D_F, A, F)$ reflects the benefits of correct prediction by the tool, correct action by the decision maker and the cost of audit whereas $u(D_F, A, \overline{F})$ reflects the costs of audit and the false positive.

This sequential process can be represented by the decision tree given in Figure 1 where all possible paths associated with different combinations of decisions and uncertain outcomes are illustrated. In Figure 1, the chance node R_1 represents the outcome of the detection tool. Based on the outcome (prediction), this is followed either by the decision node D_1 or D_2 where the decision maker chooses to audit or not and depending on the outcome of the case (reflected by the outcomes of chance nodes $R_2, \ldots, R_5$), a particular consequence is realized. For example, the consequence of path $R_1 - D_2 - R_2$ is utility $u(D_F, A, F)$.

The expected utility of the detection tool can be evaluated by rolling back the decision tree in the usual manner; see for example [28]. This is achieved by taking expectations at the chance nodes and maximizing expected utility at the decision node. For example, at chance node R_2 we compute the expected utility as

$$u(R_2) = P_1 u(D_F, A, F) + (1 - P_1)u(D_F, A, \overline{F}).$$

Similarly, at R_3 we obtain expected utility $u(R_3)$ and at decision node D_2 the optimal action is chosen by maximizing the expected utility, that is,

$$u^*(D_2) = max\{u(R_2), u(R_3)\}.$$

In the bottom portion of the tree, expected utilities $u(R_4)$ and $u(R_5)$ are evaluated and the optimal action is chosen by maximizing the expected utility at D_1. This results in $u^*(D_1)$. Finally, the expected utility of the detection tool is obtained at chance node R_1 as

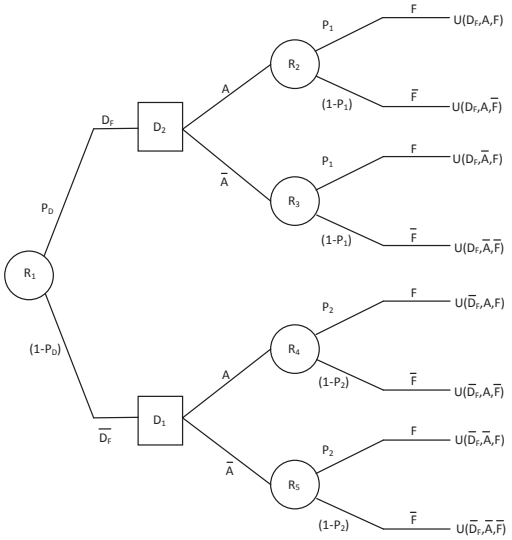

Fig. 1 Decision Tree for Detection Tool Evaluation.

$$u(R_1) = P_D u^*(D_2) + (1 - P_D)u^*(D_1).$$

If we have different fraud detection tools they can be compared by using $u(R_1)$, that is, their expected utilities. Alternatively, the decision maker can decide whether to use a particular detection tool by taking into account the cost of the tool. In this case, the utilities at the end of each path of the tree will be revised to reflect such a cost. Then the expected utility of the detection tool can be compared with the decision maker's expected utility in the absence of the tool and the optimal action can be obtained.

5 Conclusion

This paper provides a review of statistical issues in medical fraud assessment. After providing a discussion about different types of fraud, the nature of medical fraud data and availability of data sources, main statistical issues and corresponding methodologies are discussed. Most of the commonly used statistical methods include data mining techniques such as unsupervised classification and clustering that are aimed for identifying unusual patterns in data that may be indications of potential fraudulent behavior. However, such techniques are effective to deal with single-party fraud and

are not suitable for conspiracy fraud that involves more than one party.

The paper introduces co-clustering methods that are suitable for analysis of dyadic data associated with conspiracy fraud. Although such methods have been used in data mining applications in areas such as marketing, they have not been previously considered in health care fraud literature. Since expert judgment is essential in assessment of fraudulent behavior in many cases, a Bayesian framework, which can incorporate such subjective input into the analysis, is proposed. In so doing, a Bayesian co-clustering algorithm is presented for binary dyadic data using MCMC methods and its extensions are discussed.

The Bayesian methods provide a probabilistic assessment of fraud and are capable of revising the probabilistic assessments based on new data. Thus, using efficient co-clustering algorithms Bayesian methods can provide real time monitoring and analysis of fraudulent behavior. In so doing, they also provide a formalism to incorporate subjective expert knowledge into the analysis. As more government and private organizations are becoming more open to share data sources, it will be feasible to do such real time analysis and also to evaluate performance of fraud assessment methods. As discussed in Section 4, decision analytic approaches can be used for evaluation of fraud detection process. This would also be helpful in developing an optimal strategy for investigation and would minimize the costs.

References

1. Abelson, R., 4 Insurers Will Supply Health Data, New York Times (2011) http://www.nytimes.com/2011/09/20/health/policy/20health.html?_r=0
2. Agarwal, D., Merugu, S., Predictive discrete latent factor models for large scale dyadic data. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 26–35 (2007)
3. Anderson, G., Hussey, P., Comparing health system performance in oecd countries, Health Affairs, 20, 3, 219–232 (2001)
4. Barse, E., Kvarnstrom, H., and Jonsson, E., Synthesizing test data for fraud detection systems. In Computer Security Applications Conference, Proceedings of 19th Annual IEEE Conference, 384–394 (2003)
5. Binder, D.A., Bayesian Cluster Analysis, Biometrika, 65, 1, 31-38 (1978)
6. Blais, E., Bacher, J., Situational deterrence and claim padding: results from a randomized field experiment, Journal of Experimental Criminology, 3, 4, 337–352 (2007)
7. Blei, D.M., Ng, A.Y., Jordan, M.I., Latent Dirichlet Allocation, Journal of Machine Learning Research, 3, 993-1022 (2003)
8. Bolton, R. and Hand, D., Statistical fraud detection: A review, Statistical Science, 17, 3, 235–249 (2002)
9. CMS 2009, Improper medicare fee for service payments report. http://www.cms.gov/apps/er_report/preview_er_report.asp?from=public&which=long&reportID
10. CMS 2010 Financial Report. http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/downloads/highlights.pdf
11. Cheng, Y., Church, G., Biclustering of expression data. In Proceedings of the International Conference of Intelligent Systems and Molecular Biology, 8, 93-103 (2000)
12. Copeland, L., Edberg, D., Panorska, A., Wendel, J., Applying business intelligence concepts to medicaid claim fraud detection, Journal of Information Systems Applied Research, 5, 1, 51–61 (2012)
13. Davis, K., Schoen, C., Guterman, S., Shih, T., Schoenbaum, S.C., Weinbaum, I., Slowing the Growth of U.S. Health Care Expenditures: What Are the Options?, The Commonwealth Fund, 47 (2007)
14. Delamaire, L., Abdou, H., Pointon, J., Credit card fraud and detection techniques: a review, Banks and Bank Systems, 4, 2, 57–68 (2009)
15. Dionne, G., Giuliano, F., Picard, P., Optimal auditing with scoring: Theory and application to insurance fraud. Management Science, 55, 1, 58–70 (2009)
16. Ekin, T., Ieva, F., Ruggeri, F., Soyer, R., Application of Bayesian Methods in Detection of Healthcare Fraud, accepted for publication in Chemical Engineering Transaction, 33. [Online] http://business.gwu.edu/decisionsciences/i2sds/pdf/TR-2013-1.pdf (2013)
17. Ekin, T., Ieva, F., Ruggeri, F., Soyer, R., Bayesian Co-Clustering Methods for Assessment of Healthcare Fraud, Submitted, Technical Report TR-2013-4 I^2SDS - The George Washington University, Washington D.C. (2013)
18. Fawcett, T., In vivo spam filtering: a challenge problem for kdd, ACM SIGKDD Explorations Newsletter, 5, 2, 140–148 (2003)
19. Franzen, J., Bayesian Cluster Analysis: Some Extensions to Non-standard Situations, University dissertation from Stockholm : Statistiska institutionen (2008)
20. Handcock, M.S., Raftery, A.E., Tantrum, J.M., Model-based clustering for social networks, Journal of the Royal Statistical Society - Series A, 170, 2, 301–354 (2007)
21. Hartigan, J., Direct clustering of a data matrix, Journal of the American Statistical Association, 67, 337, 123–129 (1972)
22. Health Care Cost Institute, http://www.healthcostinstitute.org/ (2012)
23. Kalb, P., Health care fraud and abuse, Journal of the American Medical Association, 282, 12, 1163–1168 (1999)
24. Koh, H., Tan, G., et al., Data mining applications in healthcare, Journal of Healthcare Information Management, 19, 2, 65 (2011)
25. Laleh, N., Azgomi, M.A., A Taxonomy of Frauds and Fraud Detection Techniques, Information Systems, Technology and Management Communications in Computer and Information Science, 31, 256–267 (2009)
26. Li, J., Huang, K-Y., Jin, J. and Shi, J., A survey on statistical methods for health care fraud detection, Health Care Management Science, 11, 275–287 (2008)
27. Lin, J. and Haug, P., Data preparation framework for preprocessing clinical data in data mining. In AMIA Annual Symposium Proceedings, 489. American Medical Informatics Association (2006)
28. Lindley, D. V., Making Decisions, *Wiley* (1985)
29. Musal, R., Two models to investigate Medicare fraud within unsupervised databases, Expert Systems with Applications, 37, 12, 8628–8633 (2010)

30. Ngai, E., Hu, Y., Wong, Y., Chen, Y., Sun, X., The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, Decision Support Systems, 50, 3, 559–569 (2011)
31. National Health Care Anti Fraud Association, The Problem of Health Care Fraud: A serious and costly reality for all Americans, report of National Health Care Anti-Fraud Association (2005) http://www.nhcaa.org/resources/health-care-anti-fraud-resources/consumer-info-action.aspx
32. National Health Care Anti Fraud Association, What is Health Care Fraud? (2012) http://www.nhcaa.org/resources/health-care-anti-fraud-resources/consumer-info-action.aspx
33. Ortega, P. A., Figueroa, C. J., Ruz, G. A., A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile. Proceedings of the 2006 International Conference on Data Mining (2006)
34. Phua, C., Alahakoon, D., Lee, V., Minority report in fraud detection: classification of skewed data, ACM SIGKDD Explorations Newsletter, 6, 1, 50–59 (2004)
35. Press, S., Tanur, J., The subjectivity of scientists and the bayesian approach, Wiley Series in Probability and Statistics, Wiley (2001)
36. Rashidian, A., Joudaki, H., Vian, T., No evidence of the effect of the interventions to combat health care fraud and abuse: A systematic review of literature. PloS one, 7, 8, e41988 (2012)
37. Resdac cms data. http://www.resdac.org/cms-data.
38. Shan, H., Bayesian Co-Clustering. 8th IEEE International Conference on Data Mining, 530-539 (2008)
39. Shin, H. and Park, H. and Lee, J. and Jhee, W.C., A scoring model to detect abusive billing patterns in health insurance claims, Expert Systems with Applications, Elsevier (2012)
40. Sokol, L., Garcia, B., West, M., Rodriguez, J., Johnson, K., Precursory steps to mining hcfa health care claims. In System Sciences, Proceedings of the 34th Annual IEEE Hawaii International Conference (2001)
41. Sparrow, M.K., License to steal: why fraud plagues America's health care system, Westview Press (1996)
42. Spiegelhalter, D., Abrams, K., Myles, J., Bayesian approaches to clinical trials and health-care evaluation, Wiley (2004)
43. Torgo, L., Lopes, E., Utility-based fraud detection. In Proceedings of the 22nd international joint conference on Artificial Intelligence, 2, 1517–1522, AAAI Press (2011)
44. Ulvila, J. W., Gaffney, J. E., A decision analysis method for evaluating computer intrusion detection systems.Utility-based fraud detection, Decision Analysis, 1, 35–50 (2004)
45. Wang, P., Domeniconi, C., Laskey, K., Latent Dirichlet Bayesian co-clustering. In Proceedings of the European Conference on Machine Learning, 522–537 (2009)
46. Wang, P., Domeniconi, C., Laskey, K., Nonparametric Bayesian Clustering Ensembles, Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science, 6323, 435–450 (2010)
47. Wang, P., Nonparametric Bayesian Models for Unsupervised Learning, PhD Dissertation, George Mason University, Fairfax, VA (2011)
48. Wang, P., Domeniconi, C., Rangwala, H., Laskey, K., Feature Enriched Nonparametric Bayesian Co-clustering, Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science, 7301, 517–529 (2012)
49. Wang, J., Zhou, B., Yan, R., Benefits and Barriers in Mining Healthcare Industry Data, International Journal of Strategic Decision Science, doi: 10.4018/jsds.201200103 (2012)
50. Yang, W., Hwang, S., A process-mining framework for the detection of healthcare fraud and abuse, Expert Systems with Applications, 31, 1, 56–68 (2006)
51. Yang, Y., Mannino, M., An experimental comparison of real and artificial deception using a deception generation model, Decision Support Systems, 53, 3, 543–553 (2012)