journal homepage: www.intl.elsevierhealth.com/journals/cmpb

A prescription fraud detection model

Karca Duru Aral[a], *Halil Altay Güvenir*[b,*], *İhsan Sabuncuoğlu*[c], *Ahmet Ruchan Akar*[d,e]

[a] *INSEAD, Technology & Operations Management Area, Fontainebleau, France*
[b] *Department of Computer Engineering Bilkent University, Ankara, Turkey*
[c] *Department of Industrial Engineering, Bilkent University, Ankara, Turkey*
[d] *Department of Cardiovascular Surgery, Ankara University School of Medicine, Ankara, Turkey*
[e] *Ankara University Stem Cell Institute, Ankara, Turkey*

ARTICLE INFO

Article history:
Received 23 November 2010
Received in revised form
12 September 2011
Accepted 13 September 2011

Keywords:
Health care fraud
Prescription fraud
Data mining
Outlier detection

ABSTRACT

Prescription fraud is a main problem that causes substantial monetary loss in health care systems. We aimed to develop a model for detecting cases of prescription fraud and test it on real world data from a large multi-center medical prescription database. Conventionally, prescription fraud detection is conducted on random samples by human experts. However, the samples might be misleading and manual detection is costly. We propose a novel distance based on data-mining approach for assessing the fraudulent risk of prescriptions regarding cross-features. Final tests have been conducted on adult cardiac surgery database. The results obtained from experiments reveal that the proposed model works considerably well with a true positive rate of 77.4% and a false positive rate of 6% for the fraudulent medical prescriptions. The proposed model has the potential advantages including on-line risk prediction for prescription fraud, off-line analysis of high-risk prescriptions by human experts, and self-learning ability by regular updates of the integrative data sets. We conclude that incorporating such a system in health authorities, social security agencies and insurance companies would improve efficiency of internal review to ensure compliance with the law, and radically decrease human-expert auditing costs.

1. Introduction

Fraud is defined as the abuse of a profit organization's system without necessarily leading to direct legal consequences. Levi and Burrows define fraud as a mechanism through which the fraudster gains an unlawful advantage or causes unlawful loss [1]. Fraud constitutes a critical problem in many areas such as health care [2], banking [3], insurance [4], and telecommunications [5]. Prescription fraud is defined as the illegal acquisition of prescription drugs for personal use or profit, and could be observed in numerous ways. Any effort aiming to identify the fraudulent transactions in such domains is called as a fraud detection process. Recent data have suggested that traditional manual detection conducted by human experts is quite costly as a result of high expert wages, and large size of the databases. Other main drawbacks of manual detection are that individual human experts cannot recognize the newly emerged fraud patterns spread out in the database, and cannot manage to detect the fraudulent behavior the moment it is attempted. Thus, customized data mining algorithms should analyze the enormous databases of these large businesses, and then human experts can further inspect identified risky trasactions.

Having seen a yearly exponential increase in spending, abuse of health care systems is becoming more critical in

[*] *Corresponding author*. Tel.: +90 312 290 1252; fax: +90 312 266 4047.
E-mail address: guvenir@cs.bilkent.edu.tr (H.A. Güvenir).

0169-2607/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2011.09.003

Table 1 – Health care spending in Turkey by years.

Billion TL	2002	2007	2008
Total social insurance spending	7.6	20	24
Total medicament spending	4.3	8.6	10.5
Total hospital spending	2.8	10.3	13
State hospital payments by social SSA	1.8	6.4	7.5

SSA; Social Security Agency in Turkey known as SGK (Sosyal Güvenlik Kurumu).

Turkey as in many other countries [6]. As for the USA, according to General Accounting Office, annual health care expenditures have approached two trillion dollars, which is 15.3% of the Gross Domestic Product by 2007 [7]. The National Health Care Anti-Fraud Association (NHCAA) estimated that 3% of all health care spending which adds up to be $68 billion is lost to health care fraud in the United States. Other estimates are around 10% or $170 billion for this lost amount [8]. Examples for fraud in a healthcare system would be billing for services and goods that are not rendered, performing medically unnecessary operations or prescribing unnecessary medicines.

The experts from Social Security Agency (SSA, known as SGK) in Turkey commonly detect prescription fraud in their audits. Currently, while auditing the hospitals, SSA officer examines a small sample of the hospital prescriptions and then SSA charges the hospital by a proportional amount. This method is both costly to conduct and does not guarantee any efficiency coefficient. It is worth noting, however, that undetected fraud continues to be an enormous burden on the Turkish health-care system. According to Turkish Health Care Syndicate 2008 Health Care Report, fraud in health care has boomed in Turkey recently [6]. Having seen a yearly exponential increase in spending as shown in Table 1, health care systems' abuse is becoming more and more critical. In 2008, health care fraud was committed principally in Van, Eskişehir, Erzurum, Siirt, Adana, Bursa, Zonguldak, Diyarbakır, and many other cities even in the Head Center of the Tuberculosis Fighting Department. These fraudulent acts were in the form of fake medicament reports, fake invoices, billing Social Security Agency (SSA) for examinations, and treatments that were not rendered. The total cost of these fraudulent acts being millions of TL, and about 300 people were arrested regarding fraud charges recently. Indeed, Turkish healthcare laws provide significant legal sanctions for fraud and abuse control (Turkish Penal Law-26.09.2004, No: 5237/204). In contrast, the perception of the Turkish society that the prescription fraud is a victimless crime make it even more widespread and strengthen the fraudulent chain between the pharmaceutical companies, physicians, pharmacies, and patients. Since nearly half of the spending of the SSA is on medical drug payments, which summed up to 10.5 billion TL in 2008 [6], we see that the cost of the fraudulent prescriptions to the SSA is not tolerable. This type of fraud compromises of excessive medicine prescription, and disunity of patients' features with the prescribed medicines. The orthodox manual detection is conducted by a committee of assigned medical doctors in the SSA. When inspecting a hospital, a human expert goes through a relatively small sample of the prescriptions associated with the hospital. If there are fraudulent and abusive claims in the sample, then the agency charges the hospital to pay the amount acquired by multiplying the percentage of the fraudulent claims detected in the sample and the total cost of the prescriptions issued by the hospital in that inspection period. This method is both costly to conduct and does not guarantee any efficiency coefficient for the outcome.

In order to enable an automated user-friendly system to overcome the above-mentioned handicaps, in this paper, we propose a prescription fraud detection tool that is able to highlight the prescriptions that constitute higher fraud probability threshold assessed by the user. Risk measurements are calculated for cross-features in a knowledge-based setting to compare to the common practice by certain distance metrics. The system incorporates an efficient on-line structure that can be integrated with the electronic on-line prescription provision systems already in use in health care institutions. Although originally intended for prescription fraud detection, any other medical claim (blood tests, X-rays, MRI scans, biopsies, etc.) supervision constitutes promising areas of future applications of the proposed methodology. The underlying assumption for building such a system is that the fraudulent behaviors related to a cross feature are outliers when considering the total data set.

Rest of the paper is organized as follows. Section 2 provides a comprehensive literature review on fraud detection studies. This survey indicates that there are three main types of fraud detection techniques proposed for health care. These are supervised, unsupervised, and hybrid systems. Since we work on a data set without any prior knowledge on prescriptions' label to be fraudulent or not, the proposed system is considered as unsupervised. Section 3 discusses the data structure, the proposed methodology, and the related risk assessment formulations. Section 4 presents the results of computational experiments for both the off-line and on-line applications using real data. The empirical validations of the proposed system and its performance compared to a human expert are also given in this section. Finally, we give concluding remarks and further research directions in Section 5.

2. Related work

There are various resources relating to fraud detection. Fraud detection being a relatively large field, most of the studies considers outlier detection as a primary tool [9]. The investigators mainly incorporate artificial intelligence, data mining, expert systems, fuzzy logic, statistics and visualization. Nonetheless, studies on health care insurance fraud detection are limited. We can group the existing methodologies of fraud detection as being supervised, unsupervised, or as being hybrids of the above.

2.1. *Supervised approaches*

Supervised algorithms are trained by previously labeled training set of fraudulent and legitimate transactions. Then, the algorithms allocate mathematical methodologies to assign scores of similarity with the fraudulent profiles. The most popular applications of supervised algorithms are neural networks. In this context, Kim et al. propose a neural network model for telecommunication subscription fraud [10]. In another study, Barse et al. introduce a multi-layer neural network to handle synthetic database of Video-on-Demand [11]. For the credit card fraud detection problem, Syeda et al. develop a fuzzy neural network model that works on parallel machines [12]. A feed-forward radial basis function neural network with three-layers is introduced by Ghosh and Reilly [13]. This neural network is trained in two phases to assign risk scores to new credit card transactions periodically.

Maes et al. compare neural networks and Bayesian networks. Back propagation algorithm is used to train the neural networks [14]. The results indicate that even thought Bayesian networks are more accurate and require a short training time, they are slower in the application for new instances. Another Bayesian Network is developed by Ezawa and Norton, which has four stages and two parameters [15]. The authors assert that all the methods of regression, nearest neighbor, and neural networks are too slow for their data in hand.

Other methods in the literature are decision trees, rule induction, and case-based reasoning. Metan et al. introduce a real time dispatching rules selection system extracting knowledge from the data stream coming from the manufacturer [16]. The incorporated decision tree dynamically updates in response to changes in the manufacturer's conditions. Enabling a flexible and higher quality decisions, the system is tested on simulation runs which reveals that the proposed model outperforms the existing algorithms in the literature.

As for the statistical modeling, Foster and Stine employ least squares regression and stepwise selection of predictors [17]. They assert that traditional statistical methods are effective to be used for fraud detection. Belhadji et al. propose the cooperation of human experts for choosing best indicators (attributes) for fraud detection [18]. Then, the conditional probabilities of fraud for each indicator are calculated accordingly. Afterwards, Probit regressions are used to identify the most important indicators. The flexible thresholds are adjustable for customization regarding the company's fraud policy. Some other techniques in the literature incorporate expert systems, association rules, and genetic algorithms. Pejic-Bach gives an overview of profiling intelligent systems applications in fraud detection and prevention [19].

2.2. *Unsupervised approaches*

In the area of telecommunications fraud detection, Cortes et al. study temporal evolution of large dynamic graphs [20]. The graphs are built up by the sub-graphs named as Communities of Interest (COI). Exponential weighted average method is used to update sub-graphs daily. COIs are built up by the mobile phone accounts using call quantity and durations. The study yields the specifications of the telecommunication fraudsters. In medical insurance domain, Yamanishi et al. present the unsupervised SmartSifter [21]. This algorithm works with categorical and continuous variables. SmartSifter investigates statistical outliers by Hellinger distance. On automobile insurance data, Brockett et al. employ Principal Component Analysis of RIDIT scores on rank-ordered categorical attributes [22].

2.3. *Hybrid approaches*

Two sub-categories are identified in the literature as supervised hybrids and unsupervised hybrids.

2.3.1. *Supervised hybrids*

In this category, supervised neural networks, Bayesian networks, and decision trees are the methodologies mostly used to create hybrids. Chan et al. combine naive Bayes, C4.5, CART, and RIPPER classifiers [23]. The results give better efficiency on credit card transactions. Kim and Kim develop a decision tree algorithm to classify the data in hand [24]. They use a weighting function to compute fraud density, and then a back propagation neural network is used to generate a weighted risk score on credit card transactions. He et al. classify the general practitioner dataset by the *k*-nearest neighbor algorithm [25]. The optimal weights of the attributes are computed by genetic algorithms.

2.3.2. *Unsupervised hybrids*

Cortes and Pregibon propose the use of daily updated telecommunication account summaries (signatures) [20]. The fraudulent labeled signatures are then inserted to the training set. This training set is used for training the supervised algorithms such as tree, slipper, and model-averaged regression. The algorithm allows the authors to drive conclusions on the nature of the fraudulent calls. Moreover, Cortes et al. propose a graph-theoretic method [26]. This method is used to visually detect fraudulent international calls. Cahill et al. compute a risk score to each call regarding its similarity to fraudulent profiles and dissimilarity to the account's signature [27]. The signatures are updated with low-score calls. In this updating process, recent calls are given higher weight than older calls. The study by Moreau et al. indicates that supervised neural network and rule induction algorithms perform better than two types of unsupervised neural networks in identifying the shifts between short and long term account behavior profiles [28]. The investigators used the area under the receiver operating characteristic curve (AUC) as the performance measure. There are also studies in which unsupervised approaches are used to classify the insurance data into clusters for incorporating supervised approaches. A three step procedure is proposed by Williams and Huang in which: *k*-means is employed for cluster detection, C4.5 is used for decision tree rule induction, and domain knowledge, then statistical summaries and visualization tools are utilized for rule evaluation [29]. Williams employs a genetic algorithm for the second step to generate rules. This enables the user to explore the rules [30].

Brause et al. present RBF neural networks for screening the outputs of association rules for credit card transactions [31]. Ormerod et al. present a Mass Detection Tool (MDT) for detection of medical insurance fraud [32]. Ethnography is the core element of the proposal for capturing expertise to design the methodology. The MDT uses a dynamic Bayesian Belief Network of fraud indicators. Ortega et al. describes another medical claim fraud/abuse detection system based on data mining used by a Chilean private health insurance company [33]. The proposed detection system employs multilayer perceptron neural networks (MLP). Huang, et al. applies a filter-based feature selection method using inconsistency rate

Table 2 – Attributes in the database.

Feature	Type	Number of values	Explanation
Commercial name of the prescribed drug	Categorical	2659	2659 medicines of different commercial names seen in the database.
Market price of the prescribed drug	Continuous	2659	Prices of the each medicine in Turkish market in 2007 fixed by the Health Ministry.
Prescription I.D. number	Categorical	26,419	Identifying numbers for the 26,419 prescriptions in the database.
Age	Continuous	85	All ages between 0 and 85
Sex	Categorical	2	Female, male
Diagnosis	Categorical	332	332 different diagnosis seen in the database

measure and discretization, to a medical claims database to predict the adequacy of duration of antidepressant medication utilization [34].

This study differs from the existing ones in health care fraud detection in that the domain knowledge learned can be used as: (a) an on-line system to check if a given prescription carries risks of fraud and if so in what respects, (b) an off-line system to process a set of prescriptions and filter out those with a risk greater than a threshold to check further by human experts, (c) self-learning ability of the system by regular updates of the integrative data sets. The next section introduces the proposed methodology.

3. Proposed approach

In general, fraud detection research focuses on nonlinear, black-box supervised algorithms, nonetheless, we can assert that less complex, reliable and faster algorithms are needed. Given that the instances (prescriptions) in our database do not have labels as fraudulent and legitimate, we incorporate an unsupervised approach.

For auditing medical transactions, we need two tools. One is for batch screening/auditing which is an off-line system and the other is for on-line/on time transaction control. This imposes building up two systems that work interactively. Clearly, the on-line system should incorporate strategies to overcome the need for re-processing the whole batch of prescriptions in every new transaction. The data structure and size are also other design considerations. We fulfill these requirements under the assumption that the fraudulent cases are outliers in the database.

3.1. Data structure

The database in hand is already anonymized and allows us to consider the following features in prescription fraud detection: commercial name of the prescribed drug; market price of the prescribed drug, prescription number, age, sex, diagnosis for which the drug is prescribed. The characteristics of these features are given in Table 2. As we explicate the nature of the data in hand, we also see that the following features are correlated: medicine and diagnosis; medicine and age; medicine and sex; diagnosis and the total cost of drugs prescribed for this diagnosis; medicine and medicine interactions in a prescription.

Since there is no correlation between the features like age and sex; we ignore these cross-features. On the other hand, considering the interactions between diagnosis and age as well as diagnosis and sex we can reason that we do not need to include these cross features since any such diagnosis should convey specific medicines in the prescription. These specific medicines should reveal any mismatching between the diagnosis and age or sex. These arguments transform our domain of 6 dimensions to sub-domains of 2 dimensions which are illustrated by the interactions discussed above.

3.2. Methodology

These arguments transform our domain of 6 dimensions into 2 dimensional sub-domains, which are illustrated by the above-mentioned interactions. Therefore, our problem is refined to deal with five two-dimensional spaces. Working with incidence and risk matrices which are to be defined in the subsequently, and having two parts of consideration as on-line and off-line processing, our methodology's flow chart is as shown in Fig. 1.

3.3. Off-line processing

We developed a Matlab 2008A m-file, for the off-line batch processing of the database. This code processes the database to create the incidence matrices for all the domains.

- Medicine and age domain incidence matrix: *MA*.
- Medicine and sex domain incidence matrix: *MS*.
- Medicine and diagnosis domain incidence matrix: *MD*.
- Medicine and medicine domain incidence matrix: *MM*.
- Diagnosis and cost domain incidence matrix: *DC*.

An incidence matrix entry (i, j) corresponds to the number of times the ith and jth traits of the corresponding features are seen together in the database. As for the *DC* matrix, the row labels are diagnoses and column labels are indices from 1 to 204. These indices represent 5 TL (Turkish currency) intervals, but the last interval is for the diagnosis costs that are above 2500 TL. For every diagnosis within a prescription, the total costs of the corresponding medicines are calculated and the number of times a diagnosis *i*'s total cost falls into a cost interval *j* is the incidence matrix entry *DC*(*i*, *j*).

Now having all the incidence matrices in hand, the code creates risk matrices below:

- Medicine and age risks: *MAR*.
- Medicine and sex risks: *MSR*.
- Medicine and diagnosis risks: *MDR*.
- Medicine and medicine risks: *MMR*.
- Diagnosis and cost couple's risks: *DCR*.

These matrices are built up by calculating the risks for the corresponding incidences in the corresponding incidence matrices. For example, for calculating the *MSR*(*i*, *j*), we use the corresponding risk metric for *MS*(*i*, *j*). We need to keep the incidence matrices for on-line processing, so we do not directly update the incidence matrices for risk computations.

Having all the risk matrices in hand, the code goes through all the risks that are greater than the thresholds given by the user. The user can indicate any threshold he wants for any of the risk matrices keeping in mind that more prescriptions would be classified as risky when the threshold is kept small. That is, there is a tradeoff between the true positive rate and the human expert screening time. The user should predefine the level of tradeoff he is ready to accept.

Given the thresholds, the code outputs the fraudulent prescriptions by indicating which types of fraud are seen within the prescriptions. That way, the human expert has the chance to revise the marked prescriptions, which saves time and money in auditing large databases, besides having acquired a list of possible fraudulent transaction styles given the database.

3.4. *On-line processing*

The on-line prescription fraud detection tool is an interactive tool coded in Matlab that has a graphical user interface. Considering the nature of the health care sector where on-line transaction of the incoming invoices is the common practice, we can assert that this kind of an on-line tool is fundamental for instant real time auditing.

This interface is designed to enable the user to insert new prescriptions to the database and audit a new prescription without the need to re-run the off-line code. Thus, new prescription auditing can be done once the off-line code is run on the prescription database in available. Please note that since the database we used is in Turkish, all the generated listings in the on-line user interface are in Turkish. Fig. 2 shows a screenshot of the graphical user interface of the auditing tool.

As seen in Fig. 2, the user first needs to input the prescription number as well as the age and sex of the patient. Then, in the box below the user enters the prescribed drug and the corresponding diagnosis by the add button. The drug and diagnosis list boxes are populated by the Turkish drug names and diagnosis lists of the database, which are the outputs of the off-line fraud detection code. The user can choose to check to see if the input is correct by the view prescription button. If the prescription input is correctly specified, the user might choose to add the prescription directly to the database. That is achieved by fetching the corresponding rows of the incidence and risk matrices and updating those by the on-line code's input of the incoming prescription specifications. Alternatively, the user might want to audit the prescription directly. That way, input of the prescription is not used to update the incidence and risk matrices permanently. This is preferable since if the incoming prescription is fraudulent, updating the incidence and risk matrices by this input would slightly affect the performance of the code. This because increasing the number of outliers in a database would eventually lead the outliers to be the common transactions. This would hinder the tool to detect those fraudulent transactions. As a result, the user should add the incoming prescription to the database if the prescription is not fraudulent, perhaps after the auditing process. Pushing the audit button, the user instantly receives a message indicating each level of fraud risk regarding the prescription. Lastly, the new prescription button enables the user to put in a new prescription right after auditing another one.

3.5. *Risk assessment*

We introduce the risk assessment formulas, which consist of calculating risks given the incidence matrices. As stated previously, incidence matrices hold the information regarding the number of times an instance shows up in the data set.

3.5.1. *Risk metric for categorical features*

Sex, diagnosis, and prescription medicines are the un-ordered categorical features in the data set. The incidence matrix entry (*i*, *j*) is the number of times the medicine *i* is issued to the corresponding un-ordered categorical entry *j*. Medicine–Sex (*MS*), Medicine–Diagnosis (*MD*), and the Medicine–Medicine (*MM*) incidence matrices are the categorical matrices.

Let us denote the maximum incidence entry of the *i*th medicine of an incidence matrix *MF* by $Max_{MF}(i)$, where *F* represents the feature domain. $Max_{MF}(i)$ is the number of times the medicine *i* is issued to the trait that is most issued to.

At this point we introduce a risk estimation function, here after denoted as $risk_{MF}(i)$, that represents the likelihood of fraud when the *i*th medicine is prescribed for the *j*th trait. We required that function to return a real value between 0 and 1. Here, the risk value 1 will represent the highest possible risk of fraud, whereas the value 0 will represent the lowest possible risk. The highest risk value is obtained when *MF*(*i*, *j*) has the lowest value, that is the rarest case. Further, we wanted the risk function to drop exponentially, when *MF*(*i*, *j*) increased, and reach the value 0 when it is equal to $Max_{MF}(i)$, the most common case. Having tried many risk functions that satisfy these criteria, we found that the risk function in Eq. (1) was the most successful one.

$$risk_{MF}(i,j) = \frac{e^{-(MF(i,j)/Max_{MF}(i))} - e^{-1}}{1 - e^{-1}} \tag{1}$$

Then, the risk matrix of the *Medicine* and a feature domain *F* can be defined as: $MFR(i,j) = risk_{MF}(i,j)$.

The risk function in Eq. (1) employs an exponential function in order to achieve a steep trend since we preferred high values of fraud risk only for very small values of $MF(i,j)/Max_{MF}(i)$. That is, the sensitivity of the risk function to detect fraud should increase as the ratio $MS(i,j)/Max_{MS}(i)$ becomes smaller

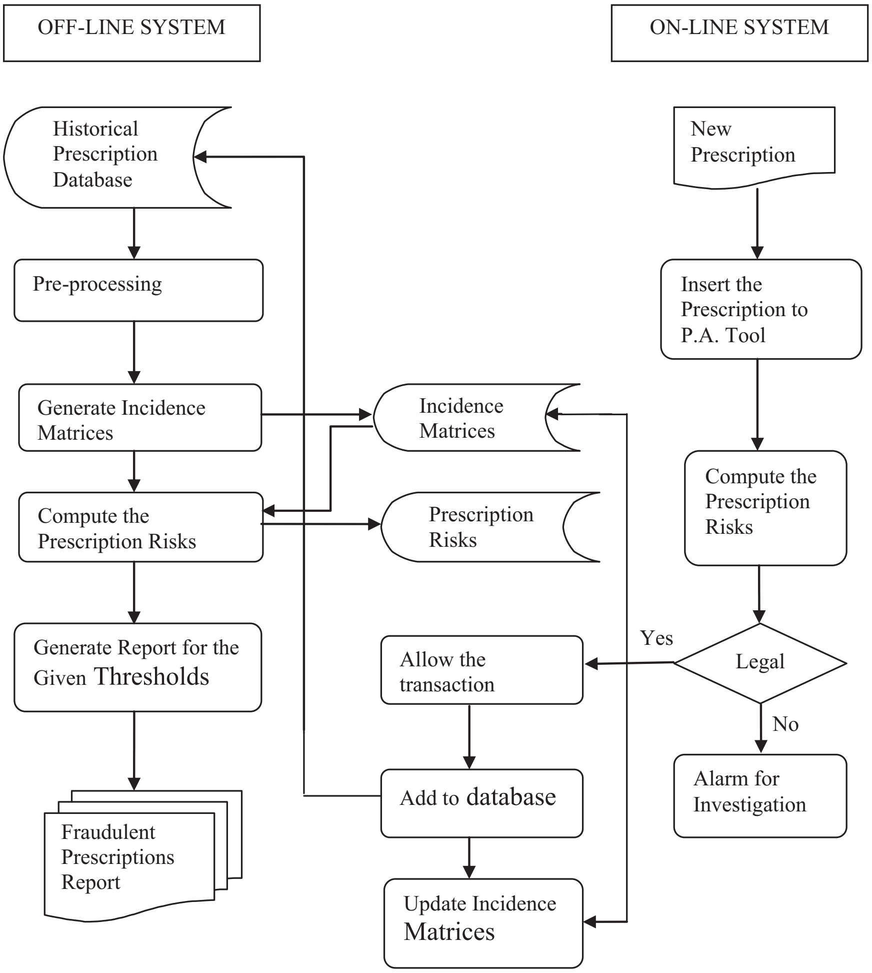
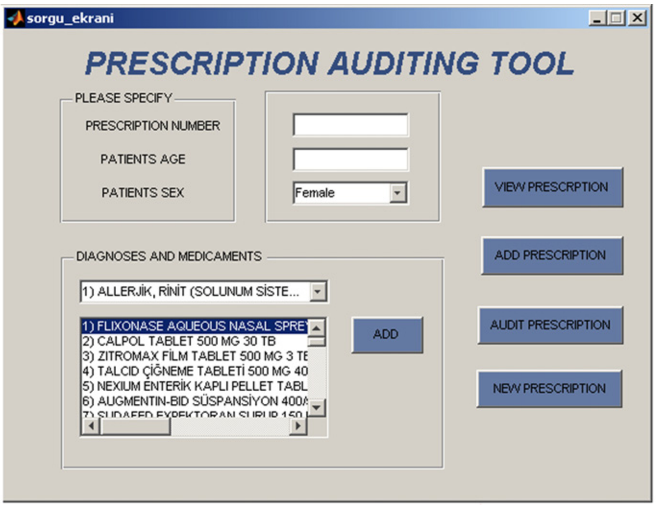

OFF-LINE SYSTEM
ON-LINE SYSTEM
Historical Prescription Database
Pre-processing
Generate Incidence Matrices
Compute the Prescription Risks
Generate Report for the Given Thresholds
Fraudulent Prescriptions Report
Incidence Matrices
Prescription Risks
Allow the transaction
Add to database
Update Incidence Matrices
New Prescription
Insert the Prescription to P.A. Tool
Compute the Prescription Risks
Yes
Legal
No
Alarm for Investigation

Fig. 1 – A schematic view of the flow chart model of the proposed system. P.A:


sorgu_ekrani

PRESCRIPTION AUDITING TOOL

PLEASE SPECIFY

PRESCRIPTION NUMBER

PATIENTS AGE

PATIENTS SEX: Female

DIAGNOSES AND MEDICAMENTS

1) ALLERJİK, RİNİT (SOLUNUM SİSTE...

1) FLIXONASE AQUEOUS NASAL SPREY
2) CALPOL TABLET 500 MG 30 TB
3) ZITROMAX FİLM TABLET 500 MG 3 TE
4) TALCID ÇİĞNEME TABLETİ 500 MG 40
5) NEXIUM ENTERİK KAPLI PELLET TABL
6) AUGMENTIN-BID SÜSPANSİYON 400/
7) SUDAFED EXPEKTORAN ŞURUP 150

Fig. 2 – Prescription auditing tool user interface.

since the derivative of e^{-x} increases as x gets smaller. We than normalize the value of $e^{-(MF(i,j)/Max_{MF}(i))}$ by subtracting $e^{-1}$and dividing by $1 - e^{-1}$ in order to get risk values between 0 and 1 for a straightforward interpretation of the risk levels. Note that here e^{-1} and $1 - e^{-1}$ are constant values.

3.5.2. *Risk metric for ordered features*

Ordered features are features over which we can make a magnitude comparison. Those are often called as continuous features. Here, we define the refined formulations for the Age and Cost ordered features of our database. Consider the Medicine – Age incidence matrix, denoted by *MA*. Let *Max*(i) and *Min*(i) denote the maximum and minimum of ages that the medicine *i* is prescribed to, respectively. In other words, $Max(i) = \{j : Max_{MA}(i) = MA(i, j)\}$ and $Min(i) = \{j : Min_{MA}(i) = MA(i, j)\}$. Then the age range of medicine i is $r_i = Max(i) - Min(i)$. The modified risk metric is:

$$risk_{MA}(i, j) = \frac{e^{-(MA(i,j)/Max_{MA}(i))} \times (1 - d_i(j)/r) - e^{-1}}{1 - e^{-1}} \qquad (2)$$

where,

$$V_i = \frac{\sum_k k \times MA(i, k)}{\sum_k MA(i, k)} \qquad \text{(centroid age for } i\text{th medicine)},$$

and

$$d_i(j) = |j - V_i| \qquad \text{(distance of the } j\text{th age to the centroid age of } i\text{th medicine)}.$$

Then, the risk matrix of the Medicine and Age domain is defined as $MAR(i, j) = risk_{MA}(i, j)$. For the Diagnosis–Cost domain, the formulation is analogous except for that we define the entry $DC(i, j)$ as the number of times the diagnosis i is prescribed medicines of total cost falling into the interval j.

4. Computational results

We develop the code of the proposed framework in Matlab 2008A release. In this system, the user can indicate any threshold he wants for any of the risk matrices keeping in mind that there is a tradeoff between the true positive rate and the human expert screening time. Given the thresholds, the code outputs the fraudulent prescriptions by indicating which types of fraud are seen within the prescriptions. That way, the human expert has the chance to revise the outputted prescriptions, which saves time and money to audit large databases. The on-line prescription fraud detection tool is an interactive tool that has a graphical user interface. This interface is designed to enable the user to insert new prescriptions to the database and audit a new prescription without the need to re-run the off-line code. We run the off-line code on the database of 87,785 prescribed drugs. The tests were run on a PC with 64 byte Core2Duo (3 GHz). The code takes 414 seconds to process the whole data set. As stated above, a run requires the user to specify riskiness thresholds of each kind of confirmation check procedure. The code reveals the prescriptions which possess higher risks than the thresholds. We have taken several runs in order to refine the preferable threshold for each of the domains.

The results indicate that the sensitivity levels of each of the criteria are different. The reason for that lies in the fact that the sizes of the incidence matrices are different from each other and thus the sparseness and intensity characteristics of each differ. That is to say, the maximum numbers in a risk matrix's row and the rows themselves change from matrix to matrix for each medicine leading to different sets of risk indicators for the corresponding features. Thus, each threshold needs a separate refinement. Knowledge inferred needs to be validated and refined by human experts [35]. We achieve this refinement in the supervision of a medical doctor who assessed the significance levels of the outputs since we are interested in building a system that produce outputs meaningful to the human expert fraud auditors who are medical doctors in Turkey. The refined model for each auditing task uses the following threshold values:

- Medicine–Diagnosis Domain: 0.85.
- Medicine–Age Domain: 0.90.
- Medicine–Sex Domain: 0.96.
- Medicine–Medicine Domain: 0.95.
- Diagnosis–Cost Domain: 0.85.

We consider false positive, false negative, and true positive rates as well as the agreement rate as performance indicators for our system. A medical doctor labeled the fraudulent prescriptions in a random sample of 249 prescriptions taken from the database. The comparison between the human expert labeling and the proposed system has led to the following results with 17 false positives, 19 false negatives, 72 true positives, and 141 true negatives. The results are summarized in Table 3. The AUC (Area Under ROC Curve) is 85.7%.

We have compared our system with two existing methods. EFD [36] performed worse with a true positive rate of 26.4%, false positive rate 5.9%, and AUC is 60%. The medical claim fraud/abuse detection system proposed by Ortega et al. [33] achieved a true positive rate of 71%, false positive rate 6%, with AUC is 82.5%.

An interesting observation about the audit results is that the prescriptions labeled as fraudulent tend to have multiple numbers of reasons for risk. For example, let us consider the prescription 1592467 whose database values are given in Table 4.

The output for this prescription is as:
Prescription Number: 1592467

- Incompatibility between Medicine: *Iliadin* Diagnosis: *Glaukoma*, Risk: 0.96.
- Incompatibility between Medicine: *Coraspin* Diagnosis: *Glaukoma*, Risk: 0.92.
- Incompatibility between Diagnosis: *Glaukoma* Cost (TL): 70, Risk: 0.87.

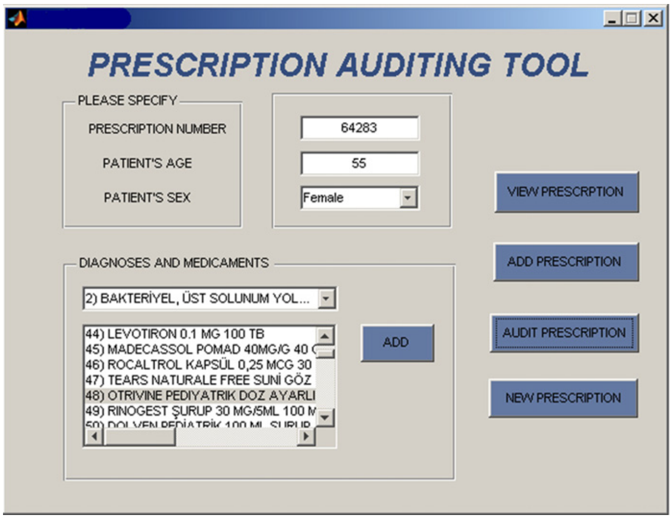
Cosopt, being an ophthalmic suspension, is a legitimate item in the prescription. Nonetheless, *Iliadin* is a nasal spray and *Coraspin* contains acetylsalicylic acid. This might be an indicator that the fraudsters tend to add several fraudulent

Table 3 – Performance indicators.

Performance indicators	Explanation	Performance
False positive rate	$\frac{\text{Number of false positives}}{\text{Total number of instances}}$	6.09%
False negative rate	$\frac{\text{Number of false negatives}}{\text{Total number of instances}}$	7.63%
True positive rate	$\frac{\text{Number of true positives}}{\text{Number of real positives}}$	77.4%
Agreement rate (accuracy)	$\frac{\text{Number of true positives}+\text{number of true negatives}}{\text{Total number of instances}}$	85.54%

Table 4 – Prescription 1592467.

Prescription no.	Drug	Age	Sex	Diagnosis	Price (TL)
1592467	*Iliadin*	57	M	*Glaukoma*	4.59
1592467	*Cosopt*	57	M	*Glaukoma*	30.80
1592467	*Cosopt*	57	M	*Glaukoma*	30.80
1592467	*Coraspin*	57	M	*Glaukoma*	2.40



PRESCRIPTION NUMBER 64283
PATIENT'S AGE 55
PATIENT'S SEX Female

2) BAKTERİYEL, ÜST SOLUNUM YOL...

44) LEVOTIRON 0.1 MG 100 TB
45) MADECASSOL POMAD 40MG/G 40 G
46) ROCALTROL KAPSÜL 0,25 MCG 30
47) TEARS NATURALE FREE SUNİ GÖZ
48) OTRIVINE PEDIYATRIK DOZ AYARLI
49) RINOGEST ŞURUP 30 MG/5ML 100 M
50) DOLVEN PEDİATRİK 100 ML ŞURUP

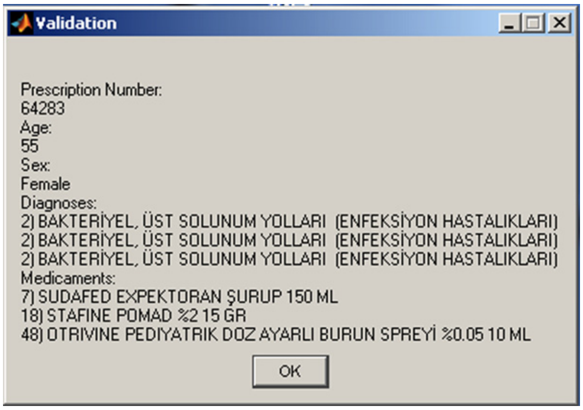
Fig. 3 – Inserting a prescription to the prescription auditing tool.

items in a prescription that could have been legitimate without those.

The on-line code can be run once having the off-line processing done. For illustrating the effectiveness of the on-line fraud detection tool, let us consider a prescription given to a 55 years old woman. Kindly note that the data base we work with is in Turkish, which means that we have Turkish listings in the on-line tool. She is diagnosed with the *upper respiration tube infection* and is given the medicines *Sudafed Syrup*, *Otrivine Pediatric Spray* and *Stafine Pomade*. The initial user interface is as seen in Fig. 3 after inputting the prescription. If the user chooses to view the prescription a message box appears as in Fig. 4. After validating the prescription input, the user might choose to add the prescription to the database. If so, the message box appears as in Fig. 5.

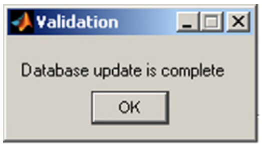
When the user chooses to audit the prescription a message box appears as in Fig. 6. Here, the Medicine and Age non-conformation risk assessments are stated in the input order of the medicines, just as the Medicine and Sex non-conformation. Considering the diagnoses, the Medicine and Diagnosis risks are seen in the screen in the appearance order of the medicine and diagnosis couples in the prescription. Lastly, we see one value for the Diagnosis and Cost non-conformation risk since there is only one diagnosis in the prescription.

Considering the prescription, where the diagnosis is *upper respiration tract infection* and the prescribed medicines are *Sudafed Syrup, Stafine Pomade* and *Otrivine Pediatric Spray,* we can state that the tool is effective to calculate no risks for the medicine and diagnosis domain for the first and the last



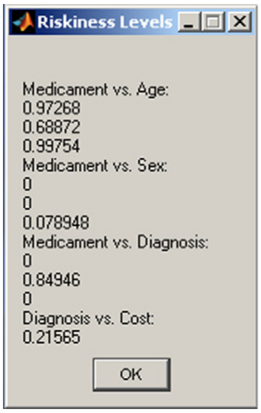
Prescription Number:
64283
Age:
55
Sex:
Female
Diagnoses:
2) BAKTERİYEL, ÜST SOLUNUM YOLLARI (ENFEKSİYON HASTALIKLARI)
2) BAKTERİYEL, ÜST SOLUNUM YOLLARI (ENFEKSİYON HASTALIKLARI)
2) BAKTERİYEL, ÜST SOLUNUM YOLLARI (ENFEKSİYON HASTALIKLARI)
Medicaments:
7) SUDAFED EXPEKTORAN ŞURUP 150 ML
18) STAFINE POMAD %2 15 GR
48) OTRIVINE PEDIYATRIK DOZ AYARLI BURUN SPREYİ %0.05 10 ML

Fig. 4 – Validation message box.



Database update is complete

Fig. 5 – Database update notification.



Medicament vs. Age:
0.97268
0.68872
0.99754
Medicament vs. Sex:
0
0
0.078948
Medicament vs. Diagnosis:
0
0.84946
0
Diagnosis vs. Cost:
0.21565

Fig. 6 – Risk assessment screen.

medicines and a high risk for the second since *Stafine Pomade* is a skin care medicine. For this second medicine we see that the tool calculates a high risk (0.85), which is expected. There is no risk associated with the sex of the patient and the medicines. Nonetheless, both *Sudafed Syrup* and *Otrivine Pediatric Spray* are pediatric medicines. Thus, the tool identifies the high risks regarding the age of the patient as 0.97 for *Sudafed Syrup* and 0.99 for the *Otrivine Pediatric Spray*.

5. Concluding remarks and further research direction

We conclude by proposing a novel model for detecting cases of prescription fraud intended to provide efficient and user-friendly platforms, and save financial resources at the institutional and national levels. Our methodology proposes dividing up the 6 dimensional features' domain into several 2 dimensional sub-domains considering the interaction levels between the features. The methodology consists of populating incidence matrices for each of the above domains and then incorporating a distance based data-mining approach. The risk metrics employed in this data-mining approach return risk measures for each of the domains mentioned above. This risk measure is scaled to be between 0 and 1, in order to give a straightforward definition of the risk level. For each of the domains, the user can specify thresholds. That way, the program alarms for only those prescriptions with risk levels higher than the thresholds.

The automated fraud detection methodology gives considerably compatible results with the human expert auditing. The system is flexible enough for an integrated on-line/on-time user interface, and its on-line incorporation is computationally inexpensive, it presents a novel and easy way to keep track of health care transactions in incidence matrices for auditing. The approach proposed here is able to handle both categorical and ordered features. The output of the system is easy to understand and interpret by human users. Besides, the system can learn and process accordingly as the input data shifts. Finally, its core methodology is adoptable to many other areas in health care and possibly in other industries.

Given the performance measurements with a true positive rate of 77.4% and a false positive rate of 6%, we can conclude that the proposed system works reasonably well for the prescription fraud detection problem. Nonetheless, further refinement of the tool would require scaling the risk outputs across all domains. This would mean that incorporating different parameters for different domains would lead to the same risk measurements across all domains. Besides, a tool can be built up where the user can specify the domains he wants to work on. Efforts must be undertaken to promote cost-effective fraud detection models for other health care practices and interventions that may have an impact on the quality of health-care.

Conflicts of interest

There is no undisclosed ethical problem or conflicts of interest related to this paper.

Acknowledgements

We thank Cagdas Baran for assistance in preparation of the types of fraud and abuse in medical practice and Murat Kurtcephe for various discussions about the topic of this paper.

REFERENCES

[1] M. Levi, M. Burrows, Measuring the impact of fraud in the UK: a conceptual and empirical journey, British Journal of Criminology 48 (3) (2008) 293–318.

[2] A.S. Kesselheim, D.M. Studdert, M.M. Mello, Whistle-blowers' experiences in fraud litigation against pharmaceutical companies, New England Journal of Medicine 362 (19) (2010) 1832–1839.

[3] R. Wheeler, S. Aitken, Multiple algorithms for fraud detection, Knowledge-Based Systems 13 (2–3) (2000) 93–99.

[4] S. Viaene, R.A. Derrig, B. Baesens, G. Dedene, A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection, Journal of Risk and Insurance 69 (3) (2002) 373–421.

[5] C.S. Hilas, P.A. Mastorocostas, An application of supervised and unsupervised learning approaches to telecommunications fraud detection, Knowledge-Based Systems 21 (7) (2008) 721–726.

[6] Turkish Health Care Syndicate 2008 Health Care Report, 2008 (Sağlıkta 2008 Raporu, Türk Sağlık Sen).

[7] J. Li, K. Huang, J. Jin, J. Shi, A survey on statistical methods for health care fraud detection, Journal of Health Care Management Science 11 (3) (2008) 275–287.

[8] USA's National Health Care Anti-Fraud Association Web Page, 2009, http://www.nhcaa.org/eweb/StartPage.aspx.

[9] X. Weng, J. Shen, Detecting outlier samples in multivariate time series dataset, Knowledge-Based Systems 21 (8) (2008) 807–812.

[10] H. Kim, S. Pang, H. Je, D. Kim, S. Bang, Constructing support vector machine ensemble, Pattern Recognition 36 (2003) 2757–2767.

[11] E. Barse, H. Kvarnstrom, E. Jonsson, Synthesizing test data for fraud detection systems, in: Proceedings of the 19th Annual Computer Security Applications Conference, 2003, pp. 384–395.

[12] M. Syeda, Y. Zhang, Y. Pan, Parallel granular neural networks for fast credit card fraud detection, in: Proceedings of the 2002 IEEE International Conference on Fuzzy Systems, 2002.

[13] R. Ghosh, D. Reilly, Credit card fraud detection with a neural-network, in: Proceedings of the Twenty-Seventh Annual Hawaii International Conference on System Sciences, 1994.

[14] S. Maes, K. Tuyls, B. Vanschoenwinkel, B. Manderick, Credit card fraud detection using Bayesian and neural networks, in: Proceedings of the 1st International NAISO Congress on Neuro Fuzzy Technologies, 2002.

[15] K. Ezawa, S. Norton, Constructing Bayesian networks to predict uncollectible telecommunications accounts, IEEE Expert 11 (5) (1996) 45–51.

[16] G. Metan, I. Sabuncuoglu, H. Pierreval, Real time selection of scheduling rules and knowledge extraction via dynamically controlled data mining, International Journal of Production Research 48 (23) (2010) 6909–6938.

[17] D. Foster, R. Stine, Variable selection in data mining: building a predictive model for bankruptcy, Journal of American Statistical Association 99 (466) (2004) 303–313.

[18] E. Belhadji, G. Dionne, F. Tarkhani, A model for the detection of insurance fraud, The Geneva Papers on Risk and Insurance 25 (4) (2000) 517–538.

[19] M. Pejic-Bach, Profiling intelligent systems applications in fraud detection and prevention: survey of research articles, in: Proceedings of International Conference on Intelligent Systems, Modelling and Simulation, 2010, pp. 80–85.

[20] C. Cortes, D. Pregibon, Signature-based methods for data streams, Data Mining and Knowledge Discovery 5 (2001) 167–182.

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

[22] P.L. Brockett, R.A. Derrig, L.L. Golden, A. Levine, M. Alpert, Fraud classification using principal component analysis of RIDITs, Journal of Risk and Insurance 69 (3) (2002) 341–371.

[23] C.L. Chan, C.H. Lan, A data mining technique combining fuzzy sets theory and Bayesian classifier – an application of auditing the health insurance fee, in: H.R. Arabnia (Ed.), Proceedings of the International Conference on Artificial Intelligence IC-AI'2001, 2001, pp. 402–408.

[24] M. Kim, T. Kim, A neural classifier with fraud density map for effective credit card fraud detection, in: Proceedings of IDEAL2002, 2002, pp. 378–383.

[25] H. He, W. Graco, X. Yao, Application of genetic algorithms and *k*-nearest neighbour method in medical fraud detection, in: Proceedings of SEAL1998, 1999, pp. 74–81.

[26] C. Cortes, D. Pregibon, C. Volinsky, Computational methods for dynamic graphs, Journal of Computational and Graphical Statistics 12 (4) (2003) 950–970.

[27] M. Cahill, F. Chen, D. Lambert, J. Pinheiro, D. Sun, Detecting fraud in the real world, in: Handbook of Massive Datasets, 2002, pp. 911–930.

[28] Y. Moreau, E. Lerouge, H. Verrelst, J. Vandewalle, C. Stormann, P. Burge, BRUTUS: a hybrid system for fraud detection in mobile communications, in: Proceedings of European Symposium on Artificial Neural Networks, 1999, pp. 447–454.

[29] G. Williams, Z. Huang, Mining the knowledge mine: the hot spots methodology for mining large real world databases, Lecture Notes in Computer Science (1997) 340–348.

[30] G. Williams, Evolutionary hot spots data mining: an architecture for exploring for interesting discoveries, in: Proceedings of PAKDD99, 1999.

[31] R. Brause, T. Langsdorf, M. Hepp, Neural data mining for credit card fraud detection, in: Proceedings of 11th IEEE International Conference on Tools with Artificial Intelligence, 1999.

[32] T. Ormerod, N. Morley, L. Ball, C. Langley, C. Spenser, Using ethnography to design a Mass Detection Tool (MDT) for the early discovery of insurance fraud, in: CHI'03 Extended Abstracts on Human Factors in Computing Systems, 2003, pp. 650–651.

[33] P. Ortega, C. Figueroa, G. Ru, A medical claim fraud/abuse detection system based on data mining: a case study in Chile, in: Proceedings of DMIN'06, 2006, pp. 224–231.

[34] S.H. Huang, L.R. Wulsin, L. Hua, J. Guo, Dimensionality reduction for knowledge discovery in medical claims database: application to antidepressant medication utilization study, Computer Methods and Programs in Biomedicine 93 (2) (2009) 115–123.

[35] T. Aydın, H.A. Güvenir, Modeling interestingness of streaming association rules as a benefit-maximizing classification problem, Knowledge Based Systems 37 (2) (2009) 1713–1718.

[36] A.J. Major, D.R. Riedinger, EFD: a hybrid knowledge/statistical-based system for the detection of fraud, Journal of Risk and Insurance 69 (3) (2002) 309–324.