DARU Journal of Pharmaceutical Sciences (2018) 26:209–214
https://doi.org/10.1007/s40199-018-0227-z

RESEARCH ARTICLE

Detecting medical prescriptions suspected of fraud using an unsupervised data mining algorithm

Mohammad Haddad Soleymani[1] · Mehdi Yaseri[1] · Farshad Farzadfar[2] · Adel Mohammadpour[3] · Farshad Sharifi[4] · Mohammad Javad Kabir[5]

Received: 31 July 2018 / Accepted: 30 October 2018 / Published online: 20 November 2018

Abstract

Nowadays, health insurance companies face various types of fraud, like phantom billing, up-coding, and identity theft. Detecting such frauds is thus of vital importance to reduce and eliminate corresponding financial losses. We used an unsupervised data mining algorithm and implemented an outlier detection model to assist the experts in detecting medical prescriptions suspected of fraud. The implementation ran medicine code, patients' sex, and patients' age variables through three successive screening steps. The proposed model is capable of detecting 25% to 100% of cases violating the standards for some medicines that are not supposed to be prescribed at the same time in one single prescription. This model can also detect medical prescriptions suspected of fraud with a sensitivity of 62.16%, specificity of 55.11%, and accuracy of 57.2%. This paper shows that data mining can help detecting potential fraud cases in medical prescriptions more quickly and accurately than by the manual inspection as well as reducing the number of medical prescriptions to be checked which will result in reducing investigators heavy workload. The results of the proposed model can also help policymakers to plan for fighting against fraudulent activities.

Keywords Fraud · Unsupervised data mining · Medical prescription · Medical insurance

Introduction

Fraud is defined as 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 [1]. Nowadays, different areas such as banking, health care, and telecommunications suffer massive financial losses because of fraudulent activities, so detecting such activities is of vital importance for them. In recent years, health care sector has become an attractive target for fraudsters due to its high amount of funds and financial resources [2, 3]. It is estimated that the loss by insurance companies and government agencies

✉ Mehdi Yaseri
myaseri@sina.tums.ac.ir

Mohammad Haddad Soleymani
m-haddadsoleymani@alumnus.tums.ac.ir

Farshad Farzadfar
f.farzadfar@ncdrc.info

Adel Mohammadpour
adel@aut.ac.ir

Farshad Sharifi
farshad.sharifi@gmail.com

Mohammad Javad Kabir
kabirmj63@gmail.com

[1] Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran

[2] Non-communicable Diseases Research Center, Endocrinology and Metabolism Population Sciences Institute, Tehran University of Medical Sciences, Tehran, Iran

[3] Department of Statistics, Faculty of Mathematics and Computer Sciences, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran

[4] Elderly Health Research Center, Endocrinology and Metabolism Population Sciences Institute, Tehran University of Medical Sciences, Tehran, Iran

[5] Health Management and Social Development Research Center, Golestan University of Medical Sciences, Gorgan, Iran

due to fraudulent health care transactions is about $100 billion, amounting to 10% of the nation's annual health care expenditure [4]. Also, The National Health Care Anti-Fraud Association (NHCAA) estimated that 3% of all health care spending ($68 billion) is lost to health care fraud in the United States [5]. Although no accurate estimate of money lost due to fraudulent activities has been reported, according to research, in most countries about 10 percent of health care costs is lost because of fraud [6]. Medical insurance companies, as one of the main subsets of health care section, face various types of fraud, which are committed by providers, insured people, and insurance companies themselves, who take unlawful money from these companies by making deceptive medical claims [3, 7]. Fraudulent activities committed by mentioned groups include activities such as phantom billing, up-coding, misrepresenting services and/or diagnoses, unbundling or exploding charges, payment or receiving kickbacks, self-referral, doctor shopping, identity theft, misuse of insurance card, falsifying reimbursements, and falsifying service statements [3, 4, 7, 8].

There are two general methods for fraud detection: 1) manual inspection, and 2) statistics-based methods [6]. With the manual inspection method, in order to identify fraudulent claims, experts have to audit a large number of documents based on specific criteria in a limited period of time. Although manual inspection is a very accurate way for fraud detection, applying it to a large volume of data is costly and time-consuming [5, 6]. For this reason, experts need to use advanced analytical tools to assist them in processing a large volume of data. One of the best available tools to do so are data mining techniques. Data mining is a relatively new concept that was introduced in the mid-1990s as a new approach to data analysis and knowledge discovery [9]. Data mining is the process of automatically discovering useful information in a large data repositories [10]. As a highly application-driven domain, data mining has incorporated many techniques from other domains such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, algorithms, high-performance computing, and many application domains [11]. According to research, applying data mining algorithms can improve fraud detection process, though excessive reliance on highly automated systems for fraud detection is not all of the solutions [1]. Therefore, it was suggested that regardless of the extent of automation, a representative sample of patients (their claims) should be tracked down as part of the fraud detection strategy [1].

Based on available information, we used an unsupervised data mining algorithm and implemented an outlier detection model to detect medical prescriptions suspected of fraud. To do so, several data sets were used which were stored in a relational database. These data sets included information about the insured people, prescribed medicines for them, as well as information about the medical providers.

Rest of this paper was organized as follows. "Methods" discussed the methods, which were developed based on six phases of Cross-industry Standard Process for Data Mining (CRISP-DM). "Results" presented the results of computational formulas using real data along with the model performance indices, which were the validations of the proposed model compared to an expert's opinion as well as the comparison of the results of the MM risk matrix with the standards for the relationship between medicines. Finally, we gave concluding remarks in "Discussion".

Methods

In order to systematically conduct data mining analyses, a general process is usually followed. Using the SAS®software, in this research, we implemented the CRISP-DM process, consisting six phases that are commonly used in data mining studies [12].

Phase 1: business understanding

As it was mentioned, fraudsters take unlawful money from medical insurance companies by making false and deceptive claims such as medical prescriptions. Since investigating all claims is a time-consuming task, a data mining algorithm was used to help experts to examine the documents more quickly and accurately to find out if they were fraudulent or not. We used the information of all medical prescriptions of the insured patients in 2013 who were covered by one of health insurance companies in Iran. In this study, it was assumed that fraudulent medical prescriptions were outliers in the investigated data set. These deviations from normality could be considered as "red flags for fraud" [8]. In this paper, the term "red flag for fraud" was used with the same meaning as the term "suspected of fraud".

Phase 2: data understanding

We used separate data sets which were stored in an electronic relational database. These data sets contained various information such as information about the insured people, prescribed medicines for them and also information about the medical providers. After checking these data sets, it was found that no information was available regarding the status of the medical prescriptions (fraudulent or legitimate). This was important for choosing the right algorithm and a suitable method for analysis.

Table 1 The characteristics of variables in the final data set

Variable	Type-scale	Description
Insurance ID	String-Nominal	Unique ID of the insured person
Sex	String-Nominal	Insured person's sex
Age	Numeric-Discrete	Insured person's age
Prescription ID	String-Nominal	Unique ID of each prescription
Provider ID	String-Nominal	Unique ID of each provider
Provider specialty	String-Nominal	The specialty of each provider
Medicine code	String-Nominal	Code of prescribed medicine
Medicine name	String-Nominal	Name of prescribed medicine

Phase 3: data preparation

In this phase, the available data sets were integrated and a suitable data set to be analyzed was created. Before doing so, the quality of every each variable in each data set was checked, i.e. invalid values were identified and corrected if possible. Records with invalid values that couldn't be corrected would be discarded before calculations. After this step, the data sets at hand were integrated and a data set containing the variables required for analysis was prepared. The characteristics of variables in the final data set is shown in Table 1. The final data set contained 156,529,417 records of medicines that were prescribed in 47,827,160 prescriptions. This information belongs to 13,668,229 insured patients in 2013, of which 6,579,454 records (48.14%) belong to women with the average age of 36.38 years and 4,828,229 records (35.33%) belong to men with the average age of 34.88 years (it should be noted that the sex values for the rest of the records were missing). Among prescriptions, the most were written by General Practitioners (16,407,902), Internists (2,708,555), Pediatricians (1,913,366), Obstetricians and Gynecologists (1,560,832) respectively and the least were written by Sports Medicine (453) and Gastroenterologists (774).

Phase 4: modeling

As it was mentioned before, no information on the status of the medical prescriptions (fraudulent or legitimate) was available in the investigated data set. Regarding this fact, we had to use an unsupervised data mining algorithm to identify medical prescriptions suspected of fraud. Also, assuming that fraudulent medical prescriptions were outliers in the data set at hand, we used a simple yet efficient outlier detection model to raise red flags for fraud in medical prescriptions. It should also be noted that when writing a prescription, providers have to consider the relationship between prescribed medicines as well as the fact that the medicines should be proper to patient's sex and age. Considering these facts, we used medicine code, patients' sex and age variables to detect medical prescriptions suspected of fraud. To do so, we followed a 3-step process [5]:

1. Computing Frequency Matrices
2. Computing Risk Matrices
3. Comparing risks with a threshold

In the first step, we computed frequency matrices for three sets of paired variables: Medicine-Medicine (MM), Medicine-Sex (MS) and Medicine-Age (MA). The elements of frequency matrices (f_{ij}) are the number of times the variable in the i^{th} row occurs against the variable in the j^{th} column. According to their row and column variables types, frequency matrices were grouped into two different types, qualitative and quantitative matrices. A qualitative matrix is the one in which both its rows and columns variables are of qualitative type. In contrast, a quantitative matrix is a matrix for which at least one of its rows and/or columns variables is of quantitative type [5]. After this step, the risk matrices were computed based on the type of each frequency matrix. The elements of risk matrices (r_{ij}) were considered to be the risk of a specific pair of variables occurring at the same time. Risk values are of quantitative type, continuous-scaled, and vary between 0 and 1. Risk values close to 1 demonstrate that it is less likely for that specific pair of variables to be seen together, on the other hand, risk values close to 0

Table 2 Medicines that interact and medicines of the same class

Medicine A	Medicine B	Status
Chlordiazepoxide	Olanzapine	Interaction
Chlorpromazine	Metoclopramide	Interaction
Gemfibrozil	Atorvastatin	Interaction
Alprazolam	Ketoconazole	Interaction
Amiodarone	Ondansetron	Interaction
Ofloxacin	Ciprofloxacin	Same Class
Doxycycline	Tetracycline	Same Class
Diltiazem	Verapamil	Same Class

indicate a pair of variables that are more likely to be seen together.

To calculate the risk for the elements of qualitative and quantitative frequency matrices, Eqs. 1 and 2 were used respectively [5].

$$r_{XY}(i,j) = \frac{exp(-\frac{f_{ij}}{max(i)}) - exp(-1)}{1 - exp(-1)} \qquad (1)$$

$$r_{XY}(i,j) = \frac{exp(-\frac{f_{ij}}{max(i)} \times (1 - \frac{D_i(j)}{R_i(j)})) - exp(-1)}{1 - exp(-1)} \qquad (2)$$

In Eqs. 1 and 2, the value of $max(i)$ is the largest element in the i^{th} row in the reference matrix. Its value demonstrates the most common pair in the i^{th} row of each of the frequency matrices. In Eq. 2, $D_i(j)$ is the absolute value of the difference between the values of the j^{th} column from their weighted mean in the i^{th} row and $R_i(j)$ is the range of the column variable in the i^{th} row [5].

$$D_i(j) = |j - V_i|$$

$$V_i = \frac{\sum_j j \times f_{ij}}{\sum_j f_{ij}}$$

$$R_i(j) = max_i(j) - min_i(j) \qquad \forall j : f_{ij} \neq 0$$

In the final step, we compared each element of the risk matrices with several thresholds consisting 0.80, 0.85, 0.90, and 0.95. After consulting with experts, we chose 0.90 as the most applicable threshold for which the calculated risk values should be compared with. Risk values greater than or equal to 0.90 indicated that the specific pairs of variables were suspected of fraud and risk values that were less than 0.90 indicated that the pairs of variables were legitimate. Using information gained from this last step, those prescriptions that consisted any pairs suspected of fraud were labeled to be "medical prescriptions suspected of fraud", i.e. a medical prescription was assumed to be suspected of fraud if it was labeled "suspected of fraud" based on at least one of the risk matrices.

Table 4 Model performance indices

Parameter	Estimate	95% Confidence Interval
Sensitivity	62.16%	(54.13 - 69.57)
Specificity	55.11%	(49.89 - 60.23)
Accuracy	57.2%	(52.82 - 61.47)
PLR[1]	1.385	(1.35 - 1.42)
NLR[2]	0.6865	(0.6575 - 0.7169)

[1] Positive likelihood ratio

[2] Negative likelihood ratio

Phase 5: evaluation

Prescribing improper and/or unnecessary medicines in one single prescription could be considered as evidence for fraudulent activities. To evaluate the performance of the model in detecting such prescriptions, the results obtained from MM risk matrix were used to check whether the existing standards about medicines that interact with each other and medicines of the same class were violated in investigated prescriptions. The name of some of the medicines that interact with each other and medicines of the same class are shown in Table 2 [13, 14]. Also based on the model results, we stratified the medical prescriptions into two strata, "suspected of fraud" and "legitimate" and took a random sample of size 500 (250 samples from each stratum) with simple random sampling without replacement (SRSWOR) method. After blinding, the selected sample was given to an expert for determination of the status of each medical prescription. Comparing the model results with expert's opinion, we calculated the Sensitivity, Specificity, Accuracy, Positive Likelihood Rate and Negative Likelihood Rate [15].

Phase 6: deployment

This model can detect medical prescriptions suspected of fraud more quickly and accurately than manual inspection, though in future it should be modified to fit the existing situation because what might be true today may not be true

Table 3 The results of using the model to demonstrate the status of medical prescriptions by paired variables

	Medicine-Medicine		Medicine-Sex		Medicine-Age	
The status of medical prescriptions	Count	Percentage	Count	Percentage	Count	Percentage
Suspected of fraud	36,745,376	76.83	46,044	0.10	3,401,177	7.11
Legitimate	11,081,784	23.17	47,781,116	99.90	44,425,983	92.89

a year from now. To deploy this model, the following points should be taken into account:

1. The medical prescriptions that were marked as "red flag for fraud" may not necessarily be fraudulent prescriptions, i.e. some other unintentional factors like providers' mistake in prescribing a specific medicine, patient's request for prescribing irrelevant medicines to their illness, errors in data entry etc., may have distracted the results. For this reason, medical prescriptions suspected of fraud should be examined by experts to confirm their status. The knowledge gained from manual examination would be helpful in improving the model performance.
2. The factors affecting the process of identifying medical prescriptions suspected of fraud (e.g. patterns of fraudsters' behavior, available data, used algorithms and available technology) are constantly changing during time. Considering this fact, the model should be revised and updated to fit the existing situation.

Results

After comparing the risk values with the threshold of 0.90, among 47,827,160 medical prescriptions, 37,098,280 records (77.57%) were identified to be suspected of fraud and the rest (22.43%) were considered to be legitimate medical prescriptions. The results of using this model to demonstrate the status of medical prescriptions are shown in Table 3. As we mentioned in the previous section, those prescriptions that were labeled "suspected of fraud" may not necessarily be fraudulent prescriptions and should be examined by experts to confirm their status.

Using the results of the risk values from MM risk matrix and the standards for the relationship between medicines, the model detected 100% of prescriptions as suspected of fraud in which Chlordiazepoxide was prescribed at the same time with Olanzapine. This equals to 100% when Chlorpromazine was prescribed with Metoclopramide, 50% when Gemfibrozil was prescribed with Atorvastatin, 83.33% when Alprazolam was prescribed with Ketoconazole, 50% when Amiodarone was prescribed with Ondansetron, 55% when Ofloxacin was prescribed with Ciprofloxacin, 87.5% when Doxycycline was prescribed with Tetracycline, and 25% when Diltiazem was prescribed with Verapamil. Based on the results, this model performed satisfactorily for detecting violations from these standards, which could be a sign of fraudulent activities.

According to the results of model performance evaluation, this model also performed acceptable with a sensitivity of 62.16%, specificity of 55.11%, and accuracy of 57.2%. The results of model performance evaluation are shown in Table 4 along with 95% confidence interval for estimated values.

We also computed the frequencies of red flags for fraud in medical prescriptions within providers' specialties groups. Identifying the groups with a high frequency of fraudulent suspected activities, it is possible to plan for dealing with such providers. According to the results, the highest frequency of medical prescriptions suspected of fraud were witten by General Practitioners (13,843,141), Internists (2,064,018), Pediatricians (1,502,579), Obstetricians and Gynecologists (1,076,087) respectively and the lowest frequency were written by Sports Medicine (170) and Gastroenterologists (571). These frequencies are shown in Table 5.

Discussion

As noted earlier, in recent years, medical insurance companies have lost large amount of money due to fraudulent activities. Using data mining techniques as an assistant tool can improve fraud detection process in these companies. With the help of data mining, it is possible to raise red flags for fraud in medical prescriptions more quickly and

Table 5 Prescriptions and prescriptions suspected of fraud by providers' specialty

	Prescriptions		Prescriptions suspected of fraud	
Providers' specialty	Count	Percentage[1]	Count	Percentage[2]
General Practitioner	16,407,902	34.31	13,843,141	84.37
Internist	2,708,555	5.66	2,064,018	76.20
Pediatrician	1,913,366	4	1,502,579	78.53
Obstetrician and Gynecologist	1,560,832	3.26	1,076,087	68.94
...	...	...	...	...
Gastroenterologist	774	0.002	571	73.77
Sports Medicine	453	0.001	170	37.53

[1] The percentage of written prescriptions of total number of prescriptions

[2] The percentage of prescriptions suspected of fraud within providers' specialty

accurately and give them to experts for detailed examination. For this purpose, we undertook the first large-scale fraud detection process using data mining algorithms over more than 47 million medical prescriptions in Iran.

Using an unsupervised data mining algorithm and implementing a model for outlier detection, more than 77% of investigated medical prescriptions were labeled to be “suspected of fraud”. This model could detect 25 to 100 percent of medical prescriptions that violated the standards for the relationship between medicines. Also computing the model performance indices, this model performed acceptable with a sensitivity higher than 60%, specificity and accuracy about 55% and 57% respectively.

Based on the results, using this model, medical insurance companies can detect medical prescriptions suspected of fraud in less time and with higher accuracy than manual investigation. It also helps to reduce the number of medical prescriptions to be checked, which will result in reducing investigators heavy workload. However it should be noted that the model should be revised and modified to fit in the existing situation due to the effect of unintentional factors (e.g. providers’ mistake in prescribing a specific medicine, patient’s request for prescribing irrelevant medicines to their illness, errors in data entry etc.) and also due to changes in factors that affect the process of identifying medical prescriptions suspected of fraud (e.g. patterns of fraudsters’ behavior, available data, used algorithms and available technology). This model can also help policy makers to identify fraudulent behavior patterns and prevent fraudsters from committing fraud in the first place and to impose penalties as a response to their illegal activities.

We recommend the following for future studies:

1) Using other unsupervised data mining algorithms to raise red flags for fraud in medical prescriptions, 2) Examine other variables at hand to find out if they are useful in detecting medical prescriptions suspected of fraud, 3) Using supervised data mining algorithms, based on the results obtained from this research, and 4) Using network analysis methods to identify fraud networks.

References

1. Arash R, Hossein J, Taryn V. No evidence of the effect of the interventions to combat health care fraud and abuse: a systematic review of literature. PloS one. Public Libr Sci. 2012;7(8):e41988.
2. Aral KD. Prescription Fraud detection via data mining: a methodology proposal. Ankara: Bilkent University; 2009.
3. Li J, Huang K-Y, Jin J, Shi J. A survey on statistical methods for health care fraud detection. Health care management science. Springer. 2008;11(3):275–287.
4. Medical Fraud Detection Through Data Mining. Megaputer intelligence. 2002.
5. Aral KD, Güvenir HA, Sabuncuoğlu İ, Akar AR. A prescription fraud detection model. Computer methods and programs in biomedicine. Elsevier. 2012;106(1):37–46.
6. Copeland L, Edberg D, Panorska AK, Wendel J. Applying business intelligence concepts to Medicaid claim fraud detection. J Inf Syst Appl Res. 2012;5(1):51.
7. Busch RS. Healthcare fraud: auditing and detection guide. New York: Wiley; 2012.
8. Baesens B, Van Vlasselaer V, Verbeke W. Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection. New York: Wiley; 2015.
9. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang J-F, Hua L. Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems. Springer. 2012;36(4):2431–2448.
10. Tan PN, Steinbach M, Kumar V. Introduction to Data Mining: Pearson Education. India: Chapter. 9; 2007, p. 624.
11. Han J, Pei J, Kamber M. Data mining: concepts and techniques. New York: Elsevier; 2011.
12. Olson DL, Delen D. Advanced data mining techniques. Berlin: Springer Science & Business Media; 2008.
13. Drug Interactions. Available from: https://online.lexi.com/lco/action/interact.
14. Same Classes Drugs Error in Single Prescription. Available from: https://www.drugs.com/.
15. Vihinen M. How to evaluate performance of prediction methods measures and their interpretation in variation effect analysis. BMC genomics. BioMed Central. 2012;13(4):S2.