How Simple Predictive Analysis of Health Care Claims Data can Detect Fraud, Waste and Abuse Threats in Health Care Insurance - The Case Study of United Arab Emirates

1st Kristijan Jankoski
Netcetera AG
Skopje, R. of N. Macedonia
kristijan.jankoski@netcetera.com

2nd Kiril Milev
Netcetera AG
Dubai, United Arab Emirates
kiril.milev@netcetera.com

3rd Gjorgji Madjarov
Faculty of Computer Science and Engineering
SS. Cyril and Methodius University,
Elevate Global LLC
Skopje, R. of N. Macedonia
gjorgji.madjarov@finki.ukim.mk

***Abstract*—The usage of unethical practices which does not follow prescribed clinical standards and leads to the unnecessarily high expenditure for health care (waste, abuse and fraud) is increasing day by day in the Middle East countries. Reports show that about 30% of health care companies expenditures are based on a fraudulent medical claim. The rule-based approaches and expert systems that are used traditionally for tackling the health care waste, abuse and fraud (WAF) are very limited and require experts with extensive knowledge of medicine and expertise in the domain itself. The predictive analysis can be more flexible and less susceptible to some of the problems encountered with rules-based systems by focusing on the outcomes rather than the entire decision making process. In this paper, we present how simple predictive analysis and unsupervised learning on health care claims data can be used for detecting waste, abuse, and fraud threats in health care insurance in UAE. Our focus is to detect abnormal behavior of the clinicians from different specialties from different medical providers using the patterns made on the diagnosis and activity level prescription. The results obtained from the experiments performed on over 370K medical claims showed that only 0.007% of the clinicians caused potentially over 10% of the WAF marked claims. 27 clinicians marked with the analysis and scored as being most suspicious by the auditors made total of 4.929 claims.**

***Index Terms*—health care, waste, abuse, fraud, machine learning, unsupervised learning**

I. Introduction

Health care insurance is a multifaceted industry that brings together care providers, insurance companies and patients. As the industry is expected to create social benefits, there is constant pressure to contain costs while providing security and improving the health of the general population.

Misuse of the health insurance is an ongoing issue. Motivated by the financial incentives, different stakeholders are creating waste, abuse the market or even commit fraud. The volume of waste, abuse and fraud (WAF) is estimated to be in the range of 5-10% of the yearly health care expenditure [1]. This makes WAF a significant contributor to the medical inflation. Insurance claims are under continuous scrutiny by the health care payers for being one of the key tools to control health care spending.

Looking from an insurance fraud detection technique perspective, WAF is being generated by both health care providers and insured members in the Middle East. In the worst cases, a conspiracy-type of fraud involves several parties colluding in the misuse. When looking at the complexity levels, we can categorize WAF in seven levels. This starts with single transaction as the simplest one and goes up all the way to multi-party, criminal conspiracies [2].

The most prevalent types of fraud carried out by policy holders are: gaining access to or being reimbursed for services typically not covered by the policy. For clinicians and health care providers, financial gain is the main motivation with up-coding, service unbundling, and billing for unnecessary or even not rendered services. Another problem behind the seemingly rampant issue of health insurance fraud is that many businesses and their employees do not know how it looks like action. The truth is that this scam exists in many forms and it is often horribly difficult to notice. According to a study by D. Thortnton et. al. [3] some of the most common types of frauds that are the most prevalent are:

Payment for more expensive treatments than necessary - otherwise known as 'upcoding'. This includes hyperbolizing the diagnosis in a much more serious condition, in order to increase the cost of the claim.

Counterfeiting of the diagnosis: essentially examining multiple diagnoses, in order to collect procedures that are not medically needed.

Refunds for services that have not been made: this type of fraud can be achieved by falsifying claims by using real patient information, creating a false claim from scratch or by supplementing any of the actual claims with procedures that are never happened.

Misrepresentation of unnecessary treatments: this is a request for an unnecessary procedure, for example scanning

the brain with magnetic resonance, as part of a medically necessary procedure for heart surgery.

Payment procedures: perhaps the most courageous form of fraud in health insurance, whereby medical professionals perform treatments over healthy patients exclusively for the purpose of submitting a claim.

Upcoding is one of the most expensive and most sophisticated types of waste, abuse and fraud in health care. Between 2002 and 2012, this method cost publicly funded medical assistance programs around $11 billion. These are not victims of crime because they put unnecessary efforts on the social security network over which millions of citizens rely on their basic medical needs.

The fraud itself is difficult to detect. Medical diagnoses, length of work visits and complexity of treatment are left at the discretion of the health care provider. Individual cases of this criminal behavior, or even small groups of them, may be almost impossible to find or prove. Scams may be even more prestigious or harder to discover in large institutions, such as laboratories or hospitals, which have a wide range of procedural options and where insurance recovery tends to be very loose.

In the United Arab Emirates currently premiums amount to over $9 billion dollars, and with the introduction of compulsory health insurance for expatriates in Dubai (95 percent of the workforce), that figure is constantly increasing. But there is additional factor that influence the increase of insurance premiums, the waste, abuse and fraud.

Insurance waste, abuse and fraud is nothing new, but levels that are currently thought to be taking place globally are shocking. The data from the Health Insurance Counter Fraud Group, a group of 32 health insurance companies, with a mandate to detect and prevent fraud in the health sector, suggest that global losses as a result of fraudulent claims are hundreds of billions of dollars.

In this paper, we show how simple predictive analysis and unsupervised learning on health care claims data can be used for detecting waste, abuse, and fraud threats in health care insurance in UAE. Our focus is to detect abnormal behavior of the clinicians from different specialties from different medical providers using the patterns made on the diagnosis and activity level prescription. In the following section, the most relevant related work is reviewed first, and then the use case and the health care insurance system of UAE is described. The analysis that was performed on the obtained data for indicating the potential threats of waste, abuse and fraud of the medical personnel is presented in section 3. This section also provides visualization of the obtained results and their interpretation. The concluding remarks are given in section 4.

II. Related Work and Problem Statement

Traditionally, detection of health care waste, abuse and fraud is based on archaic methods that is very limited and requires experts with extensive knowledge of medicine and expertise in the domain itself [4] [5]. It is based on the work of auditors who need to manually review and identify suspected medical claims, which is a very costly and time-consuming process. They have quite limited time to process each claim by following predefined rules and procedures on certain characteristics of a claim without paying attention of a provider's behavior. This process is supported by the rule-based engines and expert systems based on the information disclosed from the past events and findings in the research.

But with the development of electronic health records and advances in artificial intelligence, other opportunities for automatic dealing with fraud have been made. Machine learning can help with the knowledge ex-traction process and create models from thousands of medical claims that can identify a much smaller subgroup for further assessment by the auditors, which significantly increases the time that any expert can dedicate to claims while he inspects. In turn, this leads to a higher rate of detected fraud.

There are different data mining approaches that are used for addressing the problem of healt care waste, abuse and fraud. The most common and well-accepted categorization that is used in this domain divides the approaches into 'supervised' and 'unsupervised' [6] [7].

Supervised approaches use samples of previously known fraudulent and non-fraudulent records. They are quite successful in detecting already known patterns of fraud and abuse. Taking this into account, these models should be regularly updated to reflect new types of fraudulent behaviors and changes in the regulations and settings [4]. Examples of the supervised methods that have been applied to health care fraud and abuse detection include neural networks [8] [9], decision tree [10] [8], and Support Vector Machine (SVM) [11].

Unsupervised methods typically compare claim's attributes to other claims and determine the level of difference by measuring the distance from a concrete distribution of claims. They are able to select anomaly records or group of similar records. Examples of the unsupervised methods that have been applied to health care fraud and abuse are clustering [12] and outlier detection [13].

A. Shafafiya standard and data description

Abu Dhabi has implemented a standard for data exchange within the health sector called Shafafiya. Any electronic transaction between 2 health care institutions must be structured and exchanged according to the imposed standard. The purpose of such provisions from 'Shafafiya' is to establish a national standard for improving the efficiency and effectiveness of the health care system by simplifying the processes themselves. The various definitions, formats and data that health systems have used for decades are now united in one standard format (Figure 1) and thus the electronic transfer of information is far more efficient and reliable. This has improved the overall quality of data and significantly reduced the administrative burden. In this way, information management methods are simplified, giving health care professionals more opportunities to provide better care and reduce the costs of patients themselves.

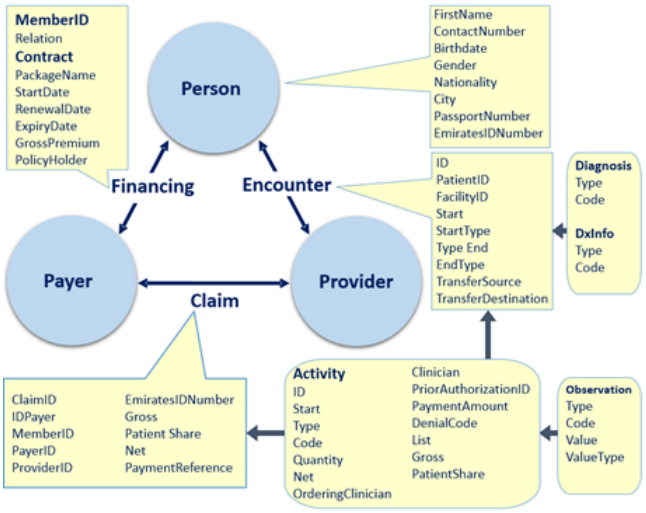

Fig. 1. Entities defined by the Shafafiya standard © Property of Department of Health (DOH)

The data that were used in this research satisfied the Shafafiya standard. We received 368.541 unique claims obtained from 1.567 medical providers for 2 years. The total number of prescribed activities on all claims were 1.224.259. The total number of investigated clinicians were 9.566 from different specialties (not given by the stakeholder). On all claims 12.384 diagnoses were detected. 35% of them were primary diagnoses, while the rest 65% were secondary diagnoses. All diagnoses were coded according to the ICD10 standard.

III. Analysis of Waste, Abuse, and Fraud Threats

Health care claims contain cleanly structured data elements that can be used as input for predictive modeling. These elements include information about the insured member with their medical condition, the medical procedures and services performed on the patient, the prescribed medications, time, date and location of the services, and others.

One representation of the problem space for waste, abuse and fraud in health insurance could be the multi-layer graph representation (Figure 2). Each data element in the transaction can be represented by one or more nodes in the graph. Nodes are linked to one or more of the other nodes, from the same or from a different type. Each edge in the graph has an assigned weight based on the relationship of the specific nodes. After determining the nodes and edges of the initial graph, additional layers can be added that represent different abstractions for the transactions. Once the problem is defined in this way, one should represent the nodes and edges using descriptive attributes and start the learning process of the machine learning based approach.

This iterative process of knowledge discovery uses a combination of different data analysis and visualizations. Conveying the information and the gained knowledge to an individual during the analysis, even when working with highly skilled actuaries, is one of the key tasks. Therefore, good data visualization is just as important as the data analysis.

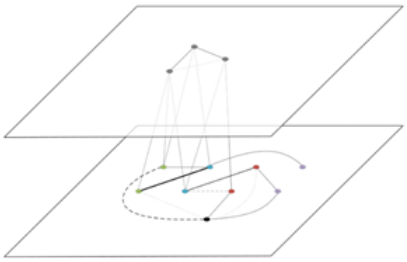

Fig. 2. Multi-layer graph representation of the entities and their interactions

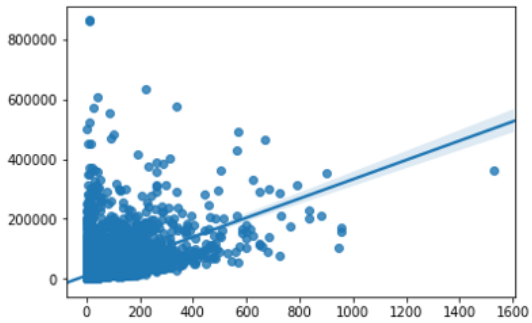

Fig. 3. Comparison

Defining the baseline in behavior analysis mandates proper analysis and definition of peer groups first. For example, pharmaceutical transactions (eRX or PBM) have very different features than inpatient visits (hospital stay). Depending on the transaction types, this classification can be straightforward. In absence of the elements containing the required information (due to the private and personal data limitation), however, a data-driven approach should be applied. Aggregations on a clinician level is another case. Figure 3 indicates the importance of identification of peer groups among the clinicians. It shows a comparison on all clinicians (represented as blue dots on the figure) on the number of unique activities that they have prescribed and the net amount on the claims that they have made. The big bang approach for solving this problem could lead us to many wrong assumptions and conclusions. Identifying the correct peer group for comparison is a critical step in the process. To address these kinds of problems, we use unsupervised data-driven approaches that try to find hidden patterns in the data based on a similarity measure using unlabeled data.

A. Peer groups identification using unsupervised learning

One of the most common methods for unsupervised learning is cluster analysis, which is used to analyze data previews to find hidden patterns or groupings in the data. Clusters are modeled using a similarity measure that is defined on a metric, such as Euclidean distance, probable distance, etc. Therefore, clustering can be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings depend on the individual set of data and the intended use of the results. The cluster analysis as

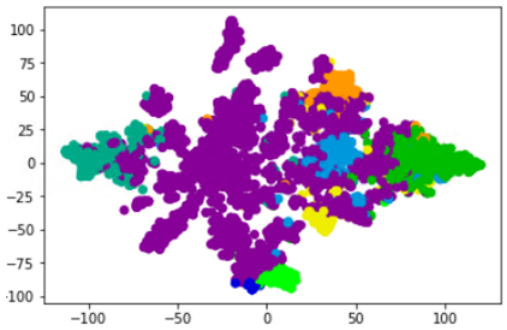

Fig. 4. t-SNE visualization of the results obtained from data-driven clustering

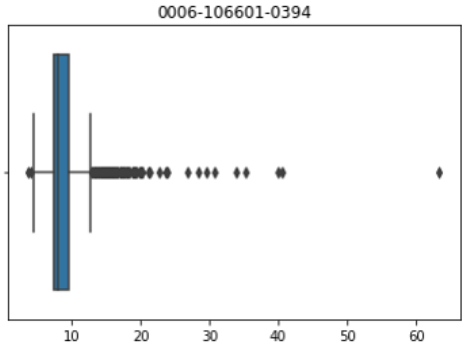

Fig. 5. Box-plot visualizations of the univariate models. Each dot represents a doctor and the amount of prescriptions he made on the 0042-114504-2481 activity

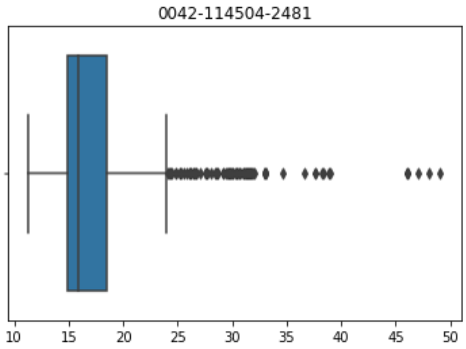

Fig. 6. Box-plot visualizations of the univariate models. Each dot represents a doctor and the amount of prescriptions he made on the 0006-106601-0394 activity

such is not an automated task, but an iterative process of knowledge discovery or interactive multi-target optimization that involves trial and failure. It is often necessary to preprocess the data and define the parameters of the model, until we achieve the desired result. One of the most commonly used clustering algorithms are: K-means and Hierarchical Agglomerative Clustering.

In our context, the use of such groups avoids the problem of comparing specialists in a medical field and doctors who are less professional. Therefore, each doctor who appears in the data set is represented as a rare vector of the frequency of the diagnoses he has prescribed, that is, the columns in such a vector are all unique diagnoses that occur in the set of data.

Then, the vectors from all doctors are concatenated into a single matrix, over which the two clustering algorithms applied. The best results are obtained using hierarchical clustering, and they are presented on Figure 4 in two dimensional space by reducing the dimensionality with t-distributed stochastic neighbor embedding (t-SNE) [14].

B. Univariate analysis

Univariate models are one of the simplest forms of data analysis. They don't deal with causes or relationships and the main goal is to describe the data. With the help of this type of model, we want to answer one of the following questions:

- How many times did the doctor prescribe a certain activity compared to other similar doctors?
- How much did a doctor earn by prescribing a particular activity compared to other similar doctors?
- How much activity has been prescribed for a given diagnosis compared to other prescribed activities?
- How much do the doctor prescribe for a diagnosis compared to other similar doctors?
- How many doctors diagnose a diagnosis compared to other similar doctors?

An example answer to some of these questions is presented on the Figures 5 and 6, where it can be clearly seen which doctors don't follow the trend of their group.

C. Bivariate analysis

Bivariate analysis is a simultaneous analysis of two variables. Here, the concept of the relationship between these variables is being investigated, regardless of whether there is any association between them, as well as the strength of that association, or whether there are differences between those variables and their significance. It is important to note that in our context, this type of analysis can detect a trend only if the data are well grouped. Using these models, the trends in 3 types of models were analyzed:

- The number of prescribed activities against the quantity of activities (Figure 7)
- The number of prescribed activities against the net amount of claims (Figure 8)
- The number of claims against the net amount of claims (Figure 9)

Then the relationship between these two variables is modeled using a linear predictive function whose unknown parameters are estimated from the data. Finally, we define a metric for the abnormality of the doctor according to this model as the z-score of the distances of doctors from the linear function Figure 10.

This predictive analysis leads us to over 36K suspicious claims. 27 clinicians marked with the analysis and scored as being most suspicious by the insurer made total of 4.929 claims. 0.007% of the clinicians caused potentially over 10% of the WAF marked claims.

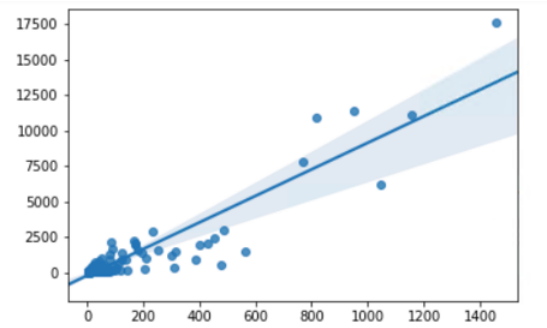

Fig. 7. Number of activities vs quantity of activities

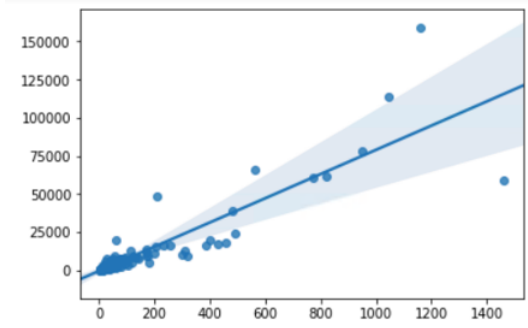

Fig. 8. Number of activities vs net amount of claims

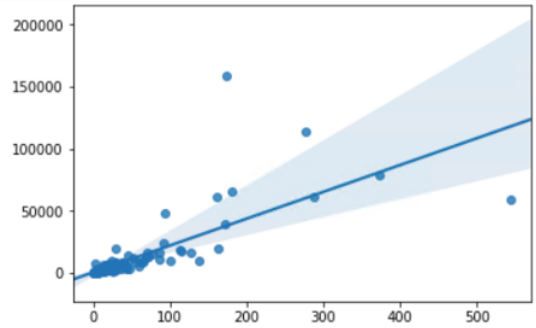

Fig. 9. Number of claims vs net amount of claims

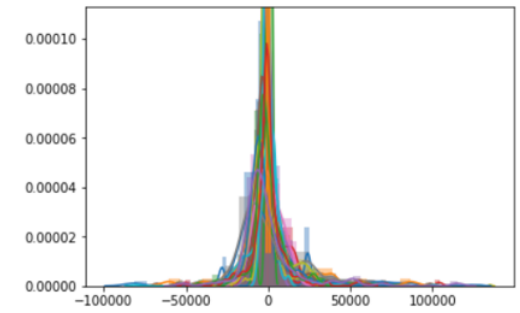

Fig. 10. Modified z-score distribution per peer group

IV. Conclusion

Health fraud significantly affects the ability of insurance companies to provide effective health care. Utilizing the power of machine learning can help detect fraudulent events, revealing perpetrators and reducing health care costs.

In this paper, we investigated various methods of machine learning to detect doctors who made claims for payment of activities that were not needed in the particular case. The results of the models applied in the practice of real data show that the data-driven methodologies used are very successful in dealing with this type of problem. They showed: over 36K suspicious out of 370K evaluated medical claims; 0.007% of the clinicians caused potentially over 10% of the WAF marked claims; 27 clinicians marked with the analysis and scored as being most suspicious by the insurer made total of 4.929 claims. It is also worth noting that predictive analysis is a much more efficient strategy than analyzing individual medical claims because it allows real-time detection over a large number of data and does not require supervision from a human auditor.

Current and further research will include improving the performance of existing models by using a larger number of data as well as using new methods based on supervised learning with a limited number of labels received by the auditors themselves. In addition, future research will dive deeper into assessing specific techniques for detection of anomalies based on other types of fraud in order to achieve a more automatic ranking and greater adaptability of the model. We hope this research will advance the state-of-the-art detection of fraud in health care and help address this important social challenge.

References

[1] J. Gee and M. Button, *The Financial Cost of Healthcare Fraud 2014: What Data from Around the World Shows*. BDO, 2014.

[2] D. Thornton, R. M. Mueller, P. Schoutsen, and J. van Hillegersberg, "Predicting healthcare fraud in medicaid: A multidimensional data model and analysis techniques for fraud detection," *Procedia Technology*, vol. 9, pp. 1252 – 1264, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2212017313002946

[3] D. Thornton, M. Brinkhuis, C. Amrit, and R. Aly, "Categorizing and describing the types of fraud in healthcare," *Procedia Computer Science*, vol. 64, pp. 713 – 720, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050915027295

[4] R. A., J. H., and V. T., "No evidence of the effect of the interventions to combat health care fraud and abuse: a systematic review of literature," *PLoS One*, vol. 7, pp. 2579–2605, 2012.

[5] J. Li, K.-Y. Huang, J. Jin, and J. Shi, "A survey on statistical methods for health care fraud detection," *Health Care Management Science*, vol. 11, no. 3, pp. 275–287, 2008. [Online]. Available: https://doi.org/10.1007/s10729-007-9045-4

[6] H. Joudaki, A. Rashidian, B. Minaei, M. Mahmoud, B. Geraili, and M. Nasiri, "Using data mining to detect health care fraud and abuse: A review of literature," *Global journal of health science*, vol. 7, p. 37879, 01 2015.

[7] R. Bauder, T. M. Khoshgoftaar, and N. Seliya, "A survey on the state of healthcare upcoding fraud analysis and detection," *Health Services and Outcomes Research Methodology*, vol. 17, no. 1, pp. 31–55, 2017. [Online]. Available: https://doi.org/10.1007/s10742-016-0154-8

[8] F.-M. Liou, Y.-C. Tang, and J.-Y. Chen, "Detecting hospital fraud and claim abuse through diabetic outpatient services," *Health Care Management Science*, vol. 11, no. 4, pp. 353–358, 2008. [Online]. Available: https://doi.org/10.1007/s10729-008-9054-y

[9] P. Ortega, C. Figueroa, and G. Ruz, "A medical claim fraud/abuse detection system based on data mining: A case study in chile," vol. 6, 01 2006, pp. 224–231.

[10] H. Shin, H. Park, J. Lee, and W. C. Jhee, "A scoring model to detect abusive billing patterns in health insurance claims," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7441 – 7450, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417412001236

[11] M. Kirlidog and C. Asuk, "A fraud detection approach with data mining in health insurance," *Procedia - Social and Behavioral Sciences*, vol. 62, pp. 989 – 994, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877042812036099

[12] T. Ekin, "Application of bayesian methods in detection of healthcare fraud," *Chemical Engineering Transactions*, vol. 33, 01 2013.

[13] D. Thornton, G. van Capelleveen, M. Poel, J. Hillegersberg, and R. Mueller, "Outlier-based health insurance fraud detection for u.s. medicaid data," *ICEIS 2014 - Proceedings of the 16th International Conference on Enterprise Information Systems*, vol. 2, pp. 684–694, 01 2014.

[14] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.