Detecting Frauds in Healthcare*

1 Introduction

The problem of fraud, waste and abuse (FWA) in prescription of drugs is a big problem in the healthcare industry. In the US, yearly costs of such undesirable prescription drugs are approximately 1% of the costs of total prescription drugs dispensed and amount to a massive $3 billion [1]. A pharmacy fraud can be committed by physicians, patients, pharmacists, prescribers among others. Identifying such malpractices can often be hard considering the ever-evolving nature of risks. Traditionally human investigators use inspections and heuristic rules to identify frauds. However, growing size of data and novel methods of frauds have proved human methods to be unfeasible.

In order to resolve the above problem, anomaly and pattern detection are common methods used to identify pharmacy frauds and abuse. Claim-volumes, claim-costs and early-refills are some key metrics which can be used to examine and classify transactions as frauds. Intuitive visualizations are often used to identify anomalous actors in the system.

Graph analysis has emerged as a popular method to identify network-based frauds. This facilitates identification of suspects based on colluding prescribers, pharmacists and patients as seen in social-network charts [2]. However, such analysis requires data points on individual prescribers, pharmacists as well as patients. Since the patient data is protected by strict state laws and cannot be accessed outside of healthcare services, the analysis in this study focuses on a more conventional technique of anomaly detection which only uses publicly available provider or prescriber data.

We use[#] k-means clustering algorithm to identify a suspected cluster of prescribers out of all the prescribers registered in the United States. A Tableau dashboard[*] is designed wherein the visuals are created to solidify the claim that the identified prescriber cluster is, in fact, outlier and can therefore be investigated in depth, rather than the traditional approach of manually vetting all prescribers. The dashboard even pinpoints outlier prescribers for inspection. Remaining of the paper is organised as follows. Section 2 contains the description of data used, the tasks to be addressed, approach followed, and the visual encodings used. In section 3, we conclude with a discussion on the novelty of the visualization and its strengths and weaknesses.

2 The Visualization

2.1 Data

Centers for Medicare and Medicaid Services (CMS) provide information on prescription drugs prescribed by health care prescribers, called as the Part D Prescriber Public Use File (PUF). Each prescriber is identified by a National Provider Identifier (NPI) and the records provide details of specific prescription advised at their discretion and paid for under the Medicare Part D Prescription Drug Program. The data is also provided at aggregated level for each healthcare prescriber. We use this Part D Prescriber Summary Table [3] in the study.

This table has more than One Million rows and has 81 features. The set of features is heterogenous and understanding the data requires domain knowledge. Fortunately, CMS provides a data-dictionary [4] for this purpose. Of these features, following are the relevant ones to this study.

The dataset contains counts of overall claims, 30-day standardized fill-counts, drug costs and beneficiary counts organised by each prescriber. Drug utilisation is classified on three further criteria. The first is the drug maker type- brand drug, generic drug and others. The second criterion is the class of drug- opioid, long-acting opioids, antibiotics and antipsychotics in the elderly. The third division is based on income level of beneficiaries- low-income subsidy and non-low-income subsidy claims. We hypothesize that these three sub-classifications of drug usage are potential markers for anomalous prescribers. Another feature, average beneficiary risk score, gives the average risk score of all beneficiaries going to a prescriber. All these features are continuous numerical.

*Visualization:
https://public.tableau.com/views/DetectingFraudsinHealthcare/Story1?:display_count=y&:origin=viz_share_link

[#]GitHub Repository: https://github.com/mohannishant6/Detecting-Frauds-in-Healthcare

The data also contains features describing the prescriber such as prescriber's specialty description, medicare-enrolment status, city, state, country. These are all nominal categorical features.

2.2 Task

- We wish to present the current overview of prescribers. For this, the task is to identify distribution of prescribers across the country and their major specialty categories. Another task is to analyse the proportion of different categories of drugs dispensed across the three criteria as mentioned previously. Finally, the distribution of patients and their risk score needs to be studied.
- We intend to cluster Healthcare Prescribers on the basis key metrics such as claim counts, costs incurred, patient risk score and others. Therefore, the task is to compare each cluster, analyse differences between metrics, and identify anomalous cluster(s).
- Within an identified cluster, the task is to compare the costs across specialty categories and states. Another task is to identify any correlation between Patient Risk and Costs. The final task is to detect outlier prescribers.

We hope to identify confounding prescriber cluster and prescribers from the visualizations. This set of prescribers can then be audited manually instead of auditing millions of prescribers, thus reducing the human efforts.

2.3 Approach

We use[#] k-means algorithm to cluster the prescribers based on the numerical features mentioned in previous section and identify optimum number of clusters using the elbow method [5]. The algorithm provided five clusters, marked as 0 through 4. Once we have tagged each prescriber with a cluster number, we import the data in Tableau and study the characteristics of each cluster. The visualization dashboard has three major sections:

- View 1: We begin with a general overview of the prescriber population- their distribution across the country, major speciality categories, average patient risk score, overall drug costs across various categories such as opioids, antibiotic drug costs and others.
- View 2: Next page focuses on the clusters. The visuals provide information on number of clusters formed, the overall prescriber counts in each cluster, as well as the distinguishing features of the clusters.
- View 3: The last section helps drill down into a cluster of choice and identify outlier prescribers.

2.4 Encodings

- View 1:
  - In the map, color encodes density of prescribers in a state
  - In the bubble chart, size of bubble encodes the count of prescribers in each specialty category
  - In donut charts, colors encode the categories as mentioned and angle encodes corresponding values of each category
  - In the bar chart towards the bottom right, size and color encode count of patients.
- View 2:
  - Colors encode each cluster for the overall counts as well as all the charts
  - Size of bars encode values for each category
- View 3:
  - In the Heat map, color encodes Average Drug Costs
  - Each circle in the scatter plot encodes a Prescriber
  - The color of circle in the scatter plot encodes the cluster number, which is same as in View 2
  - Position of circle encodes the values of Patient Risk Score and Total Drug Cost incurred by the prescriber.

*Visualization:
https://public.tableau.com/views/DetectingFraudsinHealthcare/Story1?:display_count=y&:origin=viz_share_link

[#]GitHub Repository: https://github.com/mohannishant6/Detecting-Frauds-in-Healthcare

3 Conclusion

In this work, prescriber-level data obtained from CMS was used to identify potentially fraudulent healthcare prescribers. Although, the data used in this work has been studied previously using supervised [6] and unsupervised learning [7], to the best of our knowledge, this data and the resulting prescriber clusters have not been visualized in a way that the present work does. Hence the Tableau dashboard presented is a novel piece of work.

View 1 of the dashboard aims at educating an analyst or viewer of the data and overall healthcare situation in the US, it is an exploratory visualization. After becoming familiar with the data, in View 2 we wish to visually communicate the differences between the formed clusters and the features used. As such, View 2 can be considered explanatory. The end objective of identifying one or more confounding or outlier prescriber is achieved in View 3. Hence, View 3 is again an exploratory visualization.

The strength of the presented visualization lies in successfully achieving the main task of identifying suspected prescribers. In the process, the analyst or viewer also get to appreciate the complexity of data in the field of healthcare, such as costs and patient risk scores. However, the robustness of this method cannot be guaranteed as such. For instance, it is possible that a prescriber, thus identified through the visualization, is one that caters only to high cost-incurring patients/cases only. Such a prescriber will be incorrectly identified through the proposed visualization. At the same time, it can be argued that given other features, an analyst or viewer who is adept in the healthcare field would be able to judge if the prescriber is, in fact, a confounder.

References

[1] "Pharmaceutical Fraud Solutions - Elder Research," [Online]. Available:. https://www.elderresearch.com/industries/pharmacy-fraud-analytics

[2] Liu J., Bier E.: Graph analysis for detecting fraud, waste, and abuse in healthcare data. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (2015) 3912–3919

[3] "Medicare Provider Utilization and Payment Data: Part D Prescriber Summary Table CY2017," [Online]. Available: https://data.cms.gov/Medicare-Part-D/Medicare-Provider-Utilization-and-Payment-Data-Par/psut-35i4

[4] "Part D Prescriber PUF Methodology," [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Prescriber_Methods.pdf

[5] "Elbow Method," [Online]. Available: https://www.scikit-yb.org/en/latest/api/cluster/elbow.html

[6] Herland, Matthew & Khoshgoftaar, Taghi & Bauder, Richard. Big Data fraud detection using multiple medicare data sources. Journal of Big Data. (2018) 5.

[7] Stephen F., Olumide O., Sadiku J., Jacob A.: Application of Data Mining Technique for Fraud Detection in Health Insurance Scheme Using Knee-Point K-Means Algorithm. Australian Journal of Basic and Applied Sciences (2013) 7(8): 140-144

*Visualization:
https://public.tableau.com/views/DetectingFraudsinHealthcare/Story1?:display_count=y&:origin=viz_share_link

#GitHub Repository: https://github.com/mohannishant6/Detecting-Frauds-in-Healthcare