Automatic Detection of Providers with Excess Healthcare Spending

Jiacheng Liu

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science in Computer Science and Systems

University of Washington

2018

Reading Committee:

Dr. Martine De Cock, Chair

Dr. Anderson Nascimento

Dr. Shanu Sushmita

Program Authorized to Offer Degree:
Computer Science and Systems

University of Washington

Abstract

Automatic Detection of Providers with Excess Healthcare Spending

Jiacheng Liu

Chair of the Supervisory Committee:
Professor Dr. Martine De Cock
Institute of Technology

This thesis aims to develop techniques to help large hospital systems to detect providers with excess spending. Identifying fraud, waste, and abuse resulting in superfluous expenditures associated with care delivery is central to the success of these large hospital systems and for making the cost of healthcare sustainable. In theory, such expenditures should be easily identifiable with large amounts of historical data. However, to the best of our knowledge there is no data mining framework that systematically addresses the problem of identifying unwarranted variation in expenditures on high dimensional claims data using unsupervised machine learning techniques. In this thesis, we propose methods to uncover unwarranted variation in healthcare spending by automatically extracting reference groups of peer-providers from the data and then detecting high cost outliers within these groups. Besides we also implement existing graph based techniques and compare the results with our methods. We demonstrate the utility of our proposed framework on datasets from a large ACO (Accountable Care Organization) in the Pacific Northwest of the United States to successfully identify unwarranted variation in the provision of therapeutic procedures that had previously gone unnoticed.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to prof. Martine, prof. Anderson, prof. Shanu, prof. Ankur, dr. Muhammad, dr. Carly, dr. Greg and Karthik. Without your consistent help and insightful advice, it would have been impossible to complete this thesis. Special thanks to Kensci Inc. in Seattle who provided me with this great research opportunity to study real world data. Working with so many intelligent and diligent people in Kensci has been my great pleasure and honor. Everyone has been very supportive and friendly. It will always be one of my most beautiful memories in Seattle.

DEDICATION

to my grandpa Yagu Chen

Chapter 1

INTRODUCTION

Excessive spending and waste is a perennial problem in healthcare in the United States. One of the strategies to tackle this problem is to pinpoint providers with excessive spending and to infer actionable insights from the structure of expenditure variations. Since many cost variations are explainable and justifiable by provider specialties and patients conditions, it is never an easy task even for human experts. In this thesis, we explore the possibilities of automatically identifying suspicious high cost providers utilizing unsupervised machine learning techniques.

1.1 Background

Healthcare expenditures in the U.S. exceed $2 trillion a year [19]. According to The Commonwealth Fund, the U.S. spend far more on healthcare than all other high-income countries [26]. Yet, among comparably wealthy nations, the U.S. rank lowest in terms of quality of care, resulting in poorer health outcomes [26]. Additionally, it is estimated that unnecessary spending accounts for 20% to 30% of the total medical expenditures in the U.S. [4]. Such facts necessitate solutions that can reduce inefficiencies in the healthcare system while improving care and reducing costs. To meet this challenge, Accountable Care Organizations (ACOs) were developed in alignment with the Patient Protection and Affordable Care Act as a way to incentivize the quality of care that healthcare systems and provider organizations deliver to Medicare beneficiaries [24]. ACOs are part of the CMS (Centers for Medicare & Medicaid Services) alternative-based payment program which considers innovative ways to manage Medicare patients to establish the "Triple Aim" of healthcare: improved patient satisfaction, improved population health, and lower growth expenditures. There are more than

nine hundred ACOs[1] in the U.S., both in the public and private sector, serving millions of patients across the country in a process to transition from fee-for-service to a value-based-care model for healthcare delivery in an effort to contain expenditures.

1.2 Problems and Challenges

In this thesis, we focus on identifying providers with suspicious excessive spending in terms of per-patient cost in an ACO. In theory, such high cost providers should be easily identifiable with large amounts of historical data. However, to the best of our knowledge there is no data mining framework that systematically addresses the problem of identifying unwarranted variation in expenditures on high dimensional claims data. Fortunately, with the advancement of machine learning and data mining techniques, it is possible to do that: the availability of large and well-structured data sets of claims and clinical information makes it possible to analyze variation of cost and care at scale. All stakeholders in the healthcare system, including patients, providers, and payers of healthcare, can benefit from such solutions that attempt to reign in the costs of care without compromising quality.

Accurately identifying these providers with high per-patient-spend can assist system administrators in their role, however there are many challenges:

- **High false positive rate.** Excess spending is not necessarily suspicious. It could signal many things: appropriate care for medically more complex and sicker patients, care fragmentation, wasteful or even fraudulent spending. The presence of a mix of justifiable and non-justifiable reasons for high spending, and the difficulties of automatically inferring which one is at play, easily leads to high false positive rates.

- **Heterogeneous data.** Healthcare data are usually heterogeneous, containing both numerical and categorical values. Features like a patient's age and number of visits are

[1]Muhlestein, D. et al. 2017. Growth of ACOs and alternative payment models in 2017. Health Affairs Blog. https://www.healthaffairs.org/do/10.1377/hblog20170628.060719/full/, Accessed Mar 3, 2017.

integers, while race and diagnosis codes are categorical variables. Effectively leveraging both numerical and categorical variables is definitely a challenge.

- **Lack of labels.** Labels are hard to obtain. If a provider were not consistently and systematically billing for more money, it would take lots of time to find out decisive evidence. Considering the huge amount of claims, it is not cost-effective to manually investigate every claim and every patient of every provider. The lack of ground truth labeled data poses challenges both for the learning process (one has to rely on *unsupervised* machine learning, as we do in this thesis) and for the evaluation (one has to rely on domain experts to verify the results generated by the machine learning algorithm).

1.3 Methods and Results

The word "excess" in "excess spending" intrinsically implies a threshold of costs, prompting the question "what constitutes abnormal provider costs?". A simple solution is to examine a histogram of providers' costs and study the top $k\%$ of high cost providers. Another solution is to examine the distribution of cost data and identify the provider outliers based on their deviation from the mean. The baseline method utilized in this thesis integrates these approaches and flags any providers above *upper inner fences* (UIF) as outliers. However, the threshold of "high cost" is a context-related concept. For example, it is inappropriate to compare the median patient cost of an oncologist with the median patient cost of an ophthalmologist, because the therapeutic treatments and procedures commonly associated with each of these patient cohorts may be quite different. To address this problem of appropriate relative spend, in Chapter 5 we propose a method, referred to as the *provider-centric method*, which automates the process of creating reference groups by clustering providers whose patients have similar diagnosis codes. Then, within each cluster, we identify abnormally high cost providers. It should, however, be noted that there may be additional confounders that are associated with provider spending that are not addressed in this analysis. Thus, patients with serious and complicated conditions are expected to cost more. To address such cases,

we propose a second method, referred to as the *patient-centric method*, in which we cluster patients by their medical history and demographic data. Within each patient cluster, we examine all associated providers to determine which are responsible for any abnormally high per-patient spend.

We contrast our methods reference group based methods from Chapter 5 with a recently proposed graph based approach to healthcare fraud detection that extracts bipartite patient-provider graphs from healthcare claims data and identifies suspicious providers based on abnormal patterns in these graphs (see Chapter 4).

Utilizing data from a large ACO in the Northwestern U.S. (see Chapter 3), we used all methods from Chapter 4 and 5 to examine medical claims that occurred from January 1, 2016 to June 30, 2016, reflecting the care of more than 28,000 patients. As an important part of our analysis, we discussed the results with healthcare domain experts and identified two significant billing behavior patterns associated with high-cost providers. In particular, we were able to demonstrate the utility of our proposed framework by successfully identified unwarranted variation in therapeutic procedures even in low cost claims that had previously gone unnoticed. Additional sensitivity analyses were conducted to determine the impact of varying the number of clusters and other model parameters on the provider outliers detected.

The major contributions of this thesis are as follows:

- We propose two new methods – a provider-centric method and a patient-centric method – to automate the detection of excess spending which could indicate the need for further investigation by healthcare administrators of a large ACO.

- We compare our methods with a recently proposed approach that leverages patient-provider bipartite graphs.

- The application of the proposed techniques uncovered billing patterns corresponding to abnormal provider behavior which previously could not be detected by the rule based systems commonly employed. These results may provide opportunities for administrators

to intervene on excess spending.

This thesis is organized as follows: Chapter 2 summarizes related work on healthcare anomaly detection and general unsupervised anomaly detection techniques. Chapter 3 provides a description of the data and terminology that is used in later chapters. Chapter 4 includes a description of the graph based approach to outlier detection, and the results when applied to the data from Chapter 3. Our proposed methods are introduced in Chapter 5, along with their results when applied to the data from Chapter 3. Finally, in Chapter 6 we summarize our findings and discuss possible directions for future work.

Part of the results from this thesis have been accepted for presentation:

- **Eric (Jiacheng) Liu**, Muhammad A. Ahmad, Carly Eckert, Anderson Nascimento, Martine De Cock, Karthik Padthe, Ankur Teredesai, Greg McKelvey. *Automatic Detection of Excess Healthcare Spending and Cost Variation in ACOs*, in: DMMH2018 (SDM 2018 Workshop on Data Mining for Medicine and Healthcare), 2018

Chapter 2

RELATED WORK

A variety of methods have been proposed to address the general problem of anomaly detection in data, based on statistical methods and data mining methods. We refer to the survey paper by Chandola et al. [6] for a nice overview. There is also a rich literature on the automated discovery of fraud and anomalies in healthcare settings in particular. Overviews are given in the survey papers by Jing et al. [11] and Phua et al. [7]. Because of a usual lack of ground truth labels, the use of supervised machine learning for health insurance fraud detection is not very common. For an exception, we refer to the work by Y. Shi et al. [30] who treat the task as a classification problem and apply decision trees and naive Bayes classifiers. Since the data available for our study does not include ground truth labels (see Chapter 3), in this thesis we can not rely on supervised machine learning techniques. For this reason, in the remainder of this chapter, we focus on the existing use of unsupervised and semi-supervised techniques for anomality detection in general, and in healthcare data in particular. Existing approaches include graph based, instance based, and cluster based techniques, which we describe next.

2.1 Graph Based Approaches

Anomaly detection in graph data is an active area of research. Table 2.1 has a brief summary of the literature. For static plain graphs, feature based approaches [3] and proximity based approaches [8] are adopted. Sun et al. proposed a page-rank-like algorithm to determine the outlier score for each node in a graph [27]. For attributed graphs, structured based approaches are very popular [9, 22, 29]. For a recent survey about different kinds of graph based anomaly detection techniques we refer to the work of Akoglu et al. [2].

Algorithm	Weighted Graphs	Unweighted Graphs	Attributed Graphs	Plain Graphs	Parameter-free
OddBall [3]	✓	✓	✗	✓	✓
Sun et.al. [27]	✓	✓	✗	✓	✓
Dai et.al. [8]	✓	✓	✗	✓	✓
SUBDUE [22]	✗	✓	✓	✗	✗
SUBDUE [29]	✗	✓	✓	✗	✗
SUBDUE [9]	✗	✓	✓	✗	✗

Table 2.1: Brief Summary of Graph-based Anomaly Detection Literature

When utilizing a graph based approach for anomaly detection in data, it is common to first convert the data into a graph, such as a bipartite patient-provider graph in the healthcare data analytics setting, and then examine the graph to detect anomalies within that graph structure [2]. To this end, features are extracted for each node, such as the number of nodes in the neighborhood or the entropy. Nodes with feature values above or below a threshold are flagged as outliers, leading to the detection of fraud, waste, and abuse in healthcare [3, 17]. In Chapter 4 we apply the graph based anomaly detection approach to identify outlier providers in the data described in Chapter 3.

2.2 Instance Based/Community Based Approaches

Konijn et al. proposed a subgroup discovery tool *Cortana*[1] that can be used for healthcare fraud detection [13, 14]. When investigating a specific provider, the tool assists in identifying local subgroups of patients such that the difference of quality measures between reference groups and these local subgroups is maximized. To this end, every patient is represented by a feature vector and a binary label. The feature vector indicates the treatments that the

[1] `http://datamining.liacs.nl/cortana.html`

patient has received and is used to calculate the k-nearest neighborhood. The binary label indicates whether or not the patient has visited the provider being evaluated, and serves as part of the quality measure of detected subgroups, as does the cost.

A distinguishing aspect between this approach on one hand, and the graph based and cluster based approaches described in Section 2.1 and 2.3 on the other hand, is that in the approach proposed by Konijn et al. [13, 14] providers are investigated one by one, on an individual basis, while in the graph based and cluster based approaches all providers can be simultaneously investigated. Indeed, in the approach by Konijn et al., the value of the binary label for each patient changes depending on which provider is under investigation, and so for each provider, the analysis has to be repeated. The subgroup discovery method by Konijn et al. does not scale well to scenarios where 1000s of providers have to be investigated, such as in our study, which is why we do not consider it further in this thesis.

2.3 Cluster Based Approaches

Existing outlier detection techniques which are most relevant to the methods proposed in Chapter 5, are the cluster based approaches. Hu et al. [10] proposed a framework for detecting patients with an extremely high number of healthcare visits. In the first part of their method, they use a two-stage clustering algorithm to identify typical prototypes and generate *clusters of patients* with similar utilization profiles, defined through the number of clinical visits of different types. In the second part, for each type of patient characterized by the Hierarchical Condition Categories (HCC) used in Medicare Risk Adjustment, Hu et al. utilize a regression model to estimate the expected number of visits for each patient. Statistical tests are applied to determine whether the resulting differences generated by these two methods are significant.

Work by other researchers focused on *provider based clusters*. Lin et al. [16] proposed a method to cluster general physicians and then characterize clusters with the help of domain experts. In this approach, physicians were clustered based on utilization features such as the total cost per patient visit, number of surgical cases, and average treatment fee per case. Paulo et al. [23] clustered physicians based on billed procedure codes. Both Lin et al. and

Paulo et al. targeted physicians which were identified by abnormal practice behaviors and high per-patient cost. These studies were not restricted to specific disease cohorts nor limited by patient or provider size.

Our work is also related to the methods used by Titus et al. [28] for unsupervised identification of common co-occurring pharmaceutical utilization and patient surgical events in electronic medical record data. Titus et al. used a vector-space model approach to represent patients in the vector space of Current Procedure Terminology (CPT) codes and latent semantic analysis to reduce the dimensionality. In Chapter 5, we utilize a similar vector space model to represent providers and patients as vectors in the vector space of Clinical Classification Software (CCS) codes, thereby capturing diagnostic features of patients.

In healthcare spending anomaly detection problems, the ground truth may not be available in most instances (as in our study) and comparison metrics are ill-defined, making it extremely difficult to compare the performance of unsupervised methods. In Chapter 5 we present experimental results to show that the cluster based methods proposed in this thesis are able to correctly identify suspicious providers that would have gone unnoticed with previously proposed graph based techniques.

Chapter 3

DATA

This chapter describes the format and background distribution of the data. The healthcare claims data used in this thesis comes from a large hospital system in Pacific Northwest of the United States. It pertains to care provided to more than 28,000 patients by more than 8,000 providers in 2016.

3.1 Format and Coding

The schema and coding of the healthcare data used in this thesis are specified by CMS, the "Centers for Medicare and Medicaid Services". CMS is a federal agency within the United States Department of Health and Human Services (HHS) that administers the Medicare program. The claims data used in our research were made available to us in CMS Claim Line Feed (CCLF) format. The schema for CCLF files can be found in the reference manual on the CMS website[1]. While there are four parts in CCLF data, we are only concerned with CCLF part A and part B data, which are related to inpatient and outpatient claims respectively.

Disease diagnosis are encoded using ICD-10 codes, which is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organization (WHO). Clinical Classification Software codes (CCS codes) are aggregated upon ICD-10 to provide a higher level description of diseases and symptoms. We adopt the more coarse grained CCS codes instead of the very fine grained ICD-10 codes to characterize the diagnosis of a patient.

---

[1] `https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/MSSP-Reference-Table.PDF`

data	number of patients	number of providers	number of claims	total claim amount (million)
CCLF part A (inpatient)	26,444	3,483	159,579	$99
CCLF part B (outpatient)	26,283	7,374	244,073	$22
Total	28,496	8,146	403,653	$121

Table 3.1: Statistics on claims data from Jan 2016 through Jun 2016

The provider attribution logic is explained as follows. For patient claims related to CCLF part A (inpatient claims), the associated provider is derived from the "attending provider" field in the data. For patient claims related to CCLF part B (outpatient claims), the associated provider is derived from the "rendering" provider field. Each claim has only one provider associated with it.

3.2 Data Description

We analyzed healthcare claims data from a large ACO in the Northwestern United States. The claims pertain to services provided for patient care from January 1, 2016 to June 30, 2016. The dataset consists of 403,652 claims which include 28,496 unique patients and 8,146 unique providers. The total healthcare expenditures related to these claims is approximately $121 million dollars. The data used in this study consists of inpatient claims (CCLF part A) as well as outpatient claims (CCLF part B). Most patients have claims in both part A and part B. Inpatient claims are substantially more expensive than outpatient claims: as can be inferred from Table 3.1, part A accounts for 81.8% of the total cost while only accounting for 40% of the total number of claims. The data contains patient demographics such as age and gender. The average age of patients in the dataset is 67 and over 61% of the patients are female. Figure 3.1 shows the age distribution of patients.

Each claim has a unique claim ID specific to a particular patient and associated provider. We identified providers according to their practice taxonomy (e.g., cardiology) using the Na-

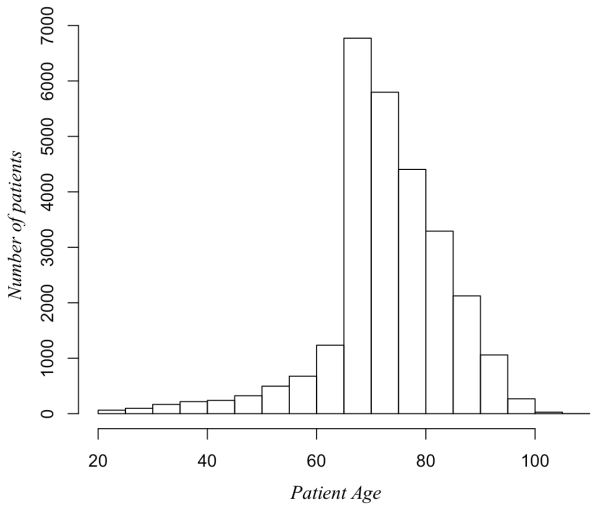

Figure 3.1: Age distribution of the 28,496 patients in the dataset

tional Provider Identifier (NPI), a unique 10-digit identification number issued to healthcare providers in the United States by the Centers for Medicare and Medicaid Services (CMS). We used the mapping logic provided by CMS[2] to map taxonomy codes, available in the NPI lookup, to specialty codes, which is a higher level of specialty categorization (e.g. internal medicine vs cardiology). During this process, if a provider was assigned two or more taxonomy codes, we only mapped his primary taxonomy code to a specialty. Thus, every provider has only one specialty. There are 63 unique provider specialties in the data and 54 of these have at least 10 associated providers. In Table 3.2, the top specialties are listed in terms of largest number of providers and in terms of average cost per patient respectively. "Internal Medicine" and "Family Practice" providers have the biggest patients group and they're top 2 expensive specialties in terms of the total cost of all of their patients. "Cardiac Surgery", "Neurosurgery" and "Hematology-Oncology" are top 3 specialties with the most expensive per patient cost, which is not surprising.

In addition to the patient demographic and provider information, each claim has patient

[2] `https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/MedicareProviderSupEnroll/Downloads/JSMTDL-08515MedicarProviderTypetoHCPTaxonomy.pdf`

Rank	Top 5 most frequent specialty	Number of providers	Number of patients
1	Internal Medicine	1,207	16,859
2	Diagnostic Radiology	765	13,449
3	Family Practice	726	14,163
4	Physician Assistant	626	9,846
5	Emergency Medicine	563	6,173
Rank	Top 5 most expensive specialty (per patient)	Average cost per patient	Number of patients
1	Cardiac Surgery	$ 11,824	163
2	Neurosurgery	$ 4,745	591
3	Hematology-Oncology	$ 4,399	1,410
4	General Surgery	$ 2,942	1,850
5	Orthopedic Surgery	$ 2,903	3,850
Rank	Top 5 most expensive specialty (total cost)	Total amount paid (millions)	Number of patients
1	Internal Medicine	$ 36.67	16,859
2	Family Practice	$ 17.99	14,163
3	Orthopedic Surgery	$ 11.18	3,850
4	Hematology-Oncology	$ 6.20	1,410
5	General Surgery	$ 5.44	1,850

Table 3.2: Provider specialties top 5

diagnosis information, encoded through ICD-10 codes. Although there can be multiple diagnoses per claim, each claim has a single primary diagnosis. ICD-10, also known as the International Statistical Classification of Diseases and Related Health Problems (ICD), is used by the World Health Organization (WHO) and has standardized medical diagnosis coding. As there are tens of thousands of ICD-10 codes, each related to a specific disease, as well as factors such as severity and chronicity, it is common practice to collapse these codes into larger groupings of diseases using Clinical Classification Software (CCS) codes. This process utilizes a mapping logic provided by CMS.[3] As opposed to the tens of thousands of ICD-10 codes, there are only 260 unique CCS codes in our data. In Chapter 5, we explain how we use these CCS codes to cluster providers and patients.

Finally, HCPCS (Healthcare Common Procedure Coding System) codes are used by health system administrators to encode the utilization of products, supplies, and services attached to a claim. For each claim in the data, all HCPCS codes billed by the provider can be identified. In this thesis, HCPCS codes are used to verify the results. Domain experts will manually check if the way how HCPCS codes are billed matches the diagnosis of a patient.

[3] `https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs_dx_icd10cm_2017.zip`

Chapter 4
GRAPH BASED OUTLIER DETECTION

Following [3] and [17], in this chapter, we apply graph based methods to detect outlier providers in the healthcare claims data described in Chapter 3. First, we build patient-provider bipartite graphs for the entire dataset, i.e. graphs in which all patients and all providers appear as nodes. Second, we extract features from the egonets of the providers in the graphs, and use them to detect outlier providers. A comparison of the experimental results obtained with the graph based methods from this chapter, and the reference group based methods from the next chapter, can be found in Chapter 5.

4.1 Patient-Provider Bipartite Graphs

Patient-provider bipartite graphs are built to extract features and detect outlier providers. In these graphs, patients and providers appear as nodes. An edge is added between a patient and a provider if the patient visited this provider at least once. There are two meaningful ways to calculate the edge weight:

- **Total.** The weight of an edge between a patient and a provider is defined as the total cost of that patient-provider pair, i.e. the total dollar amount associated with claims for that specific patient-provider pair.

- **Simple.** The weight of an edge between a patient and a provider is defined as the total number of claims found in the data for that specific patient-provider pair. Hereafter we refer to this weight assignment method as the "simple" approach, because it involves simple counts of the number of claims.

In Figure 4.1, we show such a graph constructed upon 1/10 of our data. Red dots are

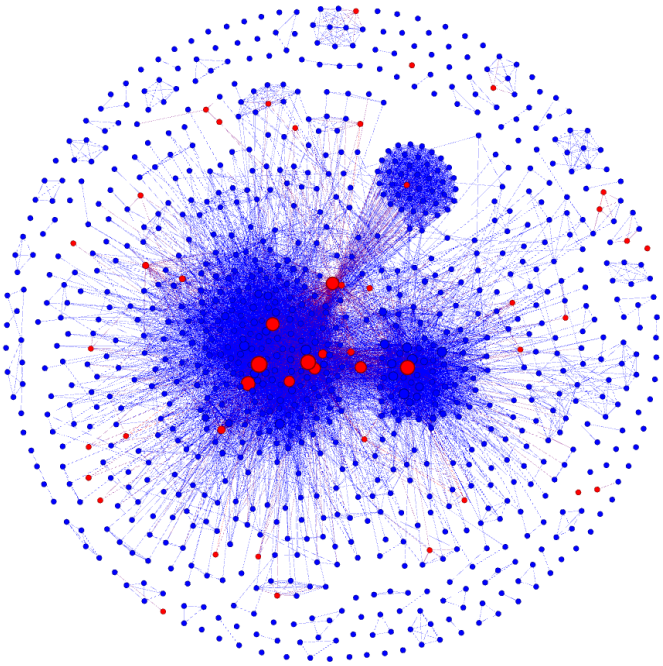

Figure 4.1: A provider-patient bipartite graph built on part of the data used in this research. Red dots are providers and blue dots represent patients. The size of the nodes is proportional to their total cost. The thickness of edges is proportional to the total claim cost of the provider-patient pair.

providers and blue ones represent patients. The size of the nodes is proportional to their total cost. There are a few very noticeable high total cost providers (the large red dots) but they also have lots of patients. This suggest that to find outlier providers, a more detailed analysis of each provider's neighborhood in the graph should be performed that takes into account his number of patients in the data, as well as how his claims are distributed across these patients because high total cost does not necessarily mean that a provider is responsible for excessive spending.

To detect outlier providers, we consider features of the egonets of the providers. An egonet of a node is its one hop neighbourhood. Figure 4.2 gives an example of an egonet. In the scope of this thesis, features like total weight and entropy ratio are extracted from every provider's egonet.

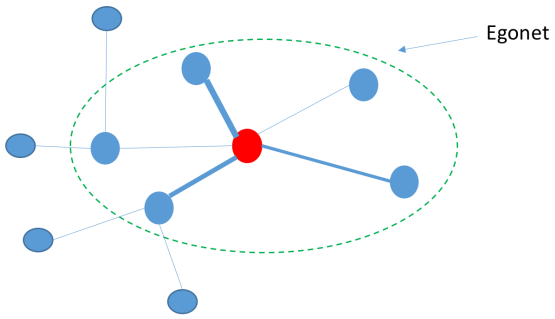

Figure 4.2: Illustration of the egonet of a provider

4.2 Edge Weight Power Law (EWPL)

The edge weight power law (EWPL) [3] states that the total edge weight W and the number of edges E of all the egonets in a graph G follow a power law, as illustrated by the following equation:

$$W \propto E^{\beta}, \beta > 1 \tag{4.1}$$

In the bipartite provider-patient graphs, the number of edges E in the egonet of a provider is the same as the number of patients of that provider (since each edge connects a provider with a patient). The weight of the egonet of a provider is either the total claim cost for that provider (in the "total" approach), or the total number of claims for that provider (in the "simple" approach). Hence in our settings, the power law could be interpreted as that the total claim cost (or the total number of claims) of a provider on one hand, and the number of patients of that provider on the other hand, should follow a power law.

In our data, we indeed observed power-law-like distributions for the provider egonets' weight (W) in terms of their size (E). Figure 4.3 plots this relationship between the size (horizontal axis) and the weight (vertical axis) of the egonet of providers. In this picture, each provider is represented as a point, and the weight for a provider's egonet is calculated as the total claim cost. Figure 4.4 shows an analogue plot where the weight for a provider's egonet is calculated as the total number of claims. Next, we can fit a line to this data in the

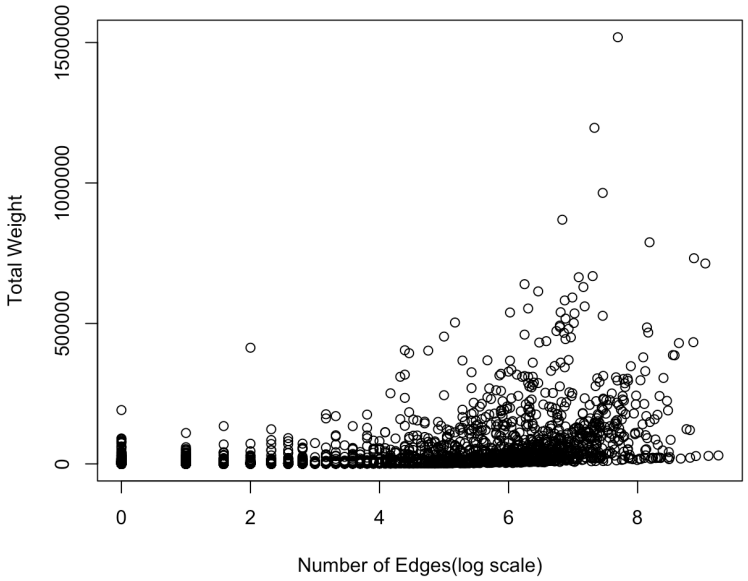

Figure 4.3: Edge weight power law illustrated for provider egonets in patient-provider graph where the edge weight is the total claim cost ("total" approach).

log-log space (in other words, learn the parameter β from Formula 4.1), and any providers with an abnormal deviation from the expected value are considered suspicious (see Section 4.5).

4.3 Egonet Entropy

While the powerlaw from Section 4.2 looks at the size and the weight of each provider's egonet as a whole, the entropy ratio (ER), whose use for detecting fraud in healthcare data was proposed in [17], measures how evenly the node (provider) associates with entities (patients) in its neighborhood, in terms of edge weights. The entropy ratio for node n is:

$$ER_n = \frac{1}{\log(|\mathcal{N}|)} \sum_{k \in \mathcal{N}} p_k \log \frac{1}{p_k} \tag{4.2}$$

where p_k is the percentage of node n's business with neighbor k out of its total business. In our context, n is a provider, k is a patient, and p_k is the percentage of the total claim cost (or, in the "simple" approach, the percentage of the total number of claims) of provider n

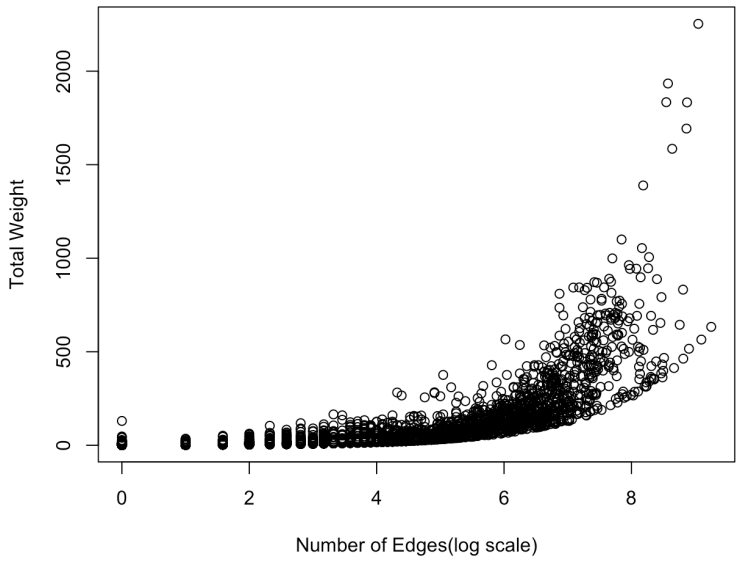

Figure 4.4: Edge weight power law illustrated for provider egonets in patient-provider graph where the edge weight is total number of claims ("simple" approach).

that is associated with patient k. The summation term is the empirical entropy, measuring the dispersion of n's business among its neighborhood $\mathcal{N}$. The entropy is further divided by $\log(|\mathcal{N}|)$ to normalize to the range $[0, 1]$. If n evenly distributes its business among $\mathcal{N}$, the entropy ratio is 1. If in contrast, n does most of its business with one neighbor, the dispersion is very skewed, resulting in an entropy ratio close to 0. Patients with dominant amount of costs or claims in a providers' egonet indicate a large variation in costs or number of claims, which suggest the practice of this provider are probably not consistent. Therefore, those providers with dominant patients and low entropy ratio are considered suspicious.

4.4 Workflow

Our workflow for the graph based approach for provider outlier detection consists of three steps:

1. **Graph Creation** Create patient-provider bipartite graph. We can choose to use either total cost ("total" approach) or total number of claims ("simple" approach) as weights.

2. **Feature Extraction** Extract egonet features like total weight, number of edges and entropy ratio for each provider egonet.

3. **Outlier Detection** For edge weight power laws, conduct linear regression using logarithmic binning in log-log space to fit the power laws and then calculate the outlier score, namely the deviation from the expected values [3, 20]. The formula for the outlier score is:

$$outlierScore(i) = \log(|y_i - (ax_i + b)| + 1) + LOF(i) \tag{4.3}$$

while,

$$y_i = log(W_i) \tag{4.4}$$

$$x_i = log(E_i) \tag{4.5}$$

a and b are coefficient and intercept of the fitted line in log-log space. W_i is the total weight of egonet i and E_i is the number of edges of egonet i .The first part evaluates the deviation from the expected power law value. The second part, LOF score [20], is implemented to measure the local deviation of a provider with respect to its neighbors. When we calculate LOF score, we determined the size of neighborhood to be 5.

For entropy ratio, the smaller the entropy ratio is, the more suspicious the provider is.

The reasons for introducing LOF score are explained as follows. The intuition is simple. In figure4.5, although point A and point B have the same deviation from their expected values, point B are supposed to be more suspicious because there are few points are around it. Thus, as a classical density based outlier detection metric, LOF scores are calculated to take this intuition into account.

Finally, results are sorted and filtered. For entropy results, trivial results (those egonets with only one edge, and hence zero entropy) are removed at this step. Thus, we have two ways to calculate edge weights ("total" and "simple") and two outlier metrics to assess each provider's egonet (based on either deviation from the expected power law relation between size and weight, or on entropy), which leads to four different methods in total.

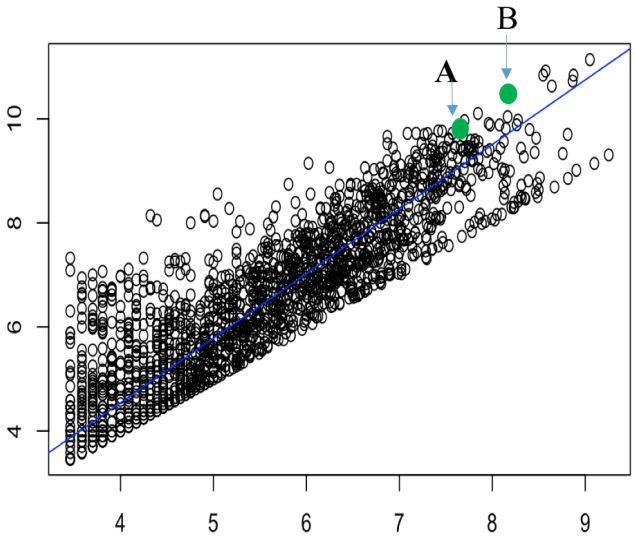

Figure 4.5: Edge weight power law illustrated for provider egonets in patient-provider graph where the edge weight is total number of claims ("simple" approach).

4.5 Experiments

We show the results for the four different graph based methods here. An evaluation of the results, and a comparison with the results from the reference group based methods of Chapter 5 are given in Section 5.6.4. In Figure 4.6 and Figure 4.7, the top 5 outlier providers are marked as green triangles. Red dots are the medians of each logarithmic bin. They are used to fit the blue line, which is the expected value for an egonet with a certain number of edges. In such a log-log space, x-axis means number of edges, basically number of patients of a provider; y-axis means the total weight of an egonet, namely total cost or total number of claims of this provider. We calculate log using base 2.

Each green triangle denotes a unique provider, which means there is no overlap between top5 suspected providers of these two methods. All 5 outlier providers in figure4.6 are providers who have a total claim cost that is higher than one would expect, based on their number of patients. On the other hand, 2 of the 5 providers marked as outliers in figure4.7 have a total number of claims that is lower than one would expect based on their number

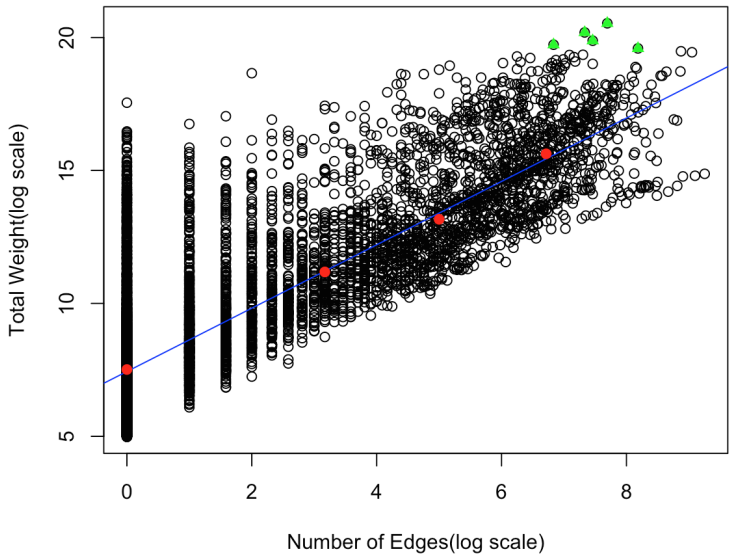

Figure 4.6: Edge weight power law in log-log scale for provider egonets in patient-provider graph where the edge weight is the total claim cost ("total" approach). The top 5 outliers are marked as green triangles.

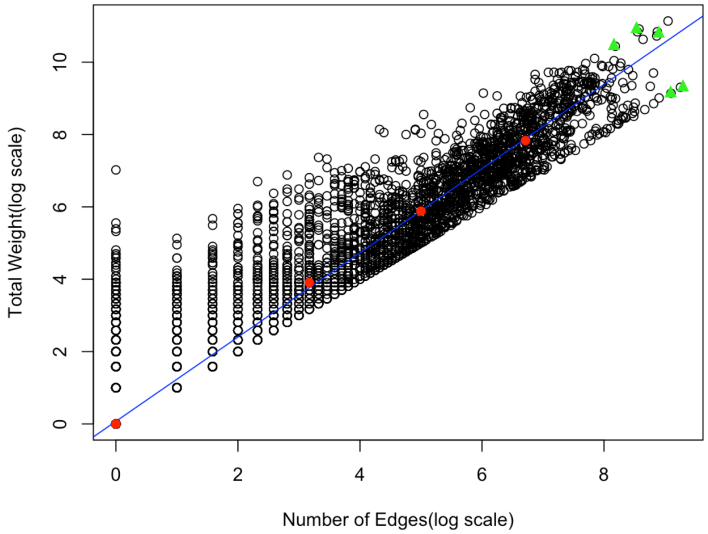

Figure 4.7: Edge weight power law in log-log scale for provider egonets in patient-provider graph where the edge weight is total number of claims ("simple" approach). The top 5 outliers are marked as green triangles.

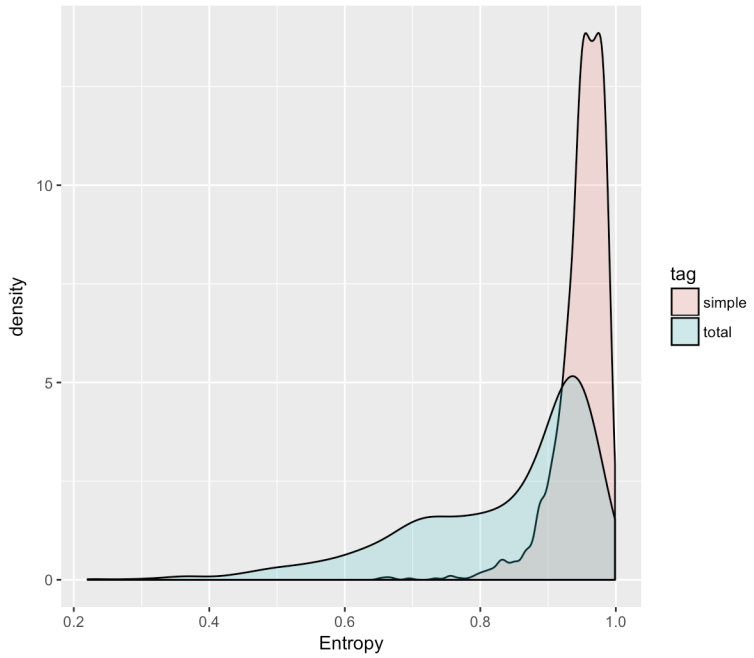

Figure 4.8: Entropy ratio distribution among providers in the data with at least 10 patients. Y-axis means the percentage of providers. As expected, most providers have high entropy.

of patients. One could easily revise 4.3 to exclude low cost and low visits outliers. However, studying these low cost examples may also provide actionable insights to lower the total healthcare cost. There is a clear lower bound in figure4.7 which fits $y = x$ in log-log space. This is quite reasonable since the every patient has at least one claim with the provider in the egonet. The parameters of the fitted blue line are listed in table4.1.

	Coefficient a	Intercept b	Formula $y = log(W, 2), x = log(E, 2)$
EWPL(total)	7.46227	1.19797	$y = 7.46227x + 1.19797$
EWPL(simple)	1.24206	-0.42810	$y = 1.24206x - 0.42810$

Table 4.1: Fitted parameters in log-log space for EWPL

Figure 4.8 shows the distribution of entropy ratio in the "total" and the "simple" approach. X-axis is entropy ratio and Y-axis means the percentage of providers. Before con-

structing the graphs, we first removed providers with less than 10 patients in the data. Since providers with a small number of patients are more likely to have an unbalanced distribution of edge weights in their egonets, which results in a small entropy ratio, they are trivial results for the entropy based outlier detection method. The extreme case would be a provider with only one patient in the data; such a provider will have zero entropy. In general, it can be expected that few providers have small entropy ratio, which is confirmed by Figure 4.8. The providers who do have low entropy, i.e. those in the long tail on the left in Figure 4.8, are potentially suspicious.

Chapter 5

REFERENCE GROUP BASED OUTLIER DETECTION

In this chapter, we propose cluster-based methods to uncover unwarranted variation in healthcare spending by automatically extracting reference groups of peer-providers from the data and then detecting high cost outliers within these groups. This chapter introduces three reference group based techniques: a specialty based, a provider centric and a patient centric method. In the last section of this chapter, we present and compare the results from both the cluster based methods (this chapter) and the graph based methods (Chapter 4) and discuss the implications.

5.1 Overview

For each provider in the data we can compute the total claims amount, i.e., the aggregate dollar amount for all healthcare claims associated with that provider from January 1, 2016 to June 30, 2016. One would expect that some providers will naturally have a higher total claim spend than others, because of specialty of practice, volume of practice, or a particular patient segment that may require more intensive (expensive) care. Our general aim is therefore to identify providers who have a total claim amount that is abnormally high *within* a reference group of peer providers. Such reference groups can be defined in various ways: they can be groups of providers of the same specialty, or groups of providers who treat patients with similar diagnoses. In Section 5.2, 5.3, and 5.4 the various methods used to define such reference groups are described, including utilizing clustering-based techniques to extract peer provider groups automatically from data. In Section 5.6 we analyze the influence of the method for reference groups definition on the outcomes of the method to identify cost variation.

Once the reference groups are established, the next step is to identify outliers in terms of claim costs within those groups. This is described in Section 5.5. In this thesis, the focus is primarily on high cost providers, i.e., those that represent an absolute high cost in addition to a relative high cost among their peers. High cost providers are specifically targeted since they are of particular interest to the ACO, in the sense that identifying these providers could enable further evaluation of spend which may have significant impact.

The overall workflow of the proposed approach is shown in Figure 5.1. First, healthcare claims data is processed and fed into the outlier detection models. Next, the model uses the data to divide the providers into (potentially overlapping) reference groups. Providers in the same group are similar according to predefined criteria (see Section 5.2, 5.3, and 5.4), hence they can be expected to have a similar total claim amount. Once the reference groups are defined and detected, using the technique described in Section 5.5, within each group a cost threshold is computed, outlier providers are identified, and their total claims related costs are estimated. Finally, results from the different reference groups are combined into an overall list of providers ranked by their total outlying claims related spend.

5.2 Specialty Based Method

This baseline method compares spending among providers with the same specialty. Patient demographics and diagnoses are not considered in this method. In each specialty provider peer-group, we single out those providers with a high median cost per patient using the metric described in Section 5.5. Specialties with less than 10 providers in our data are excluded, resulting in 54 different specialties in this component of the analysis.

5.3 Provider Centric Method

In the provider centric method we automatically extract reference groups from the data. Each reference group consists of providers who treat patients with similar diagnoses, characterized by CCS codes. To this end, we extract a *provider feature matrix* from the healthcare claims data following Algorithm 1. For each provider p, all the claims associated with that provider

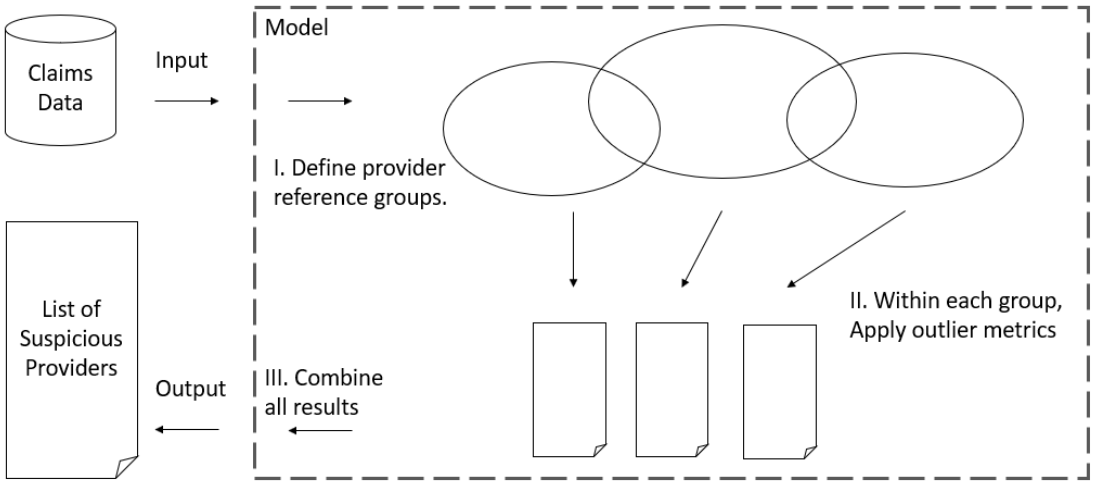

Figure 5.1: Overall workflow of the proposed approach for identifying high cost providers. Reference groups of peer providers are automatically learned from healthcare claims data – based on information about the patients they treat – and then further analyzed to detect outliers.

are collected for the study time period. Each claim contains one or more CCS diagnosis codes from a total of 260 unique CSS codes in the data. All CSS codes are concatenated into a list called *diagnosisDoc*. This list is likely to contain duplicates of CSS codes. Indeed, a provider usually has multiple patients with the same CSS codes. Additionally, the same patient may visit a provider multiple times for the same condition, leading to multiple healthcare claims with the same CSS code. Figure 5.2 illustrates the process of constructing a diagnosis document for a provider, based on the claims associated with that provider.

After creating a diagnosis document for each provider in this way, the corpus of these documents is converted into a term-frequency, inverse document-frequency (TF-IDF) matrix of the kind commonly used in the information retrieval [18] domain. In our case, each term corresponds to a CCS code, and each document corresponds to a provider, represented as a document and containing all the diagnosis codes from their claims. In particular, the TF-IDF matrix M is a matrix in which every row corresponds to a provider and every column corresponds to a diagnosis code (CCS code). The entry $M_{p,c}$ for provider p and diagnosis

Provider NPI	Claim ID	Primary CCS	Second CCS	Other CCS
1000	1	1	205	-
1000	2	176	-	-
1000	3	205	-	-
1000	4	205	135	1

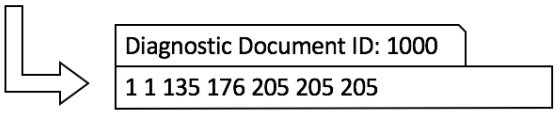

Figure 5.2: Illustration of diagnosis document creation for a provider

code c is a number between 0 and 1, representing the relative importance of CCS code c with respect to provider p. It is computed as:

$$M_{p,c} = f_{p,c} \cdot \log\left(\frac{N}{n_c}\right) \tag{5.1}$$

where $f_{p,c}$ is the number of times CCS code c appears in the claims of provider p, n_c is the number of providers that have diagnosis code c in at least one of their claims, and N is the total number of providers. Since there are 260 unique CCS code in the data, feature matrix M has 260 columns. The second factor in Equation (5.1) serves to reduce the role of CSS codes that commonly occur among many or all providers: the higher the value n_c, i.e. the more providers have patients with diagnosis code c, the less informative this code is for distinguishing among reference groups of providers.

Each row of the matrix M corresponds to a feature vector for a provider. The k-means clustering algorithm is applied with cosine distance [18] to group these feature vectors (i.e., providers) into k different clusters. Every provider appears in exactly one of the provider clusters. Note that clustering in the provider-centric method is done purely based on diagnosis codes and that no cost information is used. Section 5.5 describes how the cost information is subsequently used to detect outliers within each cluster or reference group.

Algorithm 1 Build provider feature matrix

1: **function** BUILD PROVIDER TF-IDF MATRIX
2: **Initialize** *Corpus*
3: **for** p `in` *providerlist* **do**
4: $C \leftarrow$ all claims with providerID $= p$
5: $diagnosisDoc \leftarrow$ list of all CCS codes from C
6: add $\{p : diagnosisDoc\}$ to $Corpus$
7: Build tf-idf matrix M from $Corpus$
8: **return** M

5.4 Patient Centric Method

In this method, patients are clustered based on their diagnostic history and demographic features such as age and gender, also utilizing the k-means clustering technique. First we construct a CCS feature matrix M as we do in the provider centric method. Then, we use an algorithm similar to Algorithm 1, with lines 3 and 4 replaced by

for p **in** *patientlist* **do**
$C \leftarrow$ all claims with personID $= p$

Each row of the resulting matrix M corresponds to a feature vector for a patient. We use M_p to denote the row corresponding to patient p. We define the distance between patients p_1 and p_2 as:

$$\begin{aligned} Dist(p_1, p_2) &= cosineDist(M_{p_1}, M_{p_2}) \\ &\quad +\alpha \cdot (1 - \delta(gender(p_1), gender(p_2))) \\ &\quad +\beta \cdot |age(p_1) - age(p_2)| \end{aligned}$$

in which α and β are weights to be tuned, and $\delta(x, y)$ is 1 if $x = y$, and 0 otherwise.

We cluster the patients with k-means clustering based on the distance function defined

above. Next, for each patient cluster, we derive an induced provider cluster containing all providers who cared for at least one of the patients in the cluster. In this way, we obtain k provider clusters. Note that a provider can appear in multiple clusters. Finally, within each of the k clusters, we group the providers by specialty, thereby subdividing the clusters into provider reference groups.

5.5 Outlier Definitions

The *median total cost per patient* $MedCost(p, R)$ of a provider p with respect to a reference group R is used as a metric to identify outliers. In the specialty based and provider centric methods, each provider p belongs to exactly one reference group, and the median total cost per patient for p is computed as the median of the total claim cost of all p's patients in the entire dataset, for services provided by p. In the patient centric method, a provider p can belong to multiple reference groups R, each of which are induced by a different patient cluster P. In this case, $MedCost(p, R)$ is calculated as the median total cost per patient for p restricted to patients from P.

Across all three methods, we only compute $MedCost(p, R)$ if sufficient data is available, namely if p has at least 10 patients with respect to the reference group R. This is especially relevant for the patient centric method where the number of patients of p can differ across reference groups and be substantially lower than the total number of patients of p in the whole dataset.

Given a reference group of providers R and $C = [MedCost(p) \,|\, p \in R]$, we define the threshold to identify outliers in R using Equation (5.2)

$$thresh(R) = Q_3(C) + 2 \cdot (Q_3(C) - Q_1(C)) \tag{5.2}$$

where $Q_1(C)$ is the 25^{th} percentile, and $Q_3(C)$ is the 75^{th} percentile. Any provider with a median cost per patient greater than the threshold is marked as an outlier. Figure 5.3 gives a clear illustration on how we define outliers.

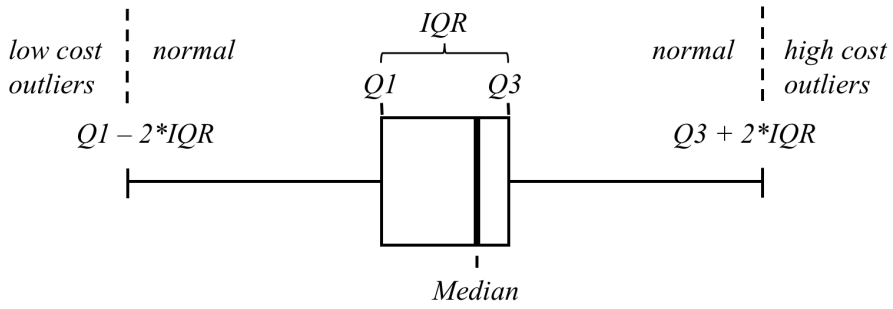

Figure 5.3: Outlier Metric Illustration

The excess spend amount of outlier provider p in group R is estimated as the amount which exceeds the threshold:

$$exc(p, R) = \frac{N_{p,R}}{2} \cdot (MedCost(p, R) - thresh(R)) \tag{5.3}$$

in which $N_{p,R}$ is the number of patients of provider p with respect to reference group R (half of which have a total claim cost for p that is at least as high as $MedCost(p, R)$). Next, $exc(p, R)$ is summed over all reference groups R to which p belongs to obtain the overall excess spend amount for outlier p. Finally, outlier providers are sorted according to their excess spend amount in descending order.

5.6 Experiments and Results

5.6.1 Comparison of the Three Reference Group based Methods

In this section we compare the results of the proposed methods for outlier provider detection when applied to the data described in Chapter 3. In the specialty based method, the providers are grouped in 54 distinct reference groups based on specialty. Unless where explicitly stated otherwise, all presented results for the provider centric method and the patient centric method are based on the creation of 10 provider clusters and 15 patient clusters respectively.

Regarding the *specialty based method*, the top side of Table 3.2 shows the top 5 specialties, out of the 54 used in the results in this section. Regarding the *provider centric method*, Table

Cluster	Top Frequent Diagnosis Codes	Explanation
0	I10	Hyper tension
	E119	Type 2 diabetes mellitus without complications
	Z0000	Encounter for general adult medical examination without abnormal findings
1	I4891	Unspecified atrial fibrillation
	I2510	Atherosclerotic heart disease of native coronary artery without angina pectoris
	I480	Paroxysmal atrial fibrillation
2	G4733	Obstructive sleep apnea
	J449	Chronic obstructive pulmonary disease
	R05	Cough
3	Z5181	Encounter for therapeutic drug level monitoring
	N186	End stage renal disease
	Z5111	Encounter for antineoplastic chemotherapy
4	C61	Malignant neoplasm of prostate
	N401	Benign prostatic hyperplasia with lower urinary tract symptoms
	N390	Urinary tract infection
5	Z1231	Encounter for screening mammogram for malignant neoplasm of breast
	K7460	Unspecified cirrhosis of liver
	Z1211	Encounter for screening for malignant neoplasm of colon
6	M545	Low back pain
	M4806	Spinal stenosis,lumbar region
	M5416	Radiculopathy, lumbar region
7	H2511	Age-related nuclear cataract, right eye
	Z01818	Encounter for other preprocedural examination
	H2512	Age-related nuclear cataract, left eye
8	L570	Actinic keratosis
	L821	Other seborrheic keratosis
	C44319	Basal cell carcinoma of skin of other parts of face
9	M069	Rheumatoid arthritis
	M1811	Unilateral primary osteoarthritis of first carpometacarpal joint, right hand
	Z471	Aftercare following joint replacement surgery

Table 5.1: Frequent diagnosis codes in provider clusters identified in the provider centric method

Cluster	Number of providers	Number of outlier providers	Total amount (million)	Excess amount (million)
0	1,500	189	$34.19	$2.68
1	693	91	$12.71	$1.57
2	713	62	$3.78	$0.06
3	1,284	161	$32.92	$1.73
4	415	50	$5.66	$0.16
5	934	124	$6.72	$0.64
6	705	42	$7.21	$0.23
7	535	72	$3.32	$0.09
8	393	41	$1.77	$0.03
9	974	82	$12.61	$0.24

Table 5.2: Summary of provider clusters identified in the provider centric method

5.1 shows the top 3 most frequently occurring ICD-10 codes in each provider cluster. Despite the fact that the clusters were automatically created from data, many of these clusters have a clearly identifiable theme, such as vascular diseases (cluster 1), pulmonary diseases (cluster 2), diseases related to the urinary system (cluster 4), spine diseases (cluster 6), ophthalmology (cluster 7), dermatology (cluster 8), and arthropathy and rheumatology (cluster 9). Table 5.2 provides summary statistics for each of the provider clusters. Finally, Table 5.3 provides a summary about the patient clusters detected by the *patient centric method*. With all providers with less than 10 patients excluded in each cluster, the total claim amount of the 15 patients clusters adds up to 110.63 million dollars.

Table 5.4 contains an overview of the number of outlier providers detected by each method, as well as the total estimated excess spend, and the total number of claims involved. Each of the three methods produces a list of outlier providers ranked in descending order in terms of estimated excess spend. Table 5.5 compares the overlap between the top

Cluster	Number of patients	Number of providers	Total amount (million)	Number of unique specialties
0	951	498	$0.27	26
1	1,245	963	$3.12	31
2	4,876	1,924	$11.89	47
3	3,278	1,085	$18.85	36
4	4,524	1,845	$39.15	43
5	1,193	576	$0.83	26
6	3,870	1,774	$19.06	40
7	2,205	1,033	$5.93	32
8	781	566	$0.98	22
9	966	740	$0.56	28
10	1187	869	$3.94	23
11	595	391	$0.64	22
12	1295	1066	$4.35	32
13	832	604	$0.75	36
14	698	521	$0.33	25

Table 5.3: Summary of patient clusters identified in the patient centric method

Method	Outlier providers	Excess amount (million)	Flagged claims
Specialty based	740	$5.9	14,010
Provider centric	914	$7.4	20,710
Patient centric	1,321	$9.0	20,599

Table 5.4: Number of detected outliers, total estimated excess amount, and total number of claims involved

20 outlier providers identified by each of the methods in terms of Jaccard similarity. Table 5.6 contains a similar comparison for the top 500. The results of the provider centric method are somewhat similar to the specialty based method, both on the top 20 and the top 500, while the results of the patient centric method are substantially different on the top 500, as indicated by the low Jaccard index values, implying smaller overlap. This phenomenon suggests that there is a group of dominant providers with unusual high spending that are detected by all methods, while at the same time the individual methods are distinct enough and focus on different aspects as they produce different results overall.

	specialty based	provider centric
provider centric	0.67	-
patient centric	0.60	0.54

Table 5.5: Jaccard index computed on top 20 outlier providers

	specialty based	provider centric
provider centric	0.6502	-
patient centric	0.1738	0.2433

Table 5.6: Jaccard index computed on top 500 outlier providers

The differences among the three methods in the specialty distribution of top 500 providers was also explored. There are 29, 40 and 20 unique specialties that appear in the top 500 outlier providers for the specialty based method, the provider centric method, and the patient centric method respectively. Table 5.7 shows the distributions of specialties in the top 500 outlier providers from each method. The provider centric method detected about 30 more general surgery providers than the other two methods did. The patient centric method nearly doubled the number of detected internal medicine physicians and family practice physicians

Specialty Based Method		Provider Centric Method		Patient Centric Method	
Specialty	Providers	Specialty	Providers	Specialty	Providers
Internal Medicine	169	Internal Medicine	150	**Internal Medicine**	**214**
Family Practice	85	Family Practice	62	**Family Practice**	**122**
Orthopedic Surgery	35	**General Surgery**	**59**	Orthopedic Surgery	68
Ophthalmology	23	Orthopedic Surgery	34	**General Surgery**	**28**
General Surgery	**19**	Ophthalmology	22	Neuropsychiatry	16
Nurse Practitioner	16	Otolaryngology	13	Nurse Practitioner	15
Neuropsychiatry	16	Radiation Oncology	13	Obstetrics & Gynecology	7
Physician Assistant	15	Neuropsychiatry	12	Emergency Medicine	6
Otolaryngology	13	Hematology-Oncology	11	Physician Assistant	3
Dermatology	11	Emergency Medicine	10	Neurosurgery	3
Others	98	Others	114	Others	18

Table 5.7: Specialty distributions of top 500 outliers

compared to the results of the other two methods. It is reasonable to conclude that the three methods produce different results, therefore, they have the potential to discover different types of outliers.

Note that some providers are marked as outliers by all three methods (#1, #2, #3, and #7) while others are only picked up by one of the three methods (highlighted in bold in Table 5.9). This illustrates that each of the methods has its use.

5.6.2 Key Findings and Analysis

An ideal way to validate the results would be to manually investigate each individual provider identified with each of the three methods. Given the large number of providers which are present in the data, it would require massive amounts of human resources to do such a comprehensive analysis. To mitigate this problem, we focused on the top 20 providers in each list. Table 5.8 summarizes the statistics for the top 20 providers identified by each

Method	Excess amount (million)	Flagged claims
Specialty based	\$2.8	3,325
Provider centric	\$3.1	3,404
Patient centric	\$3.4	3,067

Table 5.8: Results for the top 20 outlier providers

method. As indicated in Table 5.5, there is an overlap of providers in the three lists; the total number of unique providers in all three top 20 lists combined is 29.

The domain experts that we consulted, including physicians with clinical experience across multiple specialties, manually checked these provider lists to determine if there is reasonable evidence to believe their billing may warrant further evaluation. Table 5.9 displays the manually marked labels for the top 20 providers in each method. "Y" denotes a confirmed outlier, "N" denotes a false positive discovery and "?" suggests that further investigations are required.

In practice, there are various reasons to confirm a high cost anomaly. For example, some of the cost variation that was found relates to outpatient providers billing at inpatient facilities. In some claims, the therapeutic procedure codes (HCPCS codes) did not match with the patients' diagnosis codes. The mismatch between procedure codes and diagnosis codes could mean potential coding mistakes, billing errors, or potentially waste or abuse. Among the claims of confirmed abnormal high cost providers, two major patterns were discovered as follows:

- **Pattern 1. Excessive billing of physical therapeutic procedures.** Physical therapeutic codes like *97110 Therapeutic exercises* are billed many times in just one claim. We note that these codings comply with current government regulations and HCPCS modifiers requirements. Although these codes are not expensive in unit price separately, still they add up to a large amount of excess spending. In an extreme case, *97110 Therapeutic*

Specialty Based		Provider Centric		Patient Centric	
Provider	Label	Provider	Label	Provider	Label
#1	*Y*	#1	*Y*	**#25**	*Y*
#2	*Y*	#2	*Y*	#1	*Y*
#3	*Y*	#3	*Y*	#2	*Y*
#4	*N*	#4	*N*	#3	*Y*
#5	*N*	#21	*N*	#20	?
#6	*N*	#5	*N*	#6	*N*
#7	*Y*	#6	*N*	#7	*Y*
#8	*N*	#7	*Y*	#5	*N*
#9	?	#8	*N*	#12	?
#10	?	#9	?	#9	?
#11	?	**#22**	*Y*	#8	*N*
#12	?	#10	?	#10	?
#13	?	#12	?	**#26**	*Y*
#14	?	#23	?	#27	?
#15	?	#11	?	#11	?
#16	?	#14	?	#14	?
#17	*Y*	#13	?	#28	?
#18	?	#20	?	#18	?
#19	?	#15	?	#13	?
#20	?	#24	?	#29	?

Table 5.9: Results for the top 20 outlier providers. "*Y*" means confirmed outliers. "*N*" means false alarms and "?" means that further investigations are required.

exercises was billed more than 60 times in one claim. While some of the claim items were rejected by the insurance company, that total claim costs around $4,000 in total which is unusually high.

- **Pattern 2. High cost variations in hospice services.** Hospice codes such as *Q5001*, *Q5002*, *Q5003* and *Q5004* are widely found in claims of many high cost providers. About 35% of abnormal high cost claims contain hospice codes. Meanwhile, hospice items display huge cost variations. They are not always expensive items. No specific type of disease, nor geographical locations was found correlated to these codes. Further investigations are needed to identify the exact problem.

5.6.3 Sensitivity Analysis and Parameters Tuning

To measure the effect of the choice of the number of clusters k on the results, we varied the number of clusters k in the provider centric method from 1 to 20. The results are given in Figure 5.4 and 5.5. Figure 5.4 shows that as the number of clusters increases, the number of detected suspicious high cost providers remains roughly the same (represented by the dotted line). When one counts the number of abnormal providers against increasing the number of clusters, one finds that the accumulative number of providers tends to converge to a limit. Figure 5.5 shows the relation between the detected total excess dollar amount and the number of provider clusters. The figure demonstrates that the detected excess amount decreases slowly as the number of clusters increases. To conclude, the number of clusters does not greatly affect the final result of detected outlier providers in the provider centric method. Therefore, we chose the number of provider clusters to be 10 in the experiments.

For the patient centric method we varied the number of patient clusters from 5 to 30 in increments of 5. When utilizing smaller numbers of clusters, there was no sufficient variability in the specialties of the providers from a domain perspective. When utilizing larger numbers of clusters, not enough interesting outliers were discovered from the resulting clusters. 15 patient clusters was a sufficient number in terms of balancing novelty and coverage of provider

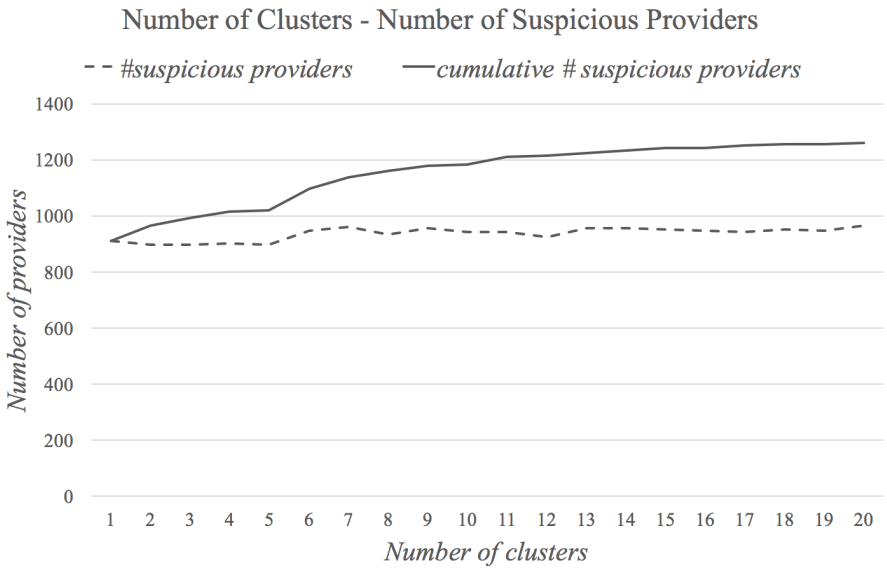

Figure 5.4: Number of detected outlier providers in terms of the number of clusters used in the provider centric method

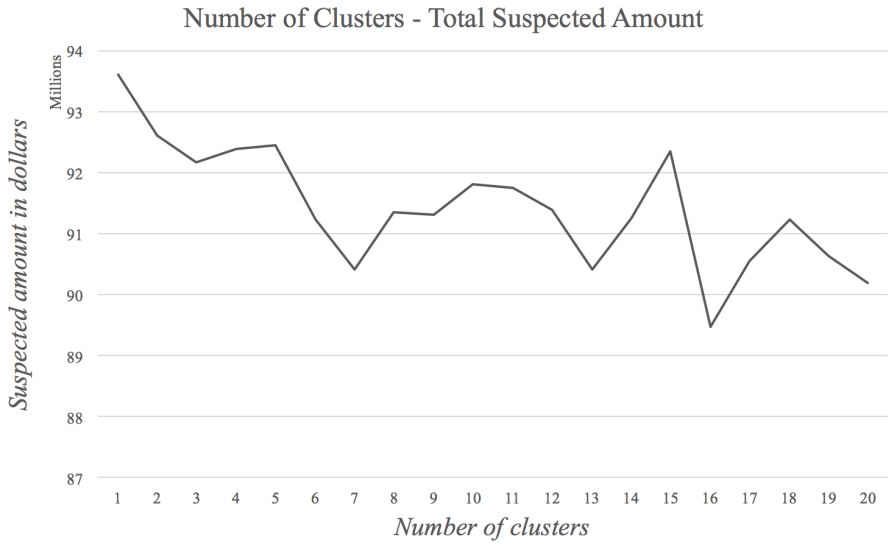

Figure 5.5: Total detected excess amount in terms of the number of clusters used in the provider centric method

	specialty based	provider centric	patient centric	EWPL (total)	EWPL (Simple)	ER (total)
provider centric	0.61	1	-	-	-	-
patient centric	0.36	0.37	1	-	-	-
EWPL (total)	0.08	0.11	0.17	1	-	-
EWPL (simple)	0.005	0.01	0.02	0.12	1	
ER(total)	0.00	0.00	0.01	0.01	0.00	1
ER(simple)	0.01	0.01	0.01	0.01	0.00	0.08

Table 5.10: Jaccard index computed on top 100 suspected providers in every method. Recall that EWPL stands for "Edge Weight Power Law" and ER for "Entropy Ratio"

specialties. It should be noted that the threshold for the optimal number of clusters may be different for different ACOs with different underlying populations.

5.6.4 Comparison with Graph Approach based Results

In this section, we compare the results of the three reference group based methods proposed in this chapter with the four graph based methods from Chapter 4. In Table 5.10 we present the Jaccard index computed for the top 100 suspected providers for each pair of methods. The first rows and columns correspond to the reference group based methods, while the remaining rows and columns correspond to the graph based methods. A higher Jaccard index value indicates a higher overlap between the results produced by the methods.

The main observations from Table 5.10 are that: (1) out of all methods, the specialty based method and the provider centric method are the most similar to each other in terms of results, which confirms what we already observed in Table 5.5 and 5.6; and (2) all other pairs of methods have a fairly low to very low Jaccard index value, which means that they yield very different final outputs. This means that the various methods have the potential to identify different kinds of outlier providers.

An important question is of course how accurate each of the methods is. In Table 5.11, we display the results of the top 10 providers in each model output, as validated by a domain expert. The results for the specialty based metric, the provider centric method, and the patient centric method correspond to what we already reported in Table 5.9. As can be seen in Table 5.11, the Edge Weight Power Law (EWPL) method based on the bipartite patient-provider graph approach that uses the total claim costs as edge weights, confirms some of the outlier providers from the reference group based methods. The other graph based methods do not successfully identify any additional outlier providers; instead their top 10 is entirely composed of false positives.

Our general conclusion based on the results in Table 5.11 is that, in deployment in a system for detecting fraud, waste, and abuse in healthcare insurance claims, one can apply all the reference group based methods as well as the EWPL (total) method, and give a high investigation priority for providers that are simultaneously marked as outliers by all of these four methods.

	specialty based	provider centric	patient centric	EWPL (total)	EWPL (simple)	Entropy Ratio (total)	Entropy Ratio (simple)
1	**Y**	**Y**	**Y**	N	N	N	N
2	**Y**	**Y**	**Y**	N	N	N	N
3	**Y**	**Y**	**Y**	N	N	N	N
4	N	N	**Y**	**Y**	N	N	N
5	N	N	N	N	N	N	N
6	N	N	N	**Y**	N	N	N
7	**Y**	N	**Y**	N	N	N	N
8	N	**Y**	N	**Y**	N	N	N
9	N	N	N	N	N	N	N
10	N	N	N	N	N	N	N

Table 5.11: Top 10 results of different methods. "Y" means worth further investigation and "N" means false positive findings.

Chapter 6

CONCLUSIONS

Excessive spending and waste is a perennial problem in healthcare in the United States. In this thesis we addressed the problem of detecting potentially wasteful and fraudulent providers through automatically mining a large dataset of healthcare insurance claims. We applied the methods proposed in this thesis to data from a large Accountable Care Organization (ACO) in the United States. It pertains to care provided to more than 28,000 patients by more than 8,000 providers in 2016 (see Chapter 3).

In Chapter 4, we implemented four graph based methods for provider outlier detection. In each case, we first transformed the data into a bipartite patient-provider graph, in which the nodes are patients and providers. An edge is added between a patient and a provider if the patient visited this provider at least once. We explored two meaningful ways to calculate the edge weight: based on the total claim cost ("total" approach), and based on a simple count of the total number of claims ("simple" approach). To detect outlier providers in this graph structure, we measured properties of the egonets of the providers. In normal circumstances, the relationship between the weight of an egonet and its size is expected to follow a power law (edge weight power law, or "EWPL" approach). Similarly, in normal circumstances, the entropy in the egonet of a provider is expected to be high (entropty ratio, or "ER" approach). Providers whose egonet deviates from these standard behaviors can be marked as outliers.

In Chapter 5, we proposed three new methods to uncover unwarranted variation in healthcare spending by automatically extracting reference groups of peer-providers from the data and then detecting high cost outliers within these groups. Such reference groups can be defined in various ways: they can be groups of providers of the same specialty (*specialty*

based method), or groups of providers who treat patients with similar diagnoses (*provider centric method* and *patient centric method*). For the later two methods we extracted peer provider groups automatically from data by representing providers and patients as vectors in a vector-space model where each dimension corresponds to a diagnosis code (CCS) code. Next we clustered these vectors with k-means clustering with cosine distance. A first interesting observation was that, despite the fact that the clusters were automatically created from data, many of these clusters had a clearly identifiable theme, such as vascular diseases, pulmonary diseases, diseases related to the urinary system, etc. Once the reference groups are established, the next step is to identify outliers in terms of claim costs within those groups. We do this by computing the median cost per patient for each provider in the reference group, and singling out those providers with an excessively high median cost per patient. Next, for each of the three reference group methods, we rank the outlier providers according to their excess spend amount in descending order.

Through validation on the data from Chapter 3 with the help of domain experts, we found that all three reference group based methods from Chapter 5 as well as the EWPL-total methods from Chapter 4 produced actionable and meaningful results. Each of these methods included between 3-5 providers in their top 10 that were confirmed by domain experts to be indeed excessively high cost providers. Among the claims of confirmed abnormal high cost providers, two major patterns were discovered: (1) excessive billing of physical therapeutic procedures, and (2) high cost variations in hospice services.

Further investigations into the claims of providers with suspicious spending are required to confirm and compare the results. Further comprehensive validation requires that detailed investigation be carried out by the ACO to determine potential system waste or abuse. Future work also includes introducing varieties of outlier detection metrics to establish thresholds, and exploring the feasibility of an embedding graph based approach into the reference group based framework.

BIBLIOGRAPHY

[1] Leman Akoglu and Christos Faloutsos. What is strange in large networks? Graph-based irregularity and fraud detection. In *Data Mining and Knowledge Discovery*, ICDM'12. IEEE, 2012.

[2] Leman Akoglu, Tong Hanghang, and Koutra Danai. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29:626–688, 2015.

[3] Leman Akoglu, Mary McGlohon, and Christos Faloutsos. Oddball: Spotting anomalies in weighted graphs. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 410–421. Springer, 2010.

[4] Donald M Berwick and Andrew D Hackbarth. Eliminating waste in US health care. *Jama*, 307(14):1513–1516, 2012.

[5] Alex Beutle, Leman Akoglu, and Christos Faloutsos. Graph-based user behavior modeling: From prediction to fraud detection. In *the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'15. ACM, 2015.

[6] Varun Chandola, Banerjee Arindam, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41.3(15):275–287, 2009.

[7] Phua Clifton, Lee Vincent, Smith Kate, and Gayler Ross. A comprehensive survey of data mining-based fraud detection research, 2010.

[8] Hanbo Dai, ZHU Feida, Peng LIM Ee, and Hwa PANG Hwee. Detecting anomalies in bipartite graphs with mutual dependency principles. In *2012 IEEE 12th International Conference on Data Mining*, ICDM'12. IEEE, 2012.

[9] Michael Davis, Weiru Liu, and Paul Miller. Detecting anomalies in graphs with numeric labels. In *the 20th ACM international conference on Information and knowledge management*. ACM, 2011.

[10] Jianying Hu, Fei Wang, Jimeng Sun, Robert Sorrentino, and Shahram Ebadollahi. A healthcare utilization analysis framework for hot spotting and contextual anomaly detection. In *AMIA Annual Symposium Proceedings*, volume 2012, page 360. American Medical Informatics Association, 2012.

[11] Li Jing, Huang Kuei-Ying, Jin Jionghua, and Jianjun Shi. A survey on statistical methods for health care fraud detection. *Health Care Management Science*, 11.3:275–287, 2008.

[12] Hossein Joudaki, Arash Rashidian, Behrouz Minaei-Bidgoli, Mahmood Mahmoodi, Bijan Geraili, Mahdi Nasiri, and Mohammad Arab. Using data mining to detect health care fraud and abuse: a review of literature. *Global Journal of Health Science*, 7(1):194, 2015.

[13] Rob M Konijn, Wouter Duivesteijn, Wojtek Kowalczyk, and Arno Knobbe. Discovering local subgroups, with an application to fraud detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2013.

[14] Rob M. Konijn, Wouter Duivesteijn, Marvin Meeng, and Arno Knobbe. Cost-based quality measures in subgroup discovery. *Journal of Intelligent Information Systems*, 45(3):337–355, 2015.

[15] Jing Li, Huang Kuei-Ying, Jin Jionghua, and Shi Jianjun. A survey on statistical methods for health care fraud detection. *Health Care Management Science*, 11(3):275–287, 2008.

[16] Chinho Lin, Lin Chun-Mei, Li Sheng-Tun, and Kuo Shu-Ching. Intelligent physician segmentation and management based on kdd approach. *Expert Systems with Applications*, 34(3):1963–1973, 2008.

[17] Juan Liu, Eric Bier, Aaron Wilson, John Alexis Guerra-Gomez, Tomonori Honda, Kumar Sricharan, Leilani Gilpin, and Daniel Davies. Graph analysis for detecting fraud, waste, and abuse in healthcare data. *AI Magazine*, 37(2):33–46, 2016.

[18] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[19] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute Report, 2011.

[20] Breunig Markus, Hans-Peter Kriegel, Raymond T. Ng, and Jorg Sander. Lof: identifying density-based local outliers. In *the 2000 ACM SIGMOD international conference on Management of data*, SIGMOD '00, pages 93–104. ACM, 2000.

[21] P B Nesbitt. Saving health care with a managed medical network. http://leanmedicalcare.blogspot.com/2012/, 2016. [Online].

[22] Caleb C. Noble and Diane J. Cook. Graph-based anomaly detection. In *the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD'03. ACM, 2003.

[23] Macedo Paulo, Araia Sewit, and Zafari Babak. Medicare fraud analytics using cluster analysis: How PROC FASTCLUS can refine the identification of peer comparison groups. SAS Global Forum, 2016. [Online].

[24] Patient Protection and Affordable Care Act. Patient protection and affordable care act. *Public Law*, 111(48):759–762, 2010.

[25] Price Waterhouse Coopers (PWC). The price of excess: identifying waste in healthcare spending, 2012.

[26] EC Schneider, DO Sarnak, D Squires, A Shah, and MM Doty. Mirror, mirror 2017: international comparison reflects flaws and opportunities for better U.S. health care. *Commonwealth Fund Reports*, 2017.

[27] Jimeng Sun, Christos Faloutsos, Huiming Qu, and Deepayan Chakrabarti. Neighborhood formation and anomaly detection in bipartite graphs. In *Fifth IEEE International Conference on Data Mining*, ICDM'05. IEEE, 2005.

[28] Alexander Titus, Rebecca Faill, and Amar Das. Automatic identification of co-occuring patient events. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 579–586. ACM, 2016.

[29] Eberle William and Lawrence Holder. Anomaly detection in data represented as graphs. *Intelligent Data Analysis*, 11(6):663–689, 2007.

[30] Shi Yong, Yingjie Tian, Gang Kou, Yi Peng, and Jianping Li. Health insurance fraud detection. In *Optimization Based Data Mining: Theory and Applications*, pages 233–235. Springer, 2011.