Automatic Detection of Excess Healthcare Spending and Cost Variation in ACOs

Eric Liu * Muhammad A. Ahmad † Carly Eckert † Anderson Nascimento *

Martine De Cock *‡ Karthik Padthe † Ankur Teredesai † Greg McKelvey †

Abstract

There are more than nine hundred Accountable Care Organizations (ACOs)[1] in the United States, 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. Identifying fraud, waste, and abuse resulting in superfluous expenditures associated with care delivery is central to the success of ACOs 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 paper 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. We demonstrate the utility of our proposed framework on datasets from a large ACO in the United States to successfully identify unwarranted variation in therapeutic procedures even in low cost claims that had previously gone unnoticed.

1 Introduction

Healthcare expenditures in the United States exceed $3 trillion a year [5]. According to The Commonwealth Fund, the U.S. spends far more on healthcare than all other high-income countries [13]. Yet, among comparably wealthy nations, the U.S. ranks lowest in terms of quality of care, resulting in poorer health outcomes [13]. 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. 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. Cost variation analysis in an effort to identify excess healthcare spending is one way to do this.

As healthcare systems are increasingly focused on reducing costs, identifying excess spending is of particular interest to healthcare entities invested in value based care, such as Accountable Care Organizations (ACOs). ACOs are patient-centered care delivery organizations that were developed in alignment with the Patient Protection and Affordable Care Act [1] as a way to incentivize the quality of care that healthcare systems and provider organizations deliver to patients under their care. Cooperation and coordination of patient care are the foci of ACOs, which operate under the construct that improved coordination of care will reduce healthcare waste, lower cost, and improve care. Systems that are able to maintain quality while constraining costs are rewarded by healthcare payers including the Centers for Medicare and Medicaid Services (CMS) depending on pre-specified arrangements. ACOs can enter into a variety of revenue risk sharing agreements, but the overall concept is that ACOs are at risk of losing program revenue if their providers overspend when caring for their associated patients, termed attributed beneficiaries. Therefore, ACOs are interested in monitoring and evaluating outlier providers related to spending on attributed beneficiaries as a way to identify such excess spend. Excess spending could signal many things: appropriate care for medically complex patients, care fragmentation, or wasteful spending. However, accurately identifying these providers with high per-patient-spend can assist system administrators in their role.

In this paper, we propose clustering-based approaches to automate the detection of cost variations in medical claims and to identify excess spending among providers. Our aim is to identify providers whose median cost per patient is abnormally high compared to other providers. We will use the terms "providers" to refer to physicians, nurse practitioners, and physician assistants throughout this paper. When linking patients to their associated providers, particularly in an ACO setting, the attribution logic must be defined. In this paper, 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

*Institute of Technology, University of Washington, Tacoma. (jiachl5@uw.edu, andclay@uw.edu, mdecock@uw.edu)

†Kensci Inc., Seattle. (muhammad@kensci.com, carly@kensci.com, karthik@kensci.com, ankur@kensci.com, greg@kensci.com)

‡Dept. of Appl. Math, Comp. Science and Statistics, Ghent University

[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. 60719, (2017).

derived from the "rendering" provider field. In this paper, each claim has only one provider associated with it.

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 paper 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, 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.

Utilizing data from a large ACO in the United States, we examined medical claims that occurred from January 1, 2016 to June 30, 2016, reflecting the care of more than 28,000 patients. We clustered all providers based on the diagnosis codes in claims attributed to them and identified high cost providers responsible for excess spending. During our analysis, we discussed the results with our healthcare domain experts and identified two significant billing behavior patterns associated with high-cost providers. 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 paper are as follows:

- We propose two 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.

- 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.

2 Related Work

There is a rich literature on automated discovery of fraud and anomalies in data, including in healthcare settings. Unsupervised and semi-supervised techniques include graph based, instance based, and cluster-based approaches, which we briefly describe.

Graph Based Approaches When utilizing this approach, the data is converted into a graph, such as a bipartite patient-provider graph, and then examined 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, 10]. While we take different approaches in this paper, applying graph based techniques remains an interesting and good direction for future work.

Instance Based Approaches Konijn et al. proposed a subgroup discovery tool *Cortana*[2] for healthcare fraud detection [7, 8]. 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 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. Considering the scale and size of our data, this method may be very computationally costly.

Cluster Based Approaches The outlier detection techniques, which are most relevant to the work presented in this paper, are the cluster-based approaches. Hu et al. [6] 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. [9] proposed

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

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 1: Statistics on claims data from Jan 2016 through Jun 2016

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. [12] 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. [14] 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 this paper, 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 and comparison metrics are ill-defined, making it extremely difficult to compare the performance of unsupervised methods.

3 Data

We analyzed healthcare claims data from a large ACO in the 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[3] 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 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.

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 National 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[4] to map taxonomy codes, available in the NPI lookup, to specialty codes, which is a higher level of specialty categorization (i.e. 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 had only one specialty. There are 63 unique provider specialties in the data and 54 of these have at least 10 associated providers. In Table 2, the top specialties are listed in terms of largest number of providers and in terms of average cost per patient respectively.

In addition to the patient demographic and provider information, each claim has patient 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.[5] As opposed to the tens of thousands of ICD-10 codes, there are only 260 unique CCS codes in our data. In Section 4, we explain how we use these CCS codes to cluster providers and patients.

HCPCS (Healthcare Common Procedure Coding System) codes are used by health system administrators to encode the utilization of products, supplies, and services at-

[3] Claims and Claims Line Feed Files format, see p. 111 in `https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/MSSP-Reference-Table.PDF`

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

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

Rank	Top 5 most frequent specialty	Number of providers
1	Internal Medicine	1,207
2	Diagnostic Radiology	765
3	Family Practice	726
4	Physician Assistant	626
5	Emergency Medicine	563
Rank	**Top 5 most expensive specialty**	**Average cost per patient**
1	Cardiac Surgery	$ 11,824
2	Neurosurgery	$ 4,745
3	Hematology-Oncology	$ 4,399
4	General Surgery	$ 2,942
5	Orthopedic Surgery	$2,903

Table 2: Provider specialties top 5

tached to a claim. For each claim in the data, all HCPCS codes billed by the provider can be identified. In the proposed technique, HCPCS codes are used to verify the results.

4 Method

For each provider in the data the total claims amount was computed, i.e., the aggregate dollar amounts 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 or 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 4.1 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 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 4.2. In this paper, 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 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)

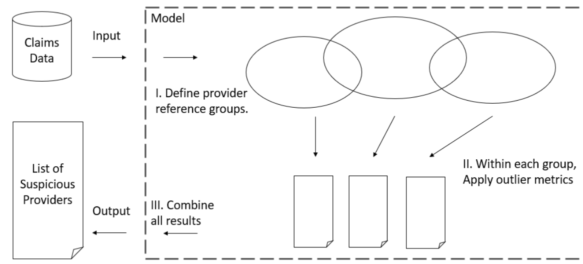

Figure 1: Overall workflow of the proposed approach for identifying high cost providers. Reference groups of peer providers are automatically learned from healthcare claims data – based on information about the patients they treat – and then further analyzed to detect outliers.

reference groups. Providers in the same group are similar according to predefined criteria (see Section 4.1), 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 4.2, 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.

4.1 Extraction of Reference Groups

Specialty Based 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 4.2. Specialties with less than 10 providers were excluded, resulting in 54 different specialties in this component of the analysis.

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 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 2 illustrates the process of constructing a diagnosis document for a provider, based on the claims associated with that provider.

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

Diagnostic Document ID: 1000

1 1 135 176 205 205 205

Figure 2: Illustration of diagnosis document creation for a 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 [11] 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 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:

$$M_{p,c} = f_{p,c} \cdot \log\left(\frac{N}{n_c}\right) \tag{4.1}$$

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 (4.1) 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 [11] 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 4.2 describes how the cost information is subsequently used to detect outliers within each cluster or reference group.

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 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

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:

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

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.

4.2 Outlier Detection 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

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

number of patients of p in the whole dataset.

Given a reference group of providers R and $C = [MedCost(p) \mid p \in R]$, we define the threshold to identify outliers in R using Equation (4.2)

$$thresh(R) = Q_3(C) + 2 \cdot (Q_3(C) - Q_1(C)) \tag{4.2}$$

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.

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

$$exc(p, R) = \frac{N_{p,R}}{2} \cdot (MedCost(p, R) - thresh(R)) \tag{4.3}$$

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 Results

5.1 Comparison of the Output of the Three Methods

In this section we compare the results of the proposed methods for outlier provider detection when applied to the data described in Section 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 2 shows the top 5 specialties, out of the 54 used in the results in this section. Regarding the *provider centric method*, Table 4 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 provides summary statistics for each of the provider clusters. Finally, Table 6 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 3 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 7 compares the overlap between the top 20 outlier providers identified by each of the methods in terms of Jaccard similarity. Table 8 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.

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 9 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 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.

5.2 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 10 summarizes the statistics for the top 20 providers identified by each method. As indicated in Table 7, 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.

Our domain experts, including physicians with clinical experience across multiple specialties, manually checked these provider lists to determine if there is reasonable evi-

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 4: 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: Summary of provider clusters identified in the provider centric method

dence to believe their billing may warrant further evaluation. Table 11 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

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	1,187	869	\$3.94	23
11	595	391	\$0.64	22
12	1,295	1,066	\$4.35	32
13	832	604	\$0.75	36
14	698	521	\$0.33	25

Table 6: Summary of patient clusters identified in the patient centric method

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

Table 7: Jaccard index computed on top 20 outlier providers

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

Table 8: Jaccard index computed on top 500 outlier providers

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 9: Specialty distributions of top 500 outliers

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 10: Results for the top 20 outlier providers

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	N
#6	N	#5	N	#6	N
#7	Y	#6	N	#7	Y
#8	N	#7	Y	#5	N
#9	N	#8	N	#12	N
#10	N	#9	N	#9	N
#11	N	#**22**	Y	#8	N
#12	N	#10	N	#10	N
#13	?	#12	N	**#26**	Y
#14	N	#23	?	#27	?
#15	?	#11	N	#11	N
#16	?	#14	N	#14	N
#17	Y	#13	?	#28	?
#18	?	#20	?	#18	?
#19	?	#15	?	#13	?
#20	?	#24	?	#29	?

Table 11: Results for the top 20 outlier providers. "Y" means confirmed outliers. "N" means false alarms and "?" means that further investigations are required.

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 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.

5.3 Sensitivity Analysis and Parameter 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 3 and 4. Figure 3 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 4 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

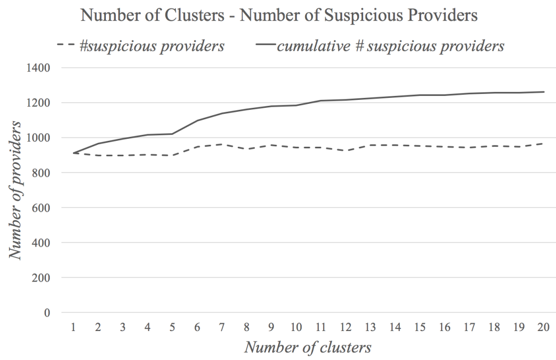

Figure 3: Number of detected outlier providers in terms of the number of clusters used in the provider centric method

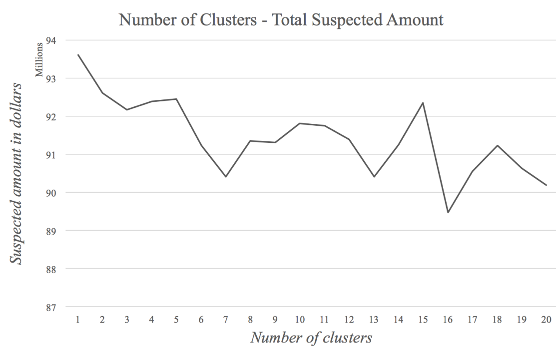

Figure 4: Total detected excess amount in terms of the number of clusters used in the provider centric method

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

6 Conclusion

In this paper we addressed the problem of detecting potentially wasteful and fraudulent providers from a large ACO in the United States. We proposed three methods and discussed their differences. Through detailed analysis, we demonstrated their potential to produce actionable and meaningful results. Further investigations into the claims of providers with suspicious spending are required to confirm and compare the results of three methods. Further comprehensive validation requires that detailed investigation be carried out by the ACO to determine potential system waste or abuse. We plan to comprehensively validate the results for a subset of providers in the future.

References

[1] *Patient protection and affordable care act*, Public law, 111 (2010), pp. 759–762.

[2] L. Akoglu, T. Hanghang, and K. Danai, *Graph based anomaly detection and description: a survey*, Data Mining and Knowledge Discovery, 29 (2015), pp. 626–688.

[3] L. Akoglu, M. McGlohon, and C. Faloutsos, *Oddball: Spotting anomalies in weighted graphs*, in PAKDD, 2010, pp. 410–421.

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

[5] Centers for Medicare Services, *National health expenditures 2016 highlights*, `https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/highlights.pdf`.

[6] J. Hu, F. Wang, J. Sun, R. Sorrentino, and S. Ebadollahi, *A healthcare utilization analysis framework for hot spotting and contextual anomaly detection*, in AMIA Annual Symposium Proceedings, 2012, p. 360.

[7] R. M. Konijn, W. Duivesteijn, W. Kowalczyk, and A. Knobbe, *Discovering local subgroups, with an application to fraud detection*, in PAKDD, 2013.

[8] R. M. Konijn, W. Duivesteijn, M. Meeng, and A. Knobbe, *Cost-based quality measures in subgroup discovery*, Journal of Intelligent Information Systems, 45 (2015), pp. 337–355.

[9] C. Lin, C. M. Lin, S. T. Li, and S. C. Kuo, *Intelligent physician segmentation and management based on kdd approach*, Expert Systems with Applications, 34 (2008), pp. 1963–1973.

[10] J. Liu, E. Bier, A. Wilson, J. A. Guerra-Gomez, T. Honda, K. Sricharan, L. Gilpin, and D. Davies, *Graph analysis for detecting fraud, waste, and abuse in healthcare data*, AI Magazine, 37 (2016), pp. 33–46.

[11] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[12] M. Paulo, A. Sewit, and Z. Babak, *Medicare fraud analytics using cluster analysis: How PROC FASTCLUS can refine the identification of peer comparison groups*. SAS Global Forum, 2016, `http://support.sas.com/resources/papers/proceedings16/10761-2016.pdf`. [Online].

[13] E. Schneider, D. Sarnak, D. Squires, A. Shah, and M. Doty, *Mirror, mirror 2017: international comparison reflects flaws and opportunities for better U.S. health care*, Commonwealth Fund Reports, (2017).

[14] A. Titus, R. Faill, and A. Das, *Automatic identification of co-occuring patient events*, in ACM-BCB, 2016, pp. 579–586.