Unsupervised Machine Learning for Explainable Medicare Fraud Detection*

Shubhranshu Shekhar  
Carnegie Mellon University, shubhras@andrew.cmu.edu

Jetson Leder-Luis  
Boston University and NBER, jetson@bu.edu

Leman Akoglu  
Carnegie Mellon University, lakoglu@andrew.cmu.edu

The US federal government spends more than a trillion dollars per year on health care, largely provided by private third parties and reimbursed by the government. A major concern in this system is overbilling, waste and fraud by providers, who face incentives to misreport on their claims in order to receive higher payments. In this paper, we develop novel machine learning tools to identify providers that overbill Medicare, the US federal health insurance program for elderly adults and the disabled. Using large-scale Medicare claims data, we identify patterns consistent with fraud or overbilling among inpatient hospitalizations. Our proposed approach for Medicare fraud detection is fully unsupervised, not relying on any labeled training data, and is explainable to end users, providing reasoning and interpretable insights into the potentially suspicious behavior of the flagged providers. Data from the Department of Justice on providers facing anti-fraud lawsuits and several case studies validate our approach and findings both quantitatively and qualitatively.

Key words: Medicare, fraud and abuse, machine learning, anomaly detection, explanation

1. Introduction

Fraud in health care is hard to detect. Insurers face information asymmetries, where physicians and patients both know more about the health care delivered than the insurer responsible for paying for that care. Providers face incentives to maximize their reimbursements from health insurance companies, and insurers must largely rely on documentation from providers themselves. This asymmetric information leads to circumstances where unscrupulous providers can choose to commit fraud.

*Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health under Award Number P30AG012810. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We also thank Lowell Taylor for the discussions on the earlier stages of the project as well as broadly sponsoring the DUA57882.
These issues are compounded in the federal health care programs, where the government is the insurer. The US federal government spends Trillions of dollars\(^1\) on health insurance, where fraud detection becomes challenging due to the sheer volume of claims being processed. Estimated US health care spending in 2019 is $3.81 Trillion\(^2\) almost as high as the world’s 4th largest GDP of Germany. National health care spending in the US is expected to grow at an average annual rate of 5.4\(^3\) from 2019 to 2028, outpacing GDP at 4.3\%. The largest of these programs is Medicare, the federal health insurance program for people of age 65 and older and the disabled. With more than $800 Billion being spent on Medicare in 2019, even small shares of waste and abuse lead to large losses. Federal Bureau of Investigations (FBI) estimates that fraud accounts for 3–10% of all billings ($24–80 billion), and the US Government Accountability Office (GAO) estimates Medicare fraud in 2019 at $46.2 Billion\(^4\). This problem has gained the attention of Medicare administrators faced with the challenge of detecting and deterring waste and fraud to ensure the program stays financially solvent (U.S. Department of Health and Human Services\(^5\) 2022).

The nature of health care fraud provides insights into how it can be detected. Health-care providers face incentives to manipulate billing to increase profits. Yet, in general, patients see multiple providers, and there are many providers in the system that do not commit fraud. Therefore, fraud detection does not rely on the verification of any particular claim, but rather detecting provider-level patterns of care that appear anomalous when considering patient characteristics, medical history, and patterns of behavior by regular non-fraudulent providers.

In this work, we develop new tools to detect Medicare overbilling or fraud. We build a machine learning (ML) framework to discover patterns and detect anomalous providers using the universe of Medicare claims database. Our method focuses on inpatient hospitalization, the largest category of spending and the highest-intensity health care provided by Medicare, which cost the US government $147 Billion in 2021. The proposed approach identifies anomalous providers based on their billing patterns, using patient-level data

---

\(^1\) https://www.advisory.com/daily-briefing/2020/04/03/health-spending


\(^3\) The growth was 36.0\% (!) in 2020 in response to the COVID-19 pandemic.

including medical history, demographics, and geography. We employ our method to identify anomalous patterns among providers and rank them in order of their suspiciousness without using any supervision, that is, not relying on any a priori labeled training data. Moreover, our approach is equipped with explanations to the suspiciousness of the flagged providers, enabling end users like auditors to use our results to guide further investigation.

Our approach is an ensemble method, utilizing three novel unsupervised detection algorithms that uncover aberrant patterns in care across different data modalities. The first component of the ensemble focuses on providers with large observed expenditures conditioned on patient characteristics and medical history. We use a regression based analysis to identify providers with large fixed effects that correspond to high spending per patient even controlling for observable medical history and location of the patient. The second component focuses on coding behavior of claims, uncovering rare ICD-10 medical coding patterns employed by providers, which is indicative of manipulation of specific codes a patient is tagged with in order to garner higher reimbursements. The third component is peer based, focusing on identifying aberrant frequency distribution patterns among a related group of hospitals, where the group of hospitals share similar patient populations and distributions of types of care.

We assemble the evidence from these three detection methods together to rank providers based on suspiciousness. We utilize instant-runoff voting (Franceschini et al. 2022) to reach an aggregate ranking based on three preference orderings for the suspiciousness of providers. This method follows an iterative procedure to include a hospital that is most suspicious based on vote count for top rank in each round. Depending on the number of votes, a flagged provider can be explained by more than one detector.

We validate our approach quantitatively by ground-truth data from the Department of Justice (DOJ). Using a corpus of thousands of DOJ press releases about fraud, we tag providers identified as fraudulent and merge these data to compare with our ranking. While only 1 in 20 hospitals are named in the DOJ Press releases, our ranking substantially improves detection over random sampling: the top 50 providers identified by our method contain 21 providers named in the same DOJ corpus, that an 8-fold lift in detection rate. We

---

5 In this work, the words provider and hospital are used interchangeably. Providers can refer to any health care service provider; we specifically study hospitals.
note that providers ranked high by our method but not listed by the DOJ are not necessarily false positives; rather, enforcement by the DOJ reflects a combination of opportunity to enforce and capacity constraints, and hence only provides partial ground-truth. The DOJ validation is a form of positive-unlabeled data, (Bekker and Davis 2020), and the overlap with our method is therefore a lower-bound of the amount of fraud successfully detected.

In summary, our proposed approach provides scalable and explainable tools to detecting fraud and abuse in health care systems. Our method does not rely on any supervision or data labeling labor, and thus can be readily employed on massive unlabeled data. As the detectors utilize different data modalities and modeling approaches, the explanations also provide different perspective and reasoning into suspicious behavior. This makes our proposed approach useful in practice, as the auditors would be presented with multiple pieces of evidence that support a case and can aid with further investigation.

We foresee that our method could be particularly effective at auditing of health care providers and guiding future enforcement, including beyond just Medicare. As our ranking provides a significant lift in detection rate than one would achieve by random sampling, it can be used to target and prioritize auditing. While our explanations cannot provide legal-standard evidence of bad behavior by providers, they can help sense-making and be used as starting points that guide deeper investigation. Overall, we anticipate that our proposed solution will have value for policymakers, auditors, and enforcers in the health care domain at large.

This paper proceeds as follows. We describe the background and institutions regarding Medicare payment, health care fraud, and fraud enforcement in Section 2 followed by a description of Medicare data in Section 3. Then, we present our detection and explanation methodologies, with an overview in Section 4. Section 5 presents the global expenditure regression-based OD model; Section 6 presents the local ICD subspace based OD model; and Section 7 presents the local/contextual peer-based excess cost OD model. Finally, we present the ensemble model detection results and multi-view explanations on several case studies in Section 8. We conclude with discussion and take-aways in Section 9.

2. Background

In this section, we discuss the institutional details of Medicare fraud. First, we describe the Medicare payment system for inpatient hospitalization, which creates incentives for fraud. Then, we discuss the various types of Medicare fraud and the ways in which it is enforced.
2.1. Medicare Payment System

Medicare uses a prospective payment system (PPS) for inpatient hospitalization, where providers are paid a fixed amount for each patient’s stay, regardless of stay length or cost. Patients are coded with diagnoses and procedure codes based on the International Classification of Diseases (ICD) system, and then based on this coding, each inpatient stay is classified into one Medicare Severity Diagnosis Related Group (DRG). Each DRG is associated with a certain fixed amount per stay, with possible small adjustments (Medpac 2021). The fixed payment for each DRG is based on the average cost of treating patients across all cases in the nation under the same DRG code.

The PPS incentivizes providers to keep the healthcare costs down (Ellis and McGuire 1986) since the provider’s profit is the difference between the fixed DRG payment and the treatment cost. This is in contrast to a reimbursement-based system, where providers would face incentive to incur higher costs for higher reimbursement. However, the PPS may lead to hospitals trying to avoid treating high-cost patients. To address such issues, PPS adjusts the DRG payment (Medpac 2021) to include provider specific factors such as provider’s wage index (geographic factor), patient case-mix to account for patient-population specific treatment cost, teaching and research expenditure, disproportionate share of low-income patients, and number of unusually costly outlier cases.

2.2. Medicare Fraud

Hospitals face incentives to miscode patients; when done intentionally or recklessly, this can qualify as fraud. Because the patient’s ICD coding dictates their DRG and ultimately the hospital reimbursement amount, hospital coding decisions directly affect hospital profits.

Fraud in inpatient hospitalization takes many forms. One well-studied form is upcoding, where hospitals miscode patients to higher severity levels of care in order to receive higher reimbursement (Dafny 2005). A second common issue is lack of medical necessity, where a patient’s health conditions do not qualify them for that care (Howard 2020). Moreover, there is a variety of conduct that can also qualify as health care fraud, discussed below, such as providing compensation to providers for referring patients.

In this paper, we are largely agnostic to which type of fraud hospitals commit, and instead focus on payment levels. In general, fraud is of greatest concern when it results in wasteful spending. Our method detects hospitals whose anomalous conduct results in higher payments, which is valuable even if it does not pinpoint which types of fraud the hospital may have committed.
2.3. Medicare Anti-Fraud Enforcement

The US federal government undertakes a number of initiatives to detect and deter waste, fraud and abuse in Medicare spending. Our method, which relies solely on claims data, is complementary to existing methodologies.

Federal law prohibits Medicare fraud and provides avenues by which fraud can be addressed through criminal and civil enforcement. The federal health care fraud statute provides criminal penalties for those who commit health care fraud, and this enforcement is compounded by criminal enforcement under the anti-kickback statute, as well as the wire fraud and racketeering statutes. Criminal Medicare fraud is prosecuted by the Department of Justice. For a deeper treatment of criminal Medicare fraud, see [Eliason et al. (2021)](#).

Civil enforcement for Medicare fraud operates through the False Claims Act, which provides an avenue for whistleblowers to come forward with information about fraud and receive compensation. Whistleblowers file their own cases in federal civil court, and the DOJ has an option to support these cases. [Leder-Luis (2020)](#) and [Howard (2020)](#) provide more information about the False Claims Act.

In addition to litigation, administrators use a variety of policy tools to limit health care waste, fraud and abuse. The Office of the Inspector General of Health and Human Services undertakes administrative actions against firms that overbill Medicare. Medicare also has a variety of auditing programs that seek to detect unnecessary or unjustified spending; see [Shi (2022)](#) for a description of the Recovery Audit Contractors program. Finally, Medicare uses regulations to target unnecessary spending, such as prior authorization requirements. Some of these regulations combat fraud while others combat waste; see [Brot-Goldberg et al. (2022)](#) and [Eliason et al. (2021)](#) for a discussion of these regulations.

In addition to the enforcement actions listed above, Medicare undertakes some data-driven investigatory work in order to detect fraud. These efforts have received little attention in academic work. Medicare claims processors work with contractors called Unified Program Integrity Coordinators (UPICs) [Noridian Healthcare Solutions (2022)](#) to audit and detect aberrant payments. In addition, Medicare uses a private-public partnership model through the Healthcare Fraud Prevention Partnership to share data and detect health care fraud with patterns similar across a variety of types of care and different health insurance programs [Healthcare Fraud Prevention Partnership (2022)](#). When fraud is identified through these data-driven efforts, investigators can refer those cases to the DOJ for civil or criminal prosecution.
In this paper, we curate for quantitative evaluation a list of hospitals that have been subject to DOJ actions at both the criminal and civil level. While there are many ways in which hospitals could have been investigated or sanctioned, being named in a DOJ press release validates that the hospital was likely committing behavior that rose to the level of criminal or civil fraud, which represents a true positive. A disclaimer, on the other hand, is that the hospitals subjected to DOJ actions likely constitute only a partial list of all fraudulent hospitals, as other unknown fraud and waste may have gone undetected, which represents a false negative.

3. Data Description

This study combines data from a variety of sources to detect anomalous provider spending behavior in Medicare and compare it to ground-truth labeling of providers that have faced anti-fraud enforcement.

Our analysis of provider behavior uses a large-scale dataset of Medicare claims. We consider patients hospitalized in 2017, and we use data from 2012 through 2016 to construct the patients’ medical history. For these years, we use 100% of samples of Fee-For-Service institutional Medicare data, including inpatient and outpatient claims, and beneficiary information including demographic information and chronic condition indicators from the Chronic Conditions Warehouse. To further understand a beneficiary’s medicare history, we use 20% of samples of carrier files, which describe physician office visits.

Table 1 describes the sample of inpatient hospitalization claims from 2017. We observe 11.2 million claims from 6.6 million beneficiaries representing 7,661 different providers. Medicare spent in total $131 billion on inpatient care in 2017, out of $710 billion total reported Medicare spending.

Table 2 describes our sample used to construct patient medical history from 2012 through 2016. We observe nearly a hundred million physician office visits and another hundred million outpatient visits per year, as well as millions of inpatient visits per year.

To understand provider characteristics, we use the Medicare Provider-of-Service files, which contain details on providers such as certification number, name, the type of Medicare services that it provides, and type of ownership (private or public). We can identify

---

6 In this work, the words patient and beneficiary are used interchangeably.

7 20% samples are the largest available for physician office visits.
Table 1  Inpatient data statistics from year 2017

<table>
<thead>
<tr>
<th>Spending</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicare total expenditure [Statista 2022]</td>
<td>$710 billion</td>
</tr>
<tr>
<td>Medicare inpatient expenditure</td>
<td>$131 billion</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Beneficiaries</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of inpatient beneficiaries</td>
<td>6.6 million</td>
</tr>
<tr>
<td>Number of inpatient claims</td>
<td>11.2 million</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Providers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of providers</td>
<td>7,661</td>
</tr>
</tbody>
</table>

Table 2  Scale of data from year 2012 to 2016 used to build medical history of patients who are 70 years or older in the inpatient claims from year 2017. The number in each cell is in millions.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician visits</td>
<td>94.7</td>
<td>100.2</td>
<td>102.8</td>
<td>107.7</td>
<td>114.2</td>
</tr>
<tr>
<td>Outpatient visits</td>
<td>81.5</td>
<td>87.3</td>
<td>90.9</td>
<td>96.8</td>
<td>104.3</td>
</tr>
<tr>
<td>Inpatient visits</td>
<td>4.0</td>
<td>4.2</td>
<td>4.4</td>
<td>5.1</td>
<td>5.8</td>
</tr>
</tbody>
</table>

patients across files using their unique beneficiary identifiers, and we identify providers by their identifiers such as the National Provider Identification (NPI) or CMS Certification Number (CCN). Further, we separately identify Academic Medical Centers based on their membership to Council of Teaching Hospitals. These providers engage in academic research, which may lead them to be ranked as anomalous due to the differences in their claim patterns from other hospitals.

The federal Department of Justice (DOJ) publishes press releases when fraud is identified in order to inform the press and the public as well as deter future fraudulent behavior. To evaluate our automated detection of suspicious providers, we utilize these press releases related to Medicare from the DOJ. To that end, we scraped from the DOJ website thousands of press releases that contain the word ‘Medicare’. Each press release corresponds to a case that the Department of Justice was involved with, often at the time of settlement.

Using partial name matching, we tag the hospitals that appear in this corpus. As the DOJ lacks both the capacity and the information to prosecute all Medicare fraud, the press releases provide only a partial list of providers that have engaged in fraudulent behavior.

4. Method Overview

The Medicare dataset comprises diverse data modalities, which provides an opportunity for modeling the fraud detection problem in various ways. For example, a provider can be represented by the DRG codes associated with its claims, the frequency of ICD codes used in its claims, or by the characteristics of the patient populations that it serves. Each modality presents us with a specific perspective of the data. These different modalities then allow us to learn comprehensive provider behavior which reveal information that cannot be completely uncovered based on only one aspect of the data, since each representation may contain information that is not reflected in others.

In this work, our goal is to estimate a suspiciousness score based on which we rank providers such that anomalous ones are ranked at the top, which may be due to their fraudulent practices. To utilize our different data modalities, we propose an unsupervised multi-view anomaly detection approach, suitable for the underlying multi-modal data. Each view (or base detector) presents itself as a different model of the anomalies, operating on a different data representation. As such, each can be seen as providing evidence that corresponds to a particular reason for detection. The explanation provided by each detector provides a unique perspective into suspicious behavior. Collectively, the evidence from these base detectors, i.e. across modalities, can be assembled systematically into an ensemble detection method.

Ensemble methods utilize multiple base detectors, where under certain accuracy and diversity conditions, they are to obtain better performance than the constituent base detector alone and produce more robust results (Aggarwal and Sathe 2017). Diversity is an important property of ensemble methods, which ensures that the base detectors make independent errors that cancel out when aggregated. Therefore, various approaches have been proposed toward promoting ensemble diversity (Kuncheva and Whitaker 2003, Nam et al. 2021). In essence, our approach utilizes the diversity of the underlying data representations to induce diversity in the ensemble.

In Figure 1, we show the different Medicare data modalities we consider and provide a high level description of the corresponding base outlier detection (OD) model that utilizes
Unsupervised ML for Explainable Medicare Fraud Detection

Medical history
Hospitals

Beneficiaries
Regression

D1
Hospitals
D2
ICD subspaces
Hospitals
D3
Peer hospitals

(a) Global OD (b) Subspace OD (c) Contextual peer-based OD

Figure 1 Multi-view anomaly detection on different Medicare data modalities – D1, D2, and D3. Model (a): Global detector based on fixed effects regression model. The coefficient of a hospital is an indicator of excess cost of care at the hospital. Model (b): Local (in ICD codes) detector in the very high dimensional ICD code frequency representation of hospitals. It explains anomalies based on feature importance, i.e. with respect to specific ICD codes. (c): Local and contextual (peer-based) detector based on comparing DRG frequency distributions. It provides a contrastive explanation in terms of excess cost of treatment when compared to peers.

The first model (a) is set up as a global regression onto cost per beneficiary (target variable) from data (denoted D1 on the figure) reflecting a beneficiary’s medical history and the hospitals that they visited. The second OD model (b) performs typical outlier detection among hospitals as represented by the frequency of ICD codes used in their claims (denoted D2). Anomalous coding may be associated with only a few ICD codes (i.e. features) at a time, rather than all. Therefore, the second model is a feature subspace detector, finding outliers locally in subsets of features. Finally, the third OD model (c) performs contextual detection, identifying hospitals that behave differently from their peers. Behavior is captured by the frequency distribution of the DRG codes assigned to each hospital’s claims (denoted D3). Here, we recognize the heterogeneity among hospitals and compare a hospital’s behavior locally, i.e. in the context of its peers with similar characteristics.

Besides detection, our proposed models can provide explanations for their flagged anomalies. This is especially important in the absence of any ground-truth labels in practice, aiding sense-making, verification and decision making (such as whether to conduct additional investigation, or whether or audit). By capitalization on different data representations, our method leads to different explanations with each OD model, enabling a multi-view
reasoning. Specifically, in Figure 1 the regression coefficient associated with a hospital in our first OD model (a) would be a direct indicator of excess spending at the hospital. The second OD model (b) quantifies feature (i.e. ICD code) importance, and can explain each flagged anomalous hospital based on the specific ICD codes that they use differently in their claims. The last OD model (c) provides contrastive explanations, through comparing DRG frequencies of a hospital to those of their peers. As the DRG code of a claim dictates cost, differences in the DRG coding distribution can be directly translated to excess cost of treatment. Importantly, the explanation can pinpoint which DRGs are most contributing to large excess cost of a hospital.

To arrive at a final anomalous ranking based on different modalities, we combine the rankings from individual detectors such that it captures the agreement among them. In effect, the ensemble approach allows us to gather evidence from multiple models, each leveraging a different data modality. Further, it can be “unrolled” to provide explanations to each flagged anomaly by each detector in the ensemble. Overall, such a multi-view detection and explanation approach takes advantage of corroborating evidences across modalities, and provides a multi-view perspective toward reasoning about suspicious behavior.

The following three sections are organized to present the details of our detection models, in terms of data setup, detection methodology and explanation.

5. Expenditure-Based Detection with Regression
The goal of provider-level analysis of expenditure is to understand which providers are associated with high spending on a beneficiary’s hospitalization. Our design considers the beneficiary’s medical history, using claims from physicians office visits, hospital outpatient visits, and hospital inpatient visits over a five-year period before the target year. The outcome or target variable is the base claim amount per beneficiary per provider in the current year. The modeling assumption is that a patient’s history over the last several years should help explain the expenditure in the current year, and therefore providers that spend high amounts on average per patient, conditional on that medical history, are anomalous.

5.1. Data Setup
Base payment amount. In our analysis, we use base payment amount computed from the Medicare inpatient claims. As explained in Section 2, the Medicare Prospective Payment System adjusts the claim payment amount to include expenses due to provider variables such as patient mix, disproportionate share of low-income patients, outlier cases,
and expenditure on education and research. These factors are generally external to the provider’s coding choice and should be excluded from analysis. Therefore, to understand provider behavior with respect to inpatient encoding, we rely on the base payment amount. The base payment amount is calculated by subtracting the reported adjustment amount from the total claim amount. While payments are also adjusted by provider location through a geographically indexed wage, we do not control for provider wage index adjustments, because the geographical factor will be picked up when controlling for patient location in our regression.

Figure 2a shows the box plot of average total claim amount per provider in the inpatient claims data (year 2017) for the top 50 DRGs, sorted by the mean of the box plot. Notice that there is large variation in the average claim amounts for each DRG. This variance across providers is reduced when the box plot instead uses the average base payment amount as shown in Figure 2b. However, there remains some variance across providers even when considering the base payment amount.

Patient representation. We represent each patient by their medical history and their covariates including location.
We consider patients from 2017 who had an inpatient hospitalization claim and are at least 70 years old. As Medicare is available for those aged 65 and older, we include patients aged 70 years or above to ensure we observe a full 5-year history. We construct the medical history based on a patient’s provider visits in the previous five years (2012–2016). We filter and join patients data from physician visits, outpatient visits, and inpatient hospitalizations in the previous five years. Each patient visit, to a physician or inpatient facility, is assigned codes based on the ICD diagnosis and treatment codes. Thus, for a patient, we collect all the unique codes that were assigned in any of the visits along with their counts.

In addition to the treatment codes, we include the chronic conditions that require regular care, associated with each patient as reported in 2016, the year before the current year. We do not include 2017 chronic conditions as those may be changed based on the code that the hospitalizations report. Including the chronic condition of a patient helps understand any comorbidities that may arise due to their medical history and ongoing chronic condition, accounting for the increase in treatment expense. It provides a comprehensive view of the past treatments received by a patient, and reflects on their medical health. Further, to account for variation due to a patient’s choice of provider, as well as geographic differences in hospital reimbursement rates, we include patient’s residential region. A patient’s region is represented by the first three digits of their zip code.

5.2. Detection Model

To estimate expected treatment expense for a patient, we employ a fixed-effects regression model with the outcome or target variable being the total base payment (which excludes external provider related variables), and the features being the aforementioned patient representation (of medical history and location). We then include as regressors variables corresponding to the count of hospitalizations at each provider. The coefficients of the provider variables give the provider fixed effects that we use to rank providers.

Note that, because we are interested in capturing the provider-level dependency of cost, we do not include treatment codes from the current year’s hospitalization. The codes of the current year’s hospitalization reflect the hospital’s coding decision, which can be an element in its fraud or overbilling behavior. Instead, the providers are added to the model to account for treatment expenses in the current year that are not reflected by the patient’s medical profile; see Figure 1(a).
Regression model specification for expenditure. Given (i) patient representation $X \in \mathbb{R}^{N \times M}$ for $N$ patients, each with a $M$-dimensional representation of historical medical profile based on the last five years (2012–2016), and (ii) the total base payment $Y$ in year 2017; the specification for expected treatment expenditure prediction is as follows.

$$Y_i = \beta_0 + X_i \beta + \sum_j \alpha_j H_{j,i} + \epsilon_i,$$  \hspace{1cm} (1)

where $Y_i$ is the total base payment expense for a patient $i$ in 2017; $X_i$ is the patient representation for $i$, $\beta$ depict regression coefficients associated with patient medical profiles and locations, $H_{j,i}$ is associated with an inpatient Medicare provider $j$ which contains total count of visits to $j$ if patient $i$ visited the provider and 0 otherwise, and $\alpha_j$'s depict the provider fixed effect regression coefficients.

Anomaly scoring. In the expenditure-based regression, a coefficient $\alpha_j$ can be interpreted as the excess treatment cost due to provider $j$ that cannot be captured by patients’ medical profile and location. As such, we can associate the magnitude and sign of this coefficient with the excess spending by a provider, and designate it as its anomaly score.

5.3. Model Explanation
The regression model’s provider ranking in order of anomalousness is easily explainable through the coefficient values. Specifically, each $\alpha_j$ used for scoring and ranking has direct interpretation as the excess expenditure on treatment for a patient when visiting the provider $j$. Therefore, the fixed effects model directly quantifies the excess dollar amount impact of a particular provider, which can be used by an auditor or investigator as model based evidence for inquiry.

5.4. Evaluation
In Figure 3, we show the estimated fixed effects, i.e. the $\alpha_j$ coefficients, for providers from our expected expenditure model. The providers with large fixed effects are ranked at the top and flagged as being of suspiciously expensive. To evaluate the effectiveness of our provider ranking, we use the partial list of known fraudulent providers based on the DOJ settlements, as described in Section 2.3.

In auditing, it is often the case that auditors have a limited budget (time and other resources) for processing red-flags and taking action. Thus, it is critical for the ranking to position the most likely fraudulent providers higher up in the ranked list. We quantitatively
evaluate the targeting of fraudulent providers using two ranking quality metrics, namely (a) Precision-Recall (PR) curve, and (b) Lift curve. The PR curve depicts the positive predictive value (precision) on the y-axis versus the true positive rate (recall) on the x-axis. In audit scenarios with limited budget, a high precision at the top of the ranked list would be useful. Similarly, lift curve measures the targeting effectiveness on y-axis when compared to a random baseline as we move along varying fractions of the ranking on x-axis.

In Figure 4, we report the PR and Lift curves for our fixed effects model, and compare its performance against two simple intuitive baselines. The baseline methods rank the providers based on average total claim amount and average base payment amount, respectively. Note that our fixed effects model is comparatively more effective at targeting fraudulent hospitals, with relatively higher precision and lift at the top positions.

In Figure 5, we report the result of a two-sample test on the fixed effect coefficients as estimated by our model for providers in the DOJ corpus versus the rest of the providers. Notice that the DOJ providers typically have larger fixed effects as compared to others, and their distribution is significantly different as the test rejects the null that the two sets of coefficients are drawn from the same distribution, with \( p < 0.001 \). We remark that the reported performance is conservative and only the lower limit on our model’s targeting ability, since many top ranked providers that are not part of DOJ ground truth may still have been involved in suspicious behavior. We report more qualitative results, and provide case studies through explanations into such flagged providers in Section 8.2, after accounting for the evidences from other models in our ensemble.

6. ICD Coding-Based Detection with Subspace Analysis

International Classification of Disease (ICD) codes are used by Medicare health care providers to characterize a patient’s medical condition and treatment. The US uses ICD-10 codes, which were developed by the World Health Organization and can be used to designate the universe of medical issues and procedure. ICD codes encode provider assessment
We report (a) Precision-Recall curve (AP: Average Precision denotes area-under-curve) and (b) Lift curve for provider ranking produced by fixed effects coefficients against two simple baselines: ranking of providers based on (1) average total claim amount and (2) average base payment amount. Dashed horizontal line 'Base' depicts the random ranking. Notice that top of the ranking is comparatively better as evidenced by higher precision and lift when recall and top data fraction are low. This is particularly helpful for auditors who would typically process only top ranked providers under limited budget.

Comparison of fixed effect coefficients for providers facing anti-fraud lawsuits (known fraudulent entities or outliers) versus the rest of the providers (normal entities). A two-sample test rejects the null hypothesis, implying significantly different distributions statistically.

of a patient based on their reason of visit to the hospital and their medical conditions, and primarily reflect the diagnoses and applied procedures for treatment. The assigned ICD codes are then used as input to a “grouper” software used by hospital billers that assigns a diagnostic code (DRG) based on the provider findings as indicated by the assigned ICD codes. As discussed above, in the Medicare PPS, the DRG code determines the reimbursement level. Consequently, ICD coding presents opportunities for miscoding, as providers may try to achieve a more expensive DRG code to obtain higher reimbursement. Therefore,
the objective of our ICD coding based analysis is to understand provider coding practices
that could reveal the coding patterns applied by providers engaging in fraudulent behavior.

6.1. Data Setup

**Provider representation.** We use inpatient claims from the year 2017 to understand
how providers assign ICD codes to each claim, and represent providers through ICD codes,
including diagnostic and procedure codes. This representation captures the coding practices
of a provider. Importantly, since providers have a choice of ICD codes, we also account for
ICD code substitutability, where a slightly similar ICD code could be used instead to yield
higher reimbursements.

To capture code substitutability, we estimate the semantic similarity of the codes within
each chapter of the ICD code hierarchy. Here, each ICD code is described by concatenating
its text description to the description of its ancestor codes within the ICD hierarchy.
Then, pairwise Jaccard distance is computed between the descriptions of the codes and
the provider representation is updated using the ICD code similarity.

Specifically, let \( X^{ICD} \in \mathbb{R}^{N_H \times M_H} \) be the matrix representation of \( N_H \) providers in terms
of \( M_H \)-dimensional ICD codes in which the entries depict the total code usage count by
provider, and \( J \in \mathbb{R}^{M_H \times M_H} \) be the ICD substitutability matrix consisting of pairwise Jaccard similarities.
Then, the provider representation \( X^{ICD_{sim}} \in \mathbb{R}^{N_H \times M_H} \) after incorporating
the code substitutability is given as \( X^{ICD_{sim}} = X^{ICD} \times J \), which re-distributes each code’s
frequency to substitutable ICD codes that are not directly reported in the claims data.

We note that \( X^{ICD_{sim}} \) is very high dimensional (> 40,000 features) as the inpatient
claims data use the ICD-10 version of these codes. However, anomalous coding of a claim
is likely covert and associate with only a few ICD codes. Therefore, we employ a feature subspace based detector for finding outliers locally among subsets of ICD codes; see
Figure 1(b).

6.2. Detection Model

We employ a suit of subspace outlier detectors on the high dimensional provider representation \( X^{ICD_{sim}} \) to find providers deviating from the majority coding practices within
certain ICD subspaces. As we are interested in ICD subspaces that are relevant for a variety
of aberrant provider practices, we utilize an ensemble of subspace detection methods that
are effective on high dimensional data. In the same spirit as with our overall approach,
the ensemble allows us to examine multiple diverse subspaces as each subspace detection method implements a different methodology for exploring candidate subspaces. In particular, our subspace ensemble uses five different state-of-the-art methods as we describe briefly below. We note that the ensemble is also flexible to incorporate other detection algorithms.

**Subspace outlier detection.** Though we represent a hospital in the high dimensional ICD space, the abnormal or aberrant behavior may be reflected only in a small, locally relevant subset of codes as pertains to stealthy behavior. Each OD algorithm in the ensemble explores local subspaces differently to provide evidences from diverse subsets. To that end, our OD model consists of the following subspace detectors:

(i) Subspace Outlier Degree (SOD) \cite{kriegel2009} locally examines each point (hospital) in the data. For each data point, it computes reference points through shared nearest neighbors. The subspace is then characterized by dimensions with low variance, lower than a provided threshold, within the identified reference set. It records the deviation of each data point from the hyperplane spanned by the mean of the identified subspace, where outliers have larger deviation.

(ii) Isolation Forest (iF) \cite{liu2008} builds a collection of randomized trees that approximate the density of data points in a random feature subspace characterized by paths in what are called “isolation trees”. Each isolation tree is constructed by recursively partitioning data using a randomly chosen point in a randomly selected dimension, until the leaf of the tree contains a single data point. Shorter paths in a tree indicate sparse regions as fewer partitions lead to leaf nodes, and points belonging to each leaf at lower depth indicate outlierness in the subspace characterized by the tree path.

(iii) Robust Random Cut Forest (RRCF) \cite{guha2016}, like iF, also constructs an ensemble of randomized trees by recursively partitioning the data. It computes the model complexity of each tree as the sum of the bits required to store the depths of each point in the tree. An outlier is defined as a point which increases the model complexity significantly when added to the tree.

(iv) Lightweight on-line detector (LODA) \cite{pervn2016} constructs a collection of histograms on random 1-dimensional projections of the data. Each data point is then associated with the negative log-likelihood based on each histogram, and data points are ranked based on their average likelihood across the 1-D histograms.
(v) RS Hash (RSHash) [Sathe and Aggarwal (2016), like LODA, is also an ensemble of histograms; however, it constructs a collection of grid-based histograms in randomly chosen subspaces, and grid sizes vary based on varying sample sizes of data. Each data point is then scored by the number of sampled points sharing the same bin in the histogram. A sparsely populated bin is indicative of outlierness.

We apply the above methods to $X^{ICD_{sim}}$, the ICD representation of providers, and identify the providers that behave abnormally in various subspaces as explored by the algorithms.

**Anomaly scoring.** Each subspace algorithm assigns an anomaly score to each provider. The scores have different scale and semantics (path length, likelihood, etc.), and thus are not directly comparable across the methods. Therefore, we aggregate the ranking of providers based on individual scoring of each subspace method. We use the instant-runoff voting technique (details in Section 8) for rank aggregation from different subspace algorithms, and provide the final ranking of hospitals by anomalousness across all subspaces.

### 6.3. Model Explanation

We explain the ranking of a subspace detector using Shapley Additive Explanation values (SHAP values), introduced in Lundberg and Lee (2017) and Lundberg et al. (2020). SHAP values estimate the feature importance in a regression model by approximating the effect of removing each feature from the model as the average of differences between the predictions of a model trained with and without the respective feature. We regress the anomaly scores from a subspace detector onto the ICD representation of providers, and then estimate the SHAP values under the regression model. The feature contributions for each observation find the most important codes that affect the anomaly score significantly. This helps us find ICD codes that are contributors to aberrant behavior for a provider.

Further, we provide dollar amount characterization of important features (ICD codes). To this end, each ICD code is mapped to the most frequent DRG code assigned for the given ICD code within the inpatient claims. Since DRG codes are determinants of the payment for care, through frequent DRG mapping, we associate dollar amount of reimbursement to ICD codes. This lends itself to understanding the dollar amount impact of an important ICD code for an anomalous provider as explained by SHAP feature importance values.
Figure 6  We report (a) Precision-Recall curve and (b) Lift curve for provider ranking produced by our ICD-10 subspace outlier detector ensemble against two simple baselines that rank the providers based on (1) average total claim amount and (2) average base payment amount. Dashed horizontal line ‘Base’ depicts the random ranking.

6.4. Evaluation

Figure 6 reports the performance of our subspace OD model in terms of the PR and Lift curves. The subspace model ranking is at least $2 \times$ better at targeting fraudulent providers compared to our two baselines, respectively based on total claim payment and base payment amounts. Moreover, it is the lower bound on the performance since the ground truth consists of providers settled with DOJ, while others may not have been caught. We present qualitative explanations for several flagged hospitals in Section 8.

7. Expenditure-Based Detection with Peer Analysis

The objective of the peer-based analysis is uncovering the local patterns of spending behavior among a related group of providers (called peers), and identifying providers deviating from the group’s expected behavior. We utilize the inpatient claims to create a profile for each provider under two complementing data modalities: (1) based on services provided by a hospital, and (2) the patients’ chronic condition profiles served by a hospital. We then find groups of related providers based on the similarity of their provider profile representation. To identify a locally aberrant behavior, each provider is represented in terms of DRG frequency distribution, capturing factors of spending for treatment. Then, the DRG representation of a given provider is compared to the summary DRG distribution of their peers; see Figure 6(c). The providers are then ranked in order of their deviation from group behavior in terms of DRG-based spending.
In short, we identify hospitals that spend excess amounts as compared to hospitals that treat similar patient populations or patients with similar medical conditions. This allows us to identify providers who are exposed to the same patient population but manage to assign more expensive DRG codes.

7.1. Data Setup

Provider representation. We construct hospital profiles to capture the nature of services provided, the characteristics of patient population served, and encoding practices that drive spending for treatment.

Provider profile – Type of services. We first examine a provider’s inpatient claims data to understand the nature of services provided. Because the DRG codes assigned by providers may be manipulated to accomplish higher reimbursement, we do not represent providers by the DRGs they use; instead, we consider the provider’s distribution into major diagnostic categories (MDC). Each MDC corresponds typically to one major body system (circulatory, digestive, etc), and can be associated with a set of medical specialties. Each MDC contains a large set of potential DRGs. Therefore, grouping providers by MDC allows us to consider providers that treat patients with similar types of medical needs, but without relying on the exact DRG codes assigned. For each provider, we collect unique MDC codes used and record the normalized count of an MDC code in the inpatient claims data in the current year.

Provider profile – Patient population. We create another profile based on patient population characteristics served by a provider. The underlying motivation for this profile is that two providers should be similar if they serve patients with similar medical conditions. To characterize the patient population at a broad level, we use the underlying chronic conditions of the patients. The chronic conditions flag whether a patient has received a previous set of services related to a chronic condition such as diabetes or ischemic heart disease. As a provider’s representation, we record the normalized count of the chronic conditions of all the patients treated at the provider.

Provider profile – Spending for care. The spending amount in each claim is directly tied to the assigned DRG code. To capture the DRG encoding practices of a provider, we represent each provider using the normalized counts of DRG codes from its inpatient claims. The DRG frequency representation allows us to compare and contrast the spendings between a hospital and its peers that provide similar services or serve similar patients.

7.2. Detection Model

**Peer identification.** We group as peers those hospitals that share similarities in the type of services provided or the patient population served. Let \( v_j \) denote the representation for provider \( j \); either based on the type of services profile using MDC codes or based on the patient population profile using chronic conditions of patients. We note that the provider representations are frequency distributions, as they depict normalized counts. Therefore, to measure the similarity between two providers \( j \) and \( k \), we use the Hellinger distance for probability distributions, which is an upper bound on the total variation distance Bar-Yossef et al. (2004), given as

\[
d_{jk} = \frac{1}{\sqrt{2}} \cdot \left\| \sqrt{v_j} - \sqrt{v_k} \right\|_2
\]  

Next, we examine the distribution of pairwise similarity values to decide on a threshold \( \tau \) to include only the most similar providers in a provider’s peer group. Then, for each provider \( j \), the providers with similarity to \( j \) above \( \tau \) constitute \( j \)’s peers, denoted \( P_j \). Notice that the peers are specified for each provider separately, rather than using any clustering algorithm. This allows us to create compact peer groups of varying sizes. We note that fixing the peer group size would be a subpar alternative, since \( j \)’s group may then include distant providers as peers, skewing the representative summary statistics of the group that \( j \) is compared to.

**Anomaly scoring.** In the Medicare PPS, the reimbursement amount for treatment is directly based on the assigned DRG code to a claim. Therefore, for anomaly scoring we utilize the provider representations over DRG codes from the inpatient claims, which consist of the normalized counts of the DRG codes used by a provider. For each provider, we have identified a group of providers (peers) with similar characteristics—type of services and patient population served—based on which we create a peer group summary in terms of distribution over DRG codes. The summary distribution is created by incorporating DRG frequencies from all the peers, weighted by their similarity to the provider under consideration. Let \( v_{j, DRG} \) be the DRG distribution for provider \( j \) with \( n_j \) claims, and \( q_{j, DRG} \) be the summary DRG distribution based on provider \( j \)’s peers, defined as follows.

\[
q_{j, DRG} = \frac{1}{Z} \sum_{k \in P_j} n_k \times (1 - d_{jk}) \times v_{k, DRG} \quad \text{where} \quad P_j = \{ k \mid (1 - d_{jk}) \geq \tau \} \quad \text{and} \quad Z = \sum_{k \in P_j} n_k \times (1 - d_{jk})
\]  

(3)
Next we tie the DRG usage frequencies to dollar amount spending by Medicare, as the former dictates the latter. To this end, let $\text{Cost}(c)$ denote the average base price of DRG code $c$ computed from the inpatient claims data from the year 2017. Then, the excess spending for treatment per claim on average for provider $j$ is given as follows.

$$ExcessSpending_j = \sum_{c \in \text{DRGs}} \text{Cost}(c) \times (v_{j,\text{index}(c)}^{\text{DRG}} - q_{j,\text{index}(c)}^{\text{DRG}})$$ \hspace{1cm} (4)$$

where $v_{j,\text{index}(c)}^{\text{DRG}}$ is the frequency corresponding to DRG code $c$ in the DRG representation $v_j^{\text{DRG}}$ for provider $j$, and $q_{j,\text{index}(c)}^{\text{DRG}}$ denotes that for DRG code $c$ in the peer group summary representation $q_j^{\text{DRG}}$.

The calculated $ExcessSpending$ amount is the anomaly score based on which the providers are ranked, as it depicts the average spending discrepancy for a provider when compared to peers of the given provider. Since we create two different peer groupings – one based on services provided, and another based on patients served – we obtain two rankings, later combined through instant-runoff voting (Section 8).

7.3. Model Explanation

The peer based OD model’s anomaly score is the estimated excess spending, which is directly interpretable as the extra dollar amount a provider charges on each claim on average as compared to what would be expected from other similar providers. Further explanation can be provided for a top-ranked provider by contrasting their frequency distribution over DRG codes against their peers. This allows auditors to have a contrastive
understanding of DRG codes used by similar providers, and to pinpoint to specific DRGs with large frequency discrepancies. Direct usage comparison of individual DRGs could point to specific codes that contribute most to the overall spending at a provider, and guide a deeper investigation of the claims associated with those specific DRG codes.

7.4. Evaluation
In Figure 7, we show the distribution of pairwise similarities between hospitals, and mark the similarity threshold at $\tau = 0.8$ which is used in our implementation for identifying peers. We exclude providers from our analysis that have less than five peers for the chosen threshold, as the estimation of excess spending could be noisy for these providers due to small peer group. Providers with large excess spending are ranked at the top and are identified as suspicious.

We use the DOJ corpus to evaluate our ranking of the providers based on excess spending. In Figure 8, we report the PR and Lift curves for our peer analysis. The ranking is also compared to the two baselines, respectively ranking providers by average total claim amount and average base payment amount. Although the peer-based ranking performance is comparable to these simple baselines, we remark that it is the lower bound on the performance. Furthermore, besides a mere ranking and unlike these simple baselines, our model can provide a nuanced explanation through DRG code frequency discrepancies, providing auditors with reasoning for potential factors driving the high spending. Finally, our model fundamentally identifies expensive hospitals as compared to their peers, which may be of interest to auditors interested in waste that may not rise to the level of fraud detected by the DOJ.

Through case studies in Section 8, we report further qualitative results and provide peer-based explanations and insights into top flagged providers after aggregating evidences from different OD models.

8. Aggregate Provider Ranking
Our ensemble method is designed to handle multi-view Medicare data. Each OD model (component) of the ensemble considers a different data modality and creates a ranked list of providers based on the evidence examined individually. To arrive at the final ranking for auditing, we merge multiple rank lists into a single ranking. Our goal is to present the aggregate ranking that is most representative of the component models. To that end, we
use instant-runoff voting (IRV) to combine results across rankings, as IRV yields ranked outcomes that better represent voting preferences (Franceschini et al. 2022).

The rank aggregation proceeds in an iterative manner, where each round utilizes IRV procedure to find a “winner” (in our case, most suspicious hospital). In each round, votes are counted for each ranking’s first choice, and a hospital with majority of votes is then ranked at top in our aggregate ranking. The rank lists across models are updated to drop the selected hospital in this round, and IRV procedure is repeated with updated rank lists in the subsequent rounds to arrive at aggregate ranking.

In our implementation, we aggregate 8 different rankings across our 3 OD models; one from the regression model, five from different subspace OD algorithms, and two from the peer-based model utilizing two separate similarity measures. To wrap up, we show the effectiveness of our final aggregate ranking for identifying fraudulent hospitals in the Medicare system through quantitative and qualitative evaluations.

8.1. Quantitative Evaluation

In Figure 9, we evaluate our aggregate ranking of hospitals using a PR curve and a Lift curve. The aggregate ranking is compared to intuitive baselines that rank hospitals based on their average reimbursements, or random auditing. Our aggregate ranking is able to target fraudulent providers on average twice as better when compared to the baseline ranking—note the area-under-curve, or average precision (AP) values on legend Figure 9(a).
While only 1 in 20 hospitals are named in the DOJ Press releases, the top 50 hospitals identified by our aggregate ranking contain 21 providers named in the DOJ corpus. That is an 8-fold lift in detection rate considering the evaluation at top 50 hospitals, with an average of 4-fold lift over random/by-chance targeting across varying data fractions as seen in Figure 9(b). Again, it is the lower bound on performance, since our ground-truth consists only of providers named in the DOJ corpus, while there may be others with yet unidentified fraudulent practices.

8.2. Qualitative Explanation: Case Studies

In this section, we present an analysis of our multi-view detectors, highlighting some of the salient aspects for the fraud detection task. In particular, we discuss how our multi-view detectors can be used to explain the aberrant patterns employed by top ranked flagged hospitals by highlighting parts of data from different views that contributed most to the ranking, which can assist in the process of auditing or deeper investigation.

We examine two top ranked providers from the aggregate ranking (see Table 3 in Appx. A): (1) the provider at rank 1 that is also named in the DOJ corpus, and (2) the provider at rank 5 which is not in our ground truth (as ranks 1–4 all are part of DOJ ground truth). In the following two case studies, we show how different models contribute evidences toward a better understanding of how each provider stands out.
Case 1: Flagged hospital named in DOJ corpus  Our aggregate ranking finds the Cleveland Clinic as the most suspicious hospital under our metrics. Here we present evidence from our 3 Outlier Detection models, where this provider is ranked at #1 by the subspace OD model, ranked at #17 by the peer-based model, and ranked at #27 by our regression-based model.

Notably, the Cleveland clinic settled with the DOJ in the years 2015 and 2021 for $1.74 million\(^{10}\) and $21 million\(^{11}\) respectively. The evidence from our models do not directly match the reason for DOJ settlements; put differently, our explanations have not been validated externally by litigation. Moreover, our data do not provide evidence of fraud by the Cleveland Clinic, nor do they substantiate claims from lawsuits against the Clinic. The existence of previous lawsuits by the DOJ against the Clinic validate that this is a provider with past bad behavior, and our metric indicates that this provider engaged in anomalous behavior that can be detected by our algorithm and merits deeper investigation.

Our regression model estimates the excess expenditure on treatment for a patient when visiting the Cleveland Clinic to be $29,844.33, which is almost 3\times the average expenditure (\(\approx$10K\)) as shown in Figure 3. This does not, by itself, indicate that the Cleveland Clinic engaged in bad behavior, as this may reflect that it performs more specialized medical procedures; although our regression accounts for the patient’s recent medical history.

\(^{10}\)https://www.justice.gov/opa/pr/32-hospitals-pay-us-more-28-million-resolve-false-claims-act-allegations-related-
kypoplasty

\(^{11}\)This is through their acquired hospital https://www.justice.gov/opa/pr/northern-ohio-health-system-agrees-pay-
over-21-million-resolve-false-claims-act-allegations

![Figure 10](image)  ICD codes contributing to suspiciousness of top ranked providers based on SHAP values
In Figure 10a, we plot the most important ICD codes that contribute to the anomaly score of the provider from the subspace OD model, based on SHAP values. The top ICD code “T782XXD” is described as “Anaphylactic shock, unspecified, subsequent encounter” which falls under the ancestor “T78” with the description: “Adverse effects, not elsewhere classified”. As such, T78 appears to be a catch-all classification for adverse effects for injuries, poisoning, and other consequences of external causes for visit. Moreover, the code T782XXD is considered exempt from reporting whether the condition is present on admission (POA) to an inpatient facility. The next ICD code “T783XXD” is under the same ancestor, T78, and is also considered exempt from reporting if POA. Similarly, the description of code “M12862” allows non-specific reasons to be used for encoding as the given description is: “Other specific arthropathies, not elsewhere classified, left knee”.

We next examine the reimbursement amounts related to these ICD codes, based on their mapping to the DRG they are most frequently associated with. The distribution of the amounts across all ICD codes is given in Figure 11. The codes T782XXD and T783XXD can be mapped to two DRG codes: 949 (Aftercare with cc/mcc) and 950 (Aftercare without cc/mcc). The reimbursement amount for DRG code 949 is about 25% more compared to DRG code 950, where T782XXD is reported most frequently against DRG code 949. Further, within the ICD-10 hierarchy, codes T782XXD and T783XXD are most expensive.

12 https://www.icd10data.com/ICD10CM/Codes/S00-T88/T66-T78/T78-
13 Here, ‘cc’ and ‘mcc’ stand for Complication or Comorbidity and Major Complication or Comorbidity, respectively.
and get at least 50% more reimbursement than any other sibling or parent code. Notably, 6 out of top 10 ICD codes contributing to anomaly score (as shown in Figure 10) have reimbursement amounts that are more than 50th percentile among all ICD codes, while 3 of them associate with DRG codes with amount above the 90th percentile (see Figure 11). All these factors explain, through specific ICD codes, associated DRGs and dollar amounts, the reasoning behind why a flagged provider stands out. These evidences provide starting points for further investigation.

In the peer-based model, the provider is flagged through peer relation of providers with respect to MDC representation. Figure 12 shows the MDC distribution of the Cleveland Clinic and its nearest peer provider. Notice that in terms of facilities and services provided as encoded by their MDC, the two hospitals are quite similar. We compare the DRG representation of the Cleveland Clinic to the summary DRG representation of all its peer hospitals over the top 50 DRG codes that are selected based on their contribution to excess spending (see Eq. 3 for excess spending estimate). As shown in Figure 13, Cleveland Clinic uses certain DRG codes more frequently than its peers as indicated by the summary distribution—starting with 219, 220, as well as 309, 310, 330. DRG codes 219 and 220 belong to “Cardiac Valve and Other Major Cardiothoracic Procedures” with reimbursement amount in top 4 most expensive within MDC 05. DRG codes 309, 310 are described as “Cardiac Arrhythmia and Conduction Disorders”, and DRG code 330 is described as “Major small and large bowel procedures with cc”. Note that the description of codes 309, 310 and 330 is specific to a particular condition, while the description for 219–220 allows for ambiguity. Ambiguity may provide opportunities for miscoding to reach for higher reimbursement.
In summary, all three OD models point to evidence from different views of the claims data that makes the top ranked hospital stand out from others, both in terms of local and global analysis. These pieces of evidence explain the ranking by shedding light into certain coding practices a provider engages in, and may be utilized in further audit processes.

**Case 2: Flagged hospital not in DOJ corpus** In the aggregate ranking, “AdventHealth Orlando” is ranked at #5 in order of suspiciousness. Although AdventHealth Orlando is not named in DOJ, we illustrate the evidences from our method to explain the ranking, where the provider is ranked at #5 by the subspace OD model, and ranked at #35 by the peers-based model.

In Figure 10b, we present the bar plot of the top 10 ICD code importances for the provider, based on SHAP values for the anomaly ranking from our subspace OD model. Note that 5 out of these top 10 ICD codes fall under ICD-10 chapter “S00-T88 Injury, poisoning and certain other consequences of external causes”. The ICD code T270XXA is most frequently mapped to DRG code 205 which is described as “Other respiratory system diagnoses with mcc”. The 3rd ranked ICD code “I70268” is described as “Atherosclerosis of native arteries of extremities with gangrene, other extremity”. Based on the descriptions of these top ICD codes, a common thread appears to be that the codes leave room for ambiguity—due to the catch-all word ‘other’ in their descriptions. Further, 7 out of 10 ICD codes have reimbursement amount larger than the 50th percentile, and 4 out of 10 have...
reimbursements larger than 90th-percentile reimbursements across all ICD codes (recall Figure 11 for the ICD price distribution).

Next we present evidence from the peer-based OD model, though the provider is not top ranked in this model. Figure 14 shows the provider and its nearest peer hospital that serve similar patient populations, represented in terms of chronic conditions of the patients. We note the almost identical distributions of chronic conditions for the provider and its nearest peer hospital. We compare the DRG distribution of the provider to the summary DRG distribution of its peers.
In Figure 15, we show the distribution over the top 50 DRG codes, where the provider’s distribution deviated from the summary distribution the most weighted by DRG reimbursement amount (see Eq. 3). We find that excess expenditure is almost entirely driven by two DRG codes, namely 291 (heart failure and shock with mcc) and 470 (major joint replacement or reattachment of lower extremity without mcc) with reimbursement costs larger than the 50th-percentile among DRG codes.

Similar to the earlier case, our models pinpoint specific ICD and DRG codes that can help jump-start further investigation, while highlighting dollar amount discrepancies that provide perspective with respect to monetary value.

9. Conclusion
The unsupervised ensemble method in this paper provides a data-driven approach to identifying fraud among hospitals using massive Medicare claims data. Our approach uses different modalities of data – including patient medical history, provider coding patterns, and provider spending – to detect anomalous behavior consistent with fraud. We validated our method quantitatively using Department of Justice (DOJ) press releases about fraud that name hospitals. We also presented qualitative case studies based on the explanations by our detection models that pinpoint specific ICD and DRG codes associated with excess spending at a provider.

Our method substantially outperforms baseline algorithms. We combine evidence from multiple unsupervised outlier detection algorithms that use different types of global and local analysis – estimating a hospital’s impact on patient expenditure, identifying few ICD codes that a hospital uses differently than the norm, and comparing a hospital’s distribution over DRGs to its peers – using which we create a final ranking of suspiciousness. While only 1 in 20 hospitals are named by the DOJ as fraudulent, 21 of our top 50 hospitals are in the same corpus, an 8-fold improvement in detection rate.

Besides the lift in detection rate, which enable higher hit-rate under limited investigation budget, our proposed fraud detection approach is fully unsupervised and explainable—two desirable properties as they relate to key human factors. Unsupervised detection waives the need for laborious labeling by human experts. Explainable tools are user-friendly and empower human-in-the-loop investigation, verification and decision making.

Therefore, the methods developed in this paper would be valuable if implemented by policymakers for detecting hospital overbilling. Medicare spends over a hundred billion of
dollars per year on hospitalizations, and even small shares of fraud can be very expensive. Our detection algorithm can be used to guide auditing by identifying which providers are committing the most egregious behavior. Moreover, our method explains which patterns drive the detection, which can facilitate auditing once a hospital is selected by allowing an investigator to focus on certain billing codes and types of care.

The methods developed here provide a set of tools whose usability extends beyond Medicare. Medicaid, the federal-state partnered low-income subsidy program, spends an additional $400 Billion per year. Moreover, private health insurance companies each face similar issues when evaluating providers who overbill for non-federally-insured patients. With health care spending at 19.7% of US GDP, tools for detecting health care fraud could find wide-ranging use.

References


A. Top ranked hospitals

We present the top 20 of the ranked hospitals by different components of our method. The hospitals named in DOJ corpus are marked by *.

Table 3  Ranking of hospitals by our methods in order of their suspiciousness. The hospitals named in DOJ corpus are marked by *.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Regression</th>
<th>Subspace</th>
<th>Peer-based</th>
<th>Aggregate Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Parkland Health And Hospital System*</td>
<td>Cleveland Clinic*</td>
<td>St. Joseph’s University Medical Center</td>
<td>Cleveland Clinic*</td>
</tr>
<tr>
<td>2.</td>
<td>Us Pain &amp; Spine Hospital</td>
<td>New Hanover Regional Medical Center*</td>
<td>Dixie Regional Medical Center</td>
<td>New Hanover Regional Medical Center*</td>
</tr>
<tr>
<td>3.</td>
<td>Minimally Invasive Surgery Hospital</td>
<td>Tampa General Hospital*</td>
<td>South Texas Health System Behavioral*</td>
<td>Tampa General Hospital*</td>
</tr>
<tr>
<td>4.</td>
<td>Westchester Medical Center*</td>
<td>Massachusetts General Hospital</td>
<td>Corpus Christi Medical Center</td>
<td>Massachusetts General Hospital</td>
</tr>
<tr>
<td>5.</td>
<td>Harris Health System</td>
<td>AdventHealthOrlando</td>
<td>Avera Sacred Heart Hospital</td>
<td>AdventHealth Orlando</td>
</tr>
<tr>
<td>6.</td>
<td>Beth Israel Deaconess Medical Center*</td>
<td>Southcoast Hospital Group, Inc</td>
<td>Camden Clark Medical Center</td>
<td>Rush University Medical Center*</td>
</tr>
<tr>
<td>7.</td>
<td>Nyu Langone Hospitals</td>
<td>Lancaster General Hospital*</td>
<td>St. Luke’s Cornwall Hospital</td>
<td>Methodist Hospital</td>
</tr>
<tr>
<td>8.</td>
<td>New York-Presbyterian Hospital</td>
<td>Beth Israel Deaconess Medical Center*</td>
<td>Faith Community Hospital</td>
<td>North Mississippi Medical Center*</td>
</tr>
<tr>
<td>9.</td>
<td>Mount Sinai Hospital*</td>
<td>Orlando Health Orlando Regional Medical Center*</td>
<td>Texas Health Harris Methodist Hospital Alliance</td>
<td>Mount Sinai Hospital*</td>
</tr>
<tr>
<td>10.</td>
<td>Jacobi Medical Center</td>
<td>Memorial Mission Hospital And Asheville Surgery Ce</td>
<td>Nyack Hospital</td>
<td>Orlando Health Orlando Regional Medical Center*</td>
</tr>
<tr>
<td>11.</td>
<td>Boston Medical Center Corporation</td>
<td>Mount Sinai Hospital*</td>
<td>St Francis-Downtown</td>
<td>Brigham And Women’s Hospital</td>
</tr>
<tr>
<td>12.</td>
<td>North Central Bronx Hospital</td>
<td>Ohio State University Hospitals</td>
<td>Eastern Niagara Hospital - Lockport Division</td>
<td>North Shore University Hospital</td>
</tr>
<tr>
<td>13.</td>
<td>University Hospital Of Brooklyn (Downstate)</td>
<td>Hackensack University Medical Center</td>
<td>Jersey Shore University Medical Center</td>
<td>Montefiore Medical Center</td>
</tr>
<tr>
<td>14.</td>
<td>Queens Hospital Center</td>
<td>St Vincent Hospital &amp; Health Services</td>
<td>Mercy Catholic Medical Center-Mercy Fitzgerald</td>
<td>Duke University Hospital*</td>
</tr>
<tr>
<td>15.</td>
<td>Montefiore Medical Center</td>
<td>York Hospital*</td>
<td>Virtua Willingboro Hospital*</td>
<td>Mercy Hospital Springfield*</td>
</tr>
<tr>
<td>16.</td>
<td>Pims Indian Hospital At Pine Ridge</td>
<td>Houston Methodist Hospital</td>
<td>Easton Hospital*</td>
<td>Baylor University Medical Center*</td>
</tr>
<tr>
<td>17.</td>
<td>Brooklyn Hospital Center - Downtown Campus</td>
<td>Northwestern Memorial Hospital</td>
<td>Cleveland Clinic*</td>
<td>Baptist Medical Center*</td>
</tr>
<tr>
<td>18.</td>
<td>Regional One Health</td>
<td>Memorial Hermann Hospital</td>
<td>Advocate Christ Hospital &amp; Medical Center</td>
<td>Uf Health Shands Hospital</td>
</tr>
<tr>
<td>19.</td>
<td>Ivinson Memorial Hospital</td>
<td>Methodist Hospital</td>
<td>Healthresource Saginaw</td>
<td>Houston Methodist Hospital</td>
</tr>
<tr>
<td>20.</td>
<td>Campbell County Memorial Hospital</td>
<td>Baylor University Medical Center*</td>
<td>St Lukes Hospital Of Kansas City</td>
<td>University Of Michigan Health System</td>
</tr>
</tbody>
</table>