Journal of Policy Analysis and Management

RESEARCH ARTICLE

Can Machine Learning Target Health Care Fraud? Evidence From Medicare Hospitalizations

Shubhranshu Shekhar[1] | Jetson Leder-Luis[2] | Leman Akoglu[3]

[1]Brandeis University, Waltham, Massachusetts, USA | [2]Boston University and NBER, Boston, Massachusetts, USA | [3]Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Corresponding author: Jetson Leder-Luis (jetson@bu.edu)

Received: 18 November 2025 | **Accepted:** 18 November 2025

Keywords: anomaly detection | explainable AI | fraud and abuse | health care | machine learning | Medicare

The United States spends more than $4 trillion per year on health care, largely conducted by private providers and reimbursed by insurers. A major concern in this system is overbilling and fraud by hospitals, who face incentives to misreport their claims to receive higher payments. In this work, we develop novel machine learning tools to identify hospitals that overbill insurers, which can be used to guide investigations and auditing of suspicious hospitals for both public and private health insurance systems. Using large-scale claims data from Medicare, the US federal health insurance program for the elderly and disabled, we identify patterns consistent with fraud among inpatient hospitalizations. Our proposed approach for fraud detection is fully unsupervised, not relying on any labeled training data, and is explainable to end users, providing interpretations for which diagnosis, procedure, and billing codes lead to hospitals being labeled suspicious. Using newly collected data from the Department of Justice on hospitals facing anti-fraud lawsuits, and case studies of suspicious hospitals, we validate our approach and findings. Our method provides a nearly fivefold lift over random targeting of hospitals. We also perform a postanalysis to understand which hospital characteristics, not used for detection, are associated with suspiciousness.

JEL Classification: I13, C19, D73, K42, M42

1 | Introduction

Fraud in health care is hard to detect. Insurers face information asymmetries, where providers know more about the health care delivered than the insurer responsible for paying for that care. Health care providers such as doctors and hospitals 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 by manipulating the provided documentation.

These issues are compounded in the U.S. federal health care programs, where the government is the insurer. The U.S. federal government spends over a trillion dollars per year on health insurance, largely paid to private firms, and fraud detection is challenging due to the sheer volume of claims being processed. In 2019, Medicare (the largest of these programs) spent $800 billion, and even small shares of fraud lead to large losses, taking away funds for valuable care. The U.S. Government Accountability Office (GAO) estimates Medicare improper payments, a measure of mistaken or inappropriately documented spending, in 2019 at $46.2 billion (GAO 2020). This problem has gained the attention of Medicare administrators faced with the challenge of detecting and deterring fraud and abuse to ensure the program stays financially solvent (U.S. Department of Health and Human Services 2022).

Machine learning poses a potential solution to the problem of health care fraud detection, but has been hampered by the

challenges of health care claims data, which are highly complex and multidimensional. The government observes health care claims including diagnostic, procedural, and billing information, amounting to tens of thousands of potential categorical codes that can be used. Moreover, methods that are based on matching known patterns of fraud, which could be used for supervised machine learning, are biased by the fact that enforcement is nonrandom, and will likely fail against the ever-changing nature of fraud in health care.

In this work, we develop and validate new machine learning tools to detect health care overbilling and fraud, which can be used to guide anti-fraud investigations. First, we construct a data formalism for understanding health care claims at the diagnostic-procedural, billing, and spending levels, which allows for the detection of rare patterns between types of claims that can be expected to be similar. We then apply recent anomaly detection tools from the computer science literature to detect anomalous providers based on their coding patterns and its effects on hospital spending. These methods rely on the fundamental idea that providers mostly behave like their peers in the absence of fraud, and that deviations from this pattern are more suspicious when they earn the hospital more money. Our approach is unsupervised, that is, does not need any a priori labeled training data, meaning it is not biased by labeled data from past enforcement, going beyond known patterns of fraud.

We apply our method using millions of claims from inpatient hospitalizations, the largest category of Medicare spending, which costs the U.S. government more than $100 Billion per year. Our detection method does not flag any particular hospitalization claim as suspicious, but rather detects hospital-level patterns of care that appear anomalous when considering patient characteristics, medical history, and patterns of behavior by other hospitals when treating similar conditions and patients. This method ranks hospitals in order of their suspiciousness, and is an "explainable" rather than a black-box method, in that it can provide explanations for which types of codes are most anomalous (i.e., potentially misused) for each hospital in our ranking. By ranking hospitals and providing explanations for each ranking, our method speaks to the government's prioritization problem of choosing hospitals for additional scrutiny (such as auditing) with limited enforcement capacity.

We validate our approach with newly collected ground-truth data from the Department of Justice (DOJ). Using a corpus of thousands of DOJ press releases about fraud, we tag hospitals ever named by the DOJ, and compare these data with our ranking. While only 1 in 12 hospitals nationwide has ever been named in the DOJ press releases, our ranking substantially improves detection over random sampling: The top 50 hospitals identified by our method contain 21 hospitals named in the same DOJ corpus, a nearly 5-fold lift in detection rate. We note that hospitals highly ranked by our method but not listed by the DOJ are not necessarily false positives; DOJ enforcement depends on opportunity and capacity constraints, providing only partial ground-truth. The DOJ validation resembles positive-unlabeled data (Bekker and Davis 2020), and the overlap with our method is therefore a lower-bound for detected fraud.

Our algorithm is an ensemble method, utilizing three novel unsupervised detection algorithms that uncover aberrant patterns in care across different levels of claims data, from the most fine-grained to the most broad. The first component focuses on the coding behavior within claims, uncovering unusual ICD-10 procedure and diagnosis coding patterns employed by hospitals, which is indicative of manipulation of a patient's codes to garner higher reimbursements. The second component is peer-based, focusing on identifying aberrant hospital stay-level billing code (DRG) patterns, compared to peer hospitals that share similar patient populations and distributions of types of care. The third component of the ensemble focuses on hospitals with large observed expenditures conditioned on patient characteristics and medical history, using a regression-based method. To assemble the evidence from these three detection methods together to rank hospitals based on suspiciousness, we utilize instant-runoff voting (Franceschini et al. 2022), which combines information from our different detectors to reach an aggregate ranking. This method follows an iterative procedure to rank the hospital that is most suspicious based on the "vote" across different detectors in each round.

The results of our analysis also provide evidence that is qualitatively consistent with detecting fraud, rather than singling out legitimate anomalies such as rare or specialty care. We compute the top ICD diagnosis and procedure codes that contribute to identifying hospitals as suspicious. These codes tend not to be rare conditions that are expensive to treat, but rather diagnoses with high payment rates and high ambiguity, indicating they can be more easily manipulated. We also provide case studies of two hospitals to show how our method can be used to dive deeper into data, and show the exact diagnosis and billing codes that make those hospitals suspicious. Additionally, we perform an explanatory analysis of the types of characteristics—not used for detection—that are correlated with hospital suspiciousness. We also explore the relationship between hospitals targeted by our method and by the DOJ, and discuss the relationship to existing methods used by the DOJ and by the Centers for Medicare and Medicaid Services for detecting improper hospital behavior.

Our approach has many potential applications for health care policymakers, auditors, and enforcers. While our explanations cannot provide legal-standard evidence of fraud by hospitals, they can be used as starting points that guide deeper investigation such as audits or claim reviews, prioritizing the most suspicious hospitals. Our method can also be readily adapted to detect overbilling in other areas of potentially fraudulent care besides hospitals, such as outpatient claims. Moreover, while the data set on which we build our method is from Medicare, we anticipate our methods will prove valuable to private insurers as well, who face nearly identical challenges in eliminating fraud from private health insurance systems.

This paper proceeds as follows: We describe background and institutions in Section 2 and our data in Section 3. Section 4 provides an overview of our detection methodology, with details in Sections 5–7. Section 8 reports the ensemble model detection results, and Section 9 provides a postanalysis that characterizes hospitals with high estimated suspiciousness. Section 10 compares to existing enforcement, and Section 11 concludes.

2 | Background

In this section, we discuss the institutional details of Medicare fraud. While many of the institutional details about Medicare claims and enforcement are specific to the federal system, the general nature of health care billing is consistent across both publicly funded and private-payer systems.

2.1 | Hospital Billing

Medicare implements a Prospective Payment System (PPS) that uses Diagnosis-Related Groups (DRGs) to determine fixed payments for hospital services based on patients' diagnoses and procedures, encouraging cost-effective care while ensuring adequate reimbursement. 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 DRG. DRGs can be surgical, if they reflect a major surgery, or medical otherwise. DRGs can also reflect patients' comorbidities and complications; for example, DRG 460 reflects noncervical spinal fusion surgery without major comorbidities and complications, while DRG 459 reflects the same procedure on patients with comorbidities and complications. Some DRGs reflect up to three levels of comorbidities and complications: no comorbidities and complications; comorbidities and complications; or major comorbidities and complications. These are all based on the ICD diagnosis and procedure codes reported as part of the visit.

Because the patient's ICD coding dictates their DRG and ultimately the hospital reimbursement amount, hospital coding decisions directly affect hospital profits. Hospitals do not receive additional reimbursement for providing higher quality services, or a higher volume of procedures for a given patient, although hospital payments are adjusted for high-level factors such as local wage variation and share of medical students trained. MedPAC (2023) presents more details about the hospital prospective payment system.

2.2 | Health Care Fraud

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; Silverman and Skinner 2004; Becker et al. 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, such as providing compensation to providers for referring patients, which qualifies as a kickback. Because hospital DRGs are based on patient diagnoses and procedures, hospitals can garner higher reimbursements by reporting additional diagnoses or comorbidities; by miscoding diagnoses to be more severe; or, in some cases, by actually performing medically unnecessary procedures to justify higher reimbursement.

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 higher levels of spending. Our method detects hospitals whose anomalous conduct results in higher payments, which is valuable for detecting hospitals where additional auditing is of highest marginal value.

2.3 | Health Care Anti-Fraud Enforcement

The U.S. government undertakes a number of initiatives to detect and deter fraud, waste, and abuse in federally funded health care spending. Our method, which relies solely on claims data, is complementary to existing methodologies. Private insurers face similar challenges and also work to detect, investigate and enforce against fraudulent providers, although they lack the full weight of the federal investigatory system.

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 DOJ. For a deeper treatment of criminal Medicare fraud, see Eliason et al. (2025).

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 (2025) and Howard (2020) provide more information about the False Claims Act and show that these whistleblowers provide high deterrence effects.

In addition to litigation, administrators use a variety of policy tools to limit health care fraud, waste 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 et al. (2022) and Eliason et al. (2025) for a discussion of these regulations.

In addition to the enforcement actions listed above, Medicare and private insurers undertake some data-driven investigatory work in order to detect fraud. These efforts have received little attention in academic work. We survey existing data-driven enforcement and compare our methods to existing methods in Section 10.

In this paper, we curate a list of hospitals that have been subject to DOJ actions at both the criminal and civil level, used for quantitative evaluation of our method. 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 may have gone undetected, which represents a false negative. We use hospitals named in press releases across all years, not just after our sample period. This is to reflect the fact that hospitals that are ever caught committing health care fraud are likely to commit fraud in the future; more than 50% of hospitals named in press releases are named more than once.

2.4 | Related Methodological Work

In addition to the economic studies listed above that discuss health care fraud, several data-centric approaches have been explored in the context of Medicare fraud. We refer the reader to R. Bauder et al. (2017), Kumaraswamy et al. (2022), and Joudaki et al. (2015) for detailed surveys on different methods.

In early work, Rosenberg et al. (2000) study upcoding within the claims data. They estimate the probability that a claim has incorrect DRG code, which they further use to identify claims to investigate and audit. Brunt (2011) studies upcoding in the physician office visits data, where he estimates the likelihood of a disease code selected for an office visit to understand upcoding practices. Fang and Gong (2017) find evidence of provider overbilling using inappropriately high number of hours worked to identify outliers.

Recently, Chandola et al. (2013) and Suresh et al. (2014) introduce methods for provider profile comparison to spot possible misuses or fraud. These works focus on introducing methods and features to represent hospital profiles for comparison; however, they do not present any conclusive results. On the other hand, Bauder and Khoshgoftaar (2018), Bauder and Khoshgoftaar (2018), Herland et al. (2018), and R. A. Bauder and Khoshgoftaar (2017) utilize publicly available excluded providers to learn models for detection of fraudulent providers. However, these approaches rely on the availability of human labeled information on fraudulent information, which is often incomplete and hard to obtain for massive Medicare data.

In contrast to earlier methods, unsupervised and explainable methods for the problem, which are more practical in the real world, have received limited attention. Luo and Gallagher (2010) analyze DRG distributions of hospitals providing services for hip replacements and heart attacks to find upcoding, with the underlying assumption that most hospitals will have similar distributions. Recently, Ekin et al. (2019) learn joint distribution of medical procedures and providers using outpatient data. The joint distribution is used to identify provider anomalies based on procedure code and usage frequency by the provider.

Most existing research uses only a fraction of the massive Medicare data, relies on labeled data on known fraud, and often does not incorporate an explanation of results that could be useful to guide deeper investigation. Our method builds upon these studies to provide a precise and explainable detection method that does not rely upon the existence of labeled data.

3 | Data Description

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

TABLE 1 | Description of inpatient data sample from year 2017.

Spending	
Medicare inpatient expenditure	$80 billion
Beneficiaries	
Number of beneficiaries	4.6 million
Number of inpatient claims	7.3 million
Hospitals	
Number of hospitals	2207

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

	2012	**2013**	**2014**	**2015**	**2016**
Physician visits	15.7	16.9	18.1	20.0	23.4
Outpatient visits	14.3	15.7	17.1	19.0	22.1
Inpatient visits	1.0	1.1	1.4	1.9	5.2

Our hospital anomaly detection method uses a large-scale dataset of Medicare claims. Data were accessed through a data use agreement with the Centers for Medicare and Medicaid Services, facilitated by the Research Data Assistant Center (ResDAC) and the National Bureau of Economic Research (NBER). These hundreds of millions of observations contain extensive data about each hospitalization and patient in the Medicare system, providing an ideal corpus with which to study hospital behavior.

We consider all patients hospitalized in 2017. We filter to inpatient acute care hospitals, whose hospital names are available from CMS Medicare Inpatient hospitals public use files, and drop hospitals that treated fewer than 11 patients to comply with data suppression rules. This leaves a sample of 2207 hospitals, and we use all patients who visited these hospitals in 2017. We use data from 2012 to 2016 to construct the patients' medical history. For these years, we use 100% samples of Fee-For-Service institutional Medicare data, including inpatient and outpatient claims, and beneficiary information including demographic information and chronic condition indicators. 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 that we analyze. We observe 7.3 million claims from 4.6 million beneficiaries representing 2207 different hospitals. We see $80 billion in inpatient spending related to our data set, out of $710 billion total reported Medicare spending (Annual Report 2022).

Table 2 describes our sample used to construct patient medical history from 2012 to 2016. We observe tens of millions physician office visits outpatient visits per year, as well as millions of inpatient visits per year. Appendix A in the Supporting Information

provides additional details about the cleaning and use of the Medicare data.

To understand hospital 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 patients across files using their unique beneficiary identifiers, and we identify hospitals by their CMS Certification Number (CCN). Further, we separately identify Academic Medical Centers based on their membership to Council of Teaching Hospitals (AAMC 2022).

The federal DOJ publishes press releases when fraud is identified in civil or criminal lawsuits, in order to inform the press and the public as well as deter future fraudulent behavior. To evaluate our automated detection of suspicious hospitals, 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 DOJ was involved with, often at the time of settlement. Using partial name matching, we tag the hospitals that appear in this corpus. This also accounts for hospital chains, when the chain name appears in both the hospital name and the press release.

As the DOJ lacks both the capacity and the information to prosecute all Medicare fraud, the press releases provide only a partial list of hospitals that have engaged in fraudulent behavior. We can consider this a form of positive-unlabeled data: while we can identify firms that have been named in a press release as having likely committed fraud, firms that are *not* named are not necessarily above suspicion. In general, we consider the DOJ data as a partial ground truth, as a sensible though possibly incomplete way to measure whether the firms identified by our metric are validated as the firms that have at some point committed fraud, and therefore deserve additional scrutiny. Appendix B in the Supporting Information provides additional details about the collection and cleaning of the DOJ corpus.

4 | Method Overview

Medicare claims data contain many levels of detailed information about hospital claims, including procedure and diagnosis codes, claim-level billing codes, and patient characteristics, which provide opportunities for modeling the fraud detection problem in various ways. For example, a hospital can be represented by the frequency of ICD (diagnosis and procedure) codes used in its claims, the DRG (billing) codes associated with its claims, or by the characteristics of the patient populations that it serves. Each data modality presents us with a specific perspective of the data, which when combined allow us to learn comprehensive hospital behavior which reveal information that cannot be completely uncovered based on only one aspect of the data.

In this work, our goal is to estimate a suspiciousness ranking for hospitals. We use an unsupervised *multi-view* anomaly detection approach, suitable for the underlying multimodal 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 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 2021). In essence, our approach utilizes the diversity of the underlying data representations to induce diversity in the ensemble.

Figure 1 shows the different Medicare data modalities we consider and provide a high level description of the corresponding base outlier detection (OD) model that utilizes it. The first OD model, Figure 1a, performs outlier detection among hospitals as represented by the frequency of ICD codes used in their claims (denoted **D1**). Anomalous coding may be associated with only a few ICD codes (i.e., features) at a time, rather than all. Therefore, this model is a feature subspace detector, finding outliers locally in subsets of features. The second OD model, Figure 1b, 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 **D2**). Here, we recognize the heterogeneity among hospitals and compare a hospital's behavior locally, that is, in the context of its peers with similar characteristics. Finally, the third OD model, Figure 1c, is set up as a global regression onto cost per beneficiary (target variable) from data (denoted **D3** on the figure) reflecting a beneficiary's medical history and the hospitals that they visited.

In addition to 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, which allows investigators to determine why a hospital is ranked as an outlier, and facilitating decisions such as whether to conduct additional investigation or to audit. By capitalizing on different data representations, our method leads to different explanations with each OD model, enabling a multiview reasoning. Specifically, in Figure 1, the selected subspace in the first OD model (a) 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 second OD model (b) 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, facilitating auditing. The regression coefficient associated with a hospital in our last OD model (c) is a direct indicator of excess spending at the hospital.

To arrive at a final anomalous ranking based on different modalities, we combine the rankings from individual detectors to capture the agreement among them. We use instant-runoff voting (Franceschini et al. 2022) to combine information from our

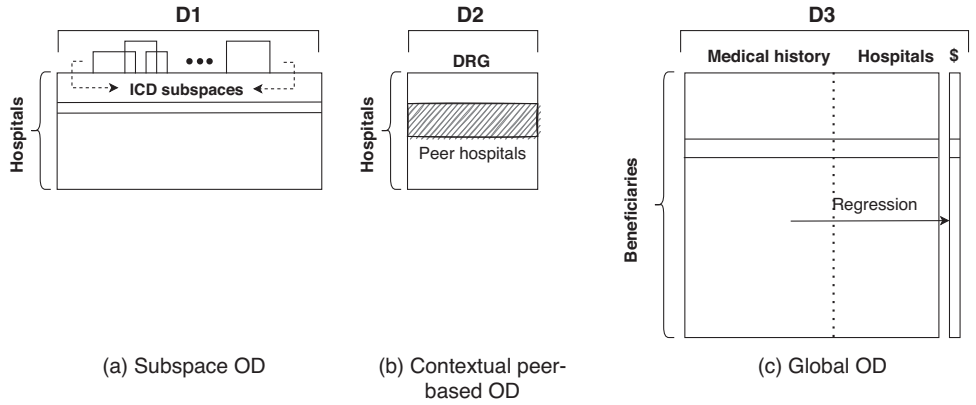


FIGURE 1 | Multi-view anomaly detection on different Medicare data modalities—D1, D2, and D3. (a) Local (in ICD codes) detector in the very high-dimensional ICD code frequency representation of hospitals. It explains anomalies based on feature importance, that is, with respect to specific ICD codes. (b) 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. (c) Global detector based on fixed effects regression model. The coefficient of a hospital is an indicator of excess cost of care at the hospital.

different detectors. Our ensemble approach allows us to gather evidence from multiple models, but can also be "unrolled" to provide explanations to each flagged anomaly by each detector in the ensemble.

There are some important caveats about our method. First, our analysis is entirely based on admitted patients. To the extent that hospitals vary in their admissions patterns, and that over-admission is a different form of health care fraud, we will not be able to detect these forms of fraud. Second, our method relies on comparisons between hospitals, and therefore is better at detecting uncommon frauds. If all hospitals regularly engage in a low-level fraud, our method will not detect it; however, we are encouraged by the fact that such diffuse frauds would be more likely to be detected by other methods, (such as whistleblowing), due to the number of institutions and individuals they would involve. Lastly, we study the annual records across hospitals collectively, without comparing trends over time; although our approach may be extended by considering historical time windows of a hospital as its "self-peer" which is out of scope of our current work.

The following three sections are organized to present the details of our three detection models, in terms of data set up, detection methodology and explanation. Then, we present the results of our aggregate ranking in Section 8.

5 | Detection Through ICD Coding Subspace Analysis

ICD codes are used by health care providers to characterize a patient's medical condition and treatment. The United States uses ICD Version 10 codes, which were developed by the World Health Organization and can be used to designate the universe of medical issues and procedures. ICD codes encode hospital assessment 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. For Medicare billing, the assigned ICDcodes are then used as input to a "grouper" software used by hospital billers that assigns a diagnostic code (DRG) based on the hospital's 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 hospitals 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 hospital coding practices that could reveal the coding patterns applied by hospitals engaging in fraudulent behavior.

5.1 | Data Setup

5.1.1 | Hospital Representation.

We use inpatient claims from the year 2017, gathering ICD codes from each claim. We represent hospitals through their reported ICD codes, including diagnostic and procedure codes.

Importantly, since hospitals have a choice of ICD codes, we must 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 description of each code within each chapter of the ICD code hierarchy. Here, the description of each ICD code is constructed 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 hospital representation is updated using the ICD code similarity.

For example, the description of ICD code J45.20 under chapter X is constructed by concatenating the descriptions of J00-J99 chapter, J40-J47 block, J45, and then the ICD code J45.20, resulting in the description given as "Diseases of the respiratory system—Chronic lower respiratory diseases—Asthma—Mild intermittent asthma uncomplicated." This code will be similar to other codes that contain the word "asthma" or "respiratory system." This representation also ensures that codes nearby in the ICD hierarchy

have somewhat similar text descriptions and are therefore near each other in Jaccard distance.

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

We note that $\boldsymbol{X}^{ICD_{sim}}$ is very high dimensional (> 40,000 features). 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. Figure 1a shows this setup.

5.2 | Detection Model

We employ a suite of subspace outlier detectors on the high-dimensional hospital representation $\boldsymbol{X}^{ICD_{sim}}$ to find hospitals 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 hospital 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 that we describe briefly below.

5.2.1 | Subspace Outlier Detection.

While 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—that is, only certain codes will be fraudulently or suspiciously substituted. Each OD algorithm in the ensemble explores local subspaces differently to provide evidence from diverse subsets. To that end, our OD model consists of the following subspace detectors: SOD (Kriegel et al. 2009), IF (Liu et al. 2008), RRCF (Guha et al. 2016), LODA (Pevnỳ 2016), and RSHASH (Sathe and Aggarwal 2016). Details of each method are included in Appendix C in the Supporting Information.

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

5.2.2 | Anomaly Scoring.

Each subspace algorithm assigns an anomaly score to each hospital. 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 hospitals 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.

5.3 | 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 feature importance 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 hospitals, 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 a hospital being ranked as an outlier.

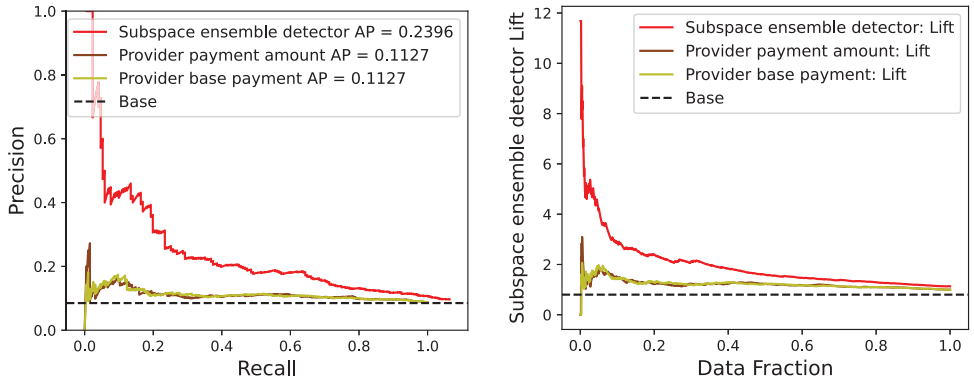
Further, we provide dollar amount characterization of important features (ICD codes). In practice, combinations of ICD codes are used to determine DRG claim codes, which correspond to payment. For our purposes, each ICD code is mapped to the most frequent DRG code assigned for the given ICD code within the inpatient claims. Since DRG codes are determinants of the payment for care, through this most-frequent DRG mapping, we associate dollar amount of reimbursement to ICD codes. This lends itself to understanding the dollar impact of an important ICD code for an anomalous hospital as explained by SHAP feature importance values.

5.4 | Evaluation

To evaluate the effectiveness of our ranking, we use the partial list of known fraudulent hospital based on the DOJ press releases described in Section 2.3, and we compare our suspicious hospitals to known fraudulent hospitals. We quantitatively evaluate the targeting of fraudulent hospitals using two ranking quality metrics, namely a Precision-Recall (PR) curve and a 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. Those responsible for detecting and enforcing fraud in Medicare have a limited budget, and therefore have to select limited targets for deeper investigation; therefore, a high precision at the **top** of the ranked list is most useful. Similarly, the 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 the *x*-axis.

Figure 2 reports the performance of our subspace OD model in terms of the PR and Lift curves, using the DOJ ground truth. The ICD subspace model ranking alone is at least 2× better at targeting fraudulent hospitals compared to our two baselines, respectively based on total claim payment and base payment amounts.



(a) Precision-recall curve

(b) Lift curve

FIGURE 2 | We report (a) Precision-Recall curve and (b) Lift curve for hospital ranking produced by our ICD-10 subspace outlier detector ensemble against two simple baselines that rank the hospitals based on (1) average total claim amount and (2) average base payment amount. Dashed horizontal line "Base" depicts the random ranking.

Our method substantially outperforms random auditing or even detection based strictly on payment amounts.

6 | Expenditure-Based Detection With Peer Analysis

Our second model is based on peer-based excess spending detection and examines the coding decisions of hospitals as compared to similar "peer" hospitals that treat similar populations. In short, we identify hospitals who are exposed to the same patient population but manage to assign more expensive DRG billing codes.

The objective of the peer-based analysis is uncovering the local patterns of spending behavior among a *related* group of hospitals called peers, and identifying hospitals deviating from the group's expected behavior. We utilize the inpatient claims to create a profile for each hospitals under two complementing data modalities, based on: (1) type of services provided by the hospital, and (2) the patients' chronic condition profiles served by a hospital. We then find groups of related hospitals based on the similarity of this representation.

To identify a locally aberrant behavior, each hospital is represented in terms of its DRG frequency distribution, which determines spending. Then, the DRG representation of a given hospital is compared to the summary DRG distribution of their peers. Figure 1c visualizes this setup. The hospitals are then ranked in order of their deviation from group behavior in terms of DRG-based spending.

6.1 | Data Setup

6.1.1 | Hospital Representation.

We construct hospital profiles to capture the nature of services provided (using major diagnostic categories or MDCs), the characteristics of patient population served (using patient's chronic conditions), and encoding practices that drive spending for treatment (using DRGs). The details of the representation are provided in Appendix D in the Supporting Information.

6.2 | Detection Model

6.2.1 | Peer Identification.

We create peer groups of hospitals that share similarities in the type of services provided or the patient population served.

Let $\boldsymbol{v}_j$ denote the representation for hospital 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 hospital representations are frequency distributions, as they depict normalized counts. Therefore, to measure the similarity between two hospitals 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

$$d_{jk} = \frac{1}{\sqrt{2}} \cdot \|\sqrt{\boldsymbol{v}_j} - \sqrt{\boldsymbol{v}_k}\|_2, \tag{1}$$

where ν is a vector of percentage of visits by MDC, computed as number of visits divided by sum of visits. We examine the distribution of pairwise similarity values to decide on a threshold τ to include only the most similar hospitals in a hospital's peer group.

For each hospital j, the hospitals with similarity to j above τ constitute j's peers, denoted $\mathcal{P}_j$. Notice that the peers are specified for each hospital 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 hospitals as peers, skewing the representative summary statistics of the group that j is compared to.

6.2.2 | 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 hospital representations over DRG codes from the inpatient claims, which consist of the normalized counts of the DRG codes used by a provider. In short, this detection mechanism assumes that hospitals who treat similar patient populations, or provide care for similar categories of illnesses and injuries, should have similar DRG distributions.

For each hospital, we create a peer group summary in terms of distribution over DRG codes among all peers, weighted by their similarity to the hospital under consideration. Let $\boldsymbol{v}_j^{DRG}$ be the DRG distribution for hospital j with n_j claims, and $\boldsymbol{q}_j^{DRG}$ be the summary DRG distribution based on hospital j's peers, defined as follows:

$$\begin{aligned}\boldsymbol{q}_j^{DRG} &= \frac{1}{Z}\sum_{k\in\mathcal{P}_j} n_k \times (1-d_{jk}) \times \boldsymbol{v}_k^{DRG} \quad \text{where} \quad \mathcal{P}_j \\ &= \{\, k \mid (1-d_{jk}) \geq \tau \,\} \text{ and } Z = \sum_{k\in\mathcal{P}_j} n_k \times (1-d_{jk}) \quad (2)\end{aligned}$$

Next we tie the DRG usage frequencies to average dollar amount spending by Medicare, as the former dictates the latter. $Cost(c)$ denotes 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:

$$ExcessSpending_j = \sum_{c\in DRGs} Cost(c) \times (\boldsymbol{v}^{DRG}_{j,\,\text{index}(c)} - \boldsymbol{q}^{DRG}_{j,\,\text{index}(c)}) \quad (3)$$

where $\boldsymbol{v}^{DRG}_{j,\,\text{index}(c)}$ is the frequency corresponding to DRG code c in the DRG representation $\boldsymbol{v}_j^{DRG}$ for provider j, and $\boldsymbol{q}^{DRG}_{j,\,\text{index}(c)}$ denotes that for DRG code c in the peer group summary representation $\boldsymbol{q}_j^{DRG}$. In short, this amount computes how much more a hospital spends because they use a different set of DRG codes than their peers, based on the average price of those DRGs.

The calculated $ExcessSpending$ amount is the anomaly score based on which the hospital are ranked, as it depicts the average spending discrepancy for a hospital when compared to peers of the given hospital. 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).

6.3 | 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 hospital charges on each claim on average as compared to what would be expected from other similar hospitals. Further explanation can be provided for a top-ranked hospital 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 hospitals, 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 hospital, and guide a deeper investigation of the claims associated with those specific DRG codes.

6.4 | Evaluation

We have included the evaluation details in Appendix E in the Supporting Information. Additionally, through case studies in Appendix H in the Supporting Information, we report qualitative results and provide peer-based explanations and insights into top flagged providers after aggregating evidences from different OD models.

7 | Expenditure-Based Detection With Massive Fixed-Effect Regression

As a third and final detector, we consider a hospital-level analysis of expenditure to understand which hospitals are associated with high spending on a beneficiary's hospitalization. The incentive of providers who commit fraud is to receive higher reimbursement, and so unexplained high expenditure is potentially suspicious. Our design detects high expenditures that are unexplained by a patient's medical history, which could reflect unnecessary or excessive billing. While any individual patient may receive entirely necessary high levels of care—for example, in response to a severe accident—when a hospital's patient population consistently shows expensive, unexplained high expenditure, this may be indicative of fraud or waste.

Our design considers expenditure as a function of a patient's medical history. We collect each 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 variable of the regression is the base claim amount per beneficiary per hospital visited in the current year. Below we provide our model specification, and in Appendix F in the Supporting Information, we give detailed description of data, algorithm and anomaly scoring.

7.1 | Regression Model Specification for Expenditure

Given (i) patient representation $\boldsymbol{X} \in \mathbb{R}^{N\times M}$ for N patients, each with a M-dimensional representation of historical medical profile based on the last 5 years (2012–2016), and (ii) the total base payment Y in year 2017; the specification for expected treatment expenditure prediction is as follows:

$$Y_i = \beta_0 + \boldsymbol{X}_i\,\boldsymbol{\beta} + \sum_j \alpha_j\, H_{j,i} + \epsilon_i \ , \quad (4)$$

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

8 | Aggregate Provider Ranking

Each outlier detection model presented above is a component of our ensemble method that considers different aspects of the data and creates a ranked list of providers. This ensemble method is designed to handle multi-view Medicare data, where different features of the data can be used to evaluate different aspects of suspiciousness. The goal of the ensemble is a single suspiciousness ranking for all providers.

To arrive at the final ranking for auditing, we merge multiple rank lists into a single ranking using instant-runoff voting (IRV). Our goal is to present the aggregate ranking that is most representative of the component models. IRV combines results across rankings in a way that best reflects the information contained across multiple models (Franceschini et al. 2022).

The rank aggregation proceeds in an iterative manner, where each round utilizes the IRV procedure to find a "winner" (in our case, most suspicious hospital). In each round, votes are counted for each component ranking's first choice, and a hospital with a 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 the IRV procedure is repeated with updated rank lists in the subsequent rounds to arrive at an aggregate ranking.

In our implementation, we aggregate eight different rankings across our three OD models; five from different subspace OD algorithms on the ICD data, two from the peer-based model utilizing the two separate similarity measures (patient populations and categories of care), and one from the regression model. Next, we show the effectiveness of our final aggregate ranking for identifying fraudulent hospitals in the Medicare system through quantitative evaluations. In addition, we consider the time required to run these models in Appendix G in the Supporting Information.

In Appendix H in the Supporting Information, we provide qualitative evaluations through case-studies on suspicious providers from our method, highlighting some of the salient aspects for the fraud detection task. We show how our method highlights parts of data from that contributed most to the ranking of a hospital as suspicious, which can assist in the process of auditing or deeper investigation. Appendix I in the Supporting Information goes further, highlighting the exact ICD codes that greatest contribute to our top-ranked provider suspiciousness. An examination of the codes indicates that providers are not being flagged for treating rare diagnoses, but rather for using diagnosis and procedure codes that reflect *ambiguity* and are relatively highly paid. The intentional use of ambiguous diagnosis codes may reflect an attempt to reclassify patients into more obscure diagnoses to achieve a higher-paid DRG code.

8.1 | Quantitative Evaluation

Figure 3 shows the evaluation of 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 hospitals on average twice as better when compared to the baseline ranking—note the area-under-curve, or average precision (AP) values on legend Figure 3a.

While only 1 in 12 hospitals is named in the DOJ press releases, the top 50 hospitals identified by our aggregate ranking contain 21 hospitals named in the DOJ corpus. That is a 4.9-fold lift in detection rate considering the evaluation at top 50 hospitals, with an average of 2-fold lift over random/by-chance targeting across varying data fractions as seen in Figure 3b. Importantly, our ground-truth consists only of hospitals named in the DOJcorpus, while there may be others with yet unidentified fraudulent practices—and therefore, our list can be used to find other hospitals not yet identified as fraudulent.

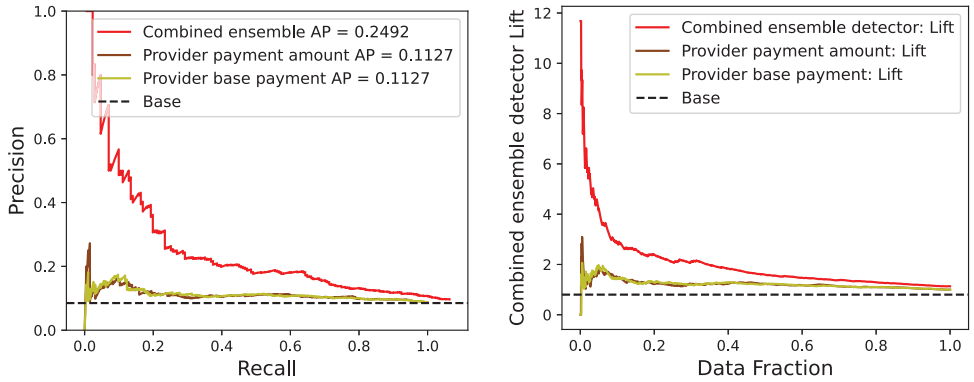
For robustness, we repeat our method on (i) claims originating from emergency room (ER), and (ii) claims excluding Medicare Advantage patients. Appendix L in the Supporting Information presents the method and results for ER data. This is a different set of patients and claims, who are less likely to have selected the hospital themselves, and therefore this addresses potential selection bias, wherein sicker patients choose more sophisticated hospitals. Despite the sample limitation, we still achieve a substantial improvement over random targeting, 1.4× lift among the top 50 hospitals. Appendix N in the Supporting Information presents the method and results after excluding patients with any Medicare Advantage (HMO) coverage. These patients are covered by third parties and therefore we may not observe their full medical history. Among this sample, our method achieves 3× lift over random targeting among the top 50 hospitals.

8.1.1 | Statistical Significance of Ranked Results

We use the $\sqrt{\epsilon}$ test proposed by Chikina et al. (2017) to evaluate the statistical significance of the ranked results. The test is based on a rigorous Markov chain framework, namely comparing our ranked result with a random ranking from a well-defined, valid claims data Markov chain. A state in our Markov chain represents possible claims data resulting from a beneficiary seeking potential treatment at a nearby hospital. We provide the details of the test and validity constraints in Appendix M in the Supporting Information. Note that due to the scale of our data, running the validity checks while constructing the Markov chain is nontrivial, details in Appendix M in the Supporting Information. We run the test on the Markov chain (with $k = 2^{30}$ steps) obtained from our experiment data. Our ranked results are statistically significant at the 5% level, $p < 0.05$.

9 | Characterizing Outlier Providers

In this section, we examine the covariates of hospitals to understand the factors that characterize an outlier hospital as detected by our model. The covariates used in the analysis depict various hospital characteristics such as hospital rating, number of unique patients served, ownership type, location, and length of stay for an inpatient visit. These features are derived from publicly available information for all Medicare hospitals and importantly are *not* included in the data used for detection.



(a) Precision-recall curve

(b) Lift curve

FIGURE 3 | We report the performance of the final ranking of providers as aggregated from eight rankings based on three different OD models. Note that aggregated ranking improves over the ranking by individual constituent experts. The proposed ensemble is on average 2× better than the random targeting of hospitals for auditing.

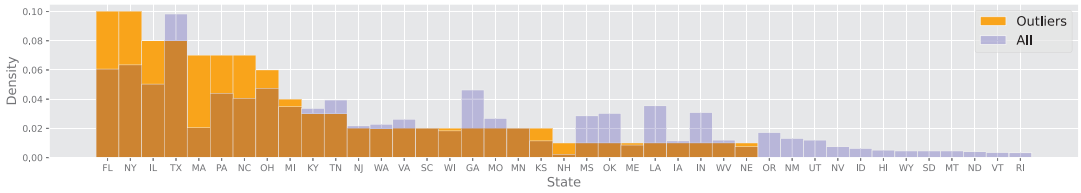
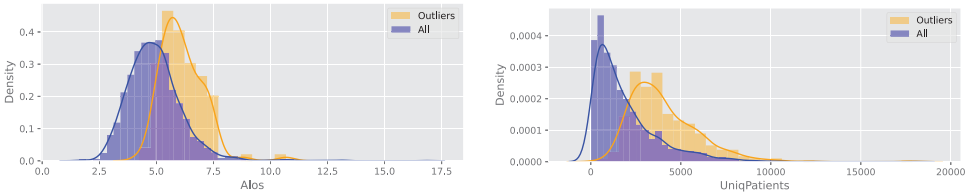


FIGURE 4 | Comparison of distributions over states for outliers and all hospitals.



(a) Average Length of Stay

(b) #Unique Patients Served

FIGURE 5 | Comparison of distributions over covariates for outliers and all hospitals.

Understanding the factors that drive outlier hospital behavior is crucial for improving the health care sector. Extensive policy reforms seek to shape the structure of the health care market, increasing regulations on providers deemed to be harmful or inefficient. By characterizing the nature of hospitals deemed suspicious by our metrics, we hope to contribute to the ongoing literature that evaluates how various interventions—for example, those targeting for-profit care—can affect fraudulent behavior.

Figure 4 compares the distributions over states where a hospital is located. Outlier hospitals are more likely to be from states Florida, New York, Illinois, and Massachusetts, and less likely to be from Texas and Georgia. This is also corroborated by the DOJ cases, where about 15% of the named hospitals are based in Florida.

Figure 5 compares the distributions across average length of stay and number of unique patients served. Ranked outlier hospitals keep inpatients longer as compared to other hospitals. This could be to justify the usage of costlier DRGs, or driven by ranked outlier hospitals receiving sicker patients; however, our metrics control for patient health characteristics. Additionally, top ranked fraudulent hospitals serve more unique patients on average. Since a large fraction of our top ranked hospitals are also named by the DOJ, it may indicate that a greater number of unique patients may provide more opportunity for perturbations in diagnosis coding resulting in higher reimbursements, or it could reflect the fact that our outliers are largely urban hospitals. We also include results for categorical covariates, for example, hospital rating, ownership type, location type in Appendix J in the Supporting Information.

10 | Comparison to Existing Enforcement

Our method stands in contrast to existing work by the DOJ to identify health care fraud. In this section, we undertake three exercises. First, we examine the use of data by the DOJ to detect health care fraud, using publicly available sources on their methods. Second, we consider which types of hospitals have been

subject to DOJ enforcement, compared to the hospitals targeted by our method, and draw contrast between these populations. Finally, we compare our method to the PEPPER Program, a set of basic statistics used by the Medicare program to flag potential improper payments in inpatient hospitalizations (Centers for Medicare and Medicaid Services 2023), and examine the relationship between those variables and the types of ICD codes flagged by our analysis.

10.1 | Existing DOJ Data-Driven Enforcement

Our method shows the ability of machine learning to detect health care fraud, raising questions about the existing usage of data analysis by the DOJ.

The DOJ conducts very limited and basic usage of data analysis to drive anti-fraud investigations. In a 2023 interview with FedTech, the DOJ describes their data analysis techniques for health care fraud detection using a basic approach as follows:

> ”The DOJ uses two types of models. In the first approach, it finds suspected fraudulent providers by examining the characteristics of medical professionals and others who were prosecuted for healthcare fraud in the past. Through analytics, they find current providers who share those characteristics…
>
> It then seeks out national outliers and ranks providers with a scoring system. For example, the DOJ investigates physicians who order more cancer genetic tests than 99.9 percent of doctors in the country.
>
> The second modeling approach is to analyze billing and other healthcare data to find trends in fraud. For example, in 2019, the DOJ saw a spike in Medicare spending on durable medical equipment.”

From this interview, we can glean that the DOJ is using basic analytics, such as sorting providers by total billing and looking for outliers, or looking at time trends. This approach has a number of limitations. First, providers are not compared to their peers, and therefore this approach is likely to be of limited value in hospitalizations, which are much more complicated than single billing codes. Moreover, this is only able to detect overt frauds (e.g., being the top biller in the country for a genetic test), rather than more subtle frauds, like manipulation of underlying diagnoses and procedure codes. Notably, this also relies on the government to manually distinguish which codes to analyze (e.g., genetic tests), while our method allows for the machine learning algorithm to highlight areas of concern without manual input.

The use of data by the DOJ is relatively new. We examined the DOJ website for its description of its health care fraud enforcement activities, and then traced this page back using the Wayback Machine (Internet Archive). In 2017, 2018, 2019, and mid-2020, the government does not mention “data” or “analytics.” Data analytics are first mentioned in September 2020 (DOJ 2020).

Next, we examine the text of the DOJ press releases we use in this paper to find hospitals that were enforced against and consider the extent of data usage. Of the 449 DOJ press releases that mention the hospitals in our data, (done through matching the names of hospitals into press releases, that is, the sample to which we are comparing), only 15 (3%) mention the keyword “data,” and 0 mention “machine learning.”

In contrast, the DOJ largely relies on whistleblowers to initiate civil False Claims Act lawsuits to target hospitals. Leder-Luis (2025) discusses these cases. Whistleblowers are generally hospital employees or other insiders, and therefore data are not used for detection, although data are sometimes used *after a case is filed* to help support the claims or estimate damages. Of the same 449 press releases, 247 of them (55%) contain “whistleblower.”

Two other avenues exist for detection of fraud, with less public details available. 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 between the federal government and private insurers to detect health care fraud with patterns similar across a variety of types of care and different health insurance programs. When fraud is identified through these data-driven efforts, investigators can refer those cases to the DOJ for civil or criminal prosecution.

10.2 | Comparison of DOJ Targets vs. Our Top Hospitals

Next, we compare the characteristics of hospitals that are subject to DOJ enforcement versus those that our algorithm flags as suspicious. Section 9 discusses the hospitals flagged by our method. We repeat these analyses, instead characterizing the hospitals targeted by the DOJ instead of our method.

Figure 6 shows the distribution of states where DOJ targets hospitals. This is in contrast to Figure 4, which shows the states targeted by our method. Notably, DOJ offices are regional and have wide autonomy over the types of cases they pursue, leading to heterogeneity in enforcement patterns across regions (Eliason et al. 2025).

As expected, we see wide heterogeneity in DOJ enforcement, with some states (notably Florida and Pennsylvania) showing heavy DOJ targeting relative to their share of the hospital population. Florida, Pennsylvania, and New York combined cover more than one third of DOJ hospital fraud litigation nationwide. This is in contrast to the states targeted by our method, shown in Figure 4. While Florida and New York are high on the list of hospitals we flag, they contribute lower shares of total enforcement, and other states like Massachusetts, Illinois, and North Carolina also contribute a much greater share of outliers than they have experienced DOJ enforcement. One potential reason is that DOJ offices are responsible for much more than hospital health care fraud enforcement, and so states with high other burdens of types of care to monitor may have limited DOJ enforcement bandwidth—Massachusetts, for instance, conducts

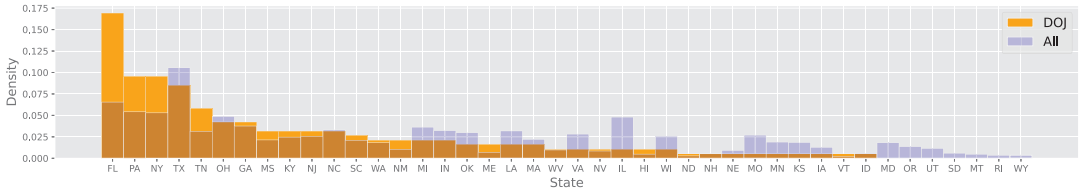
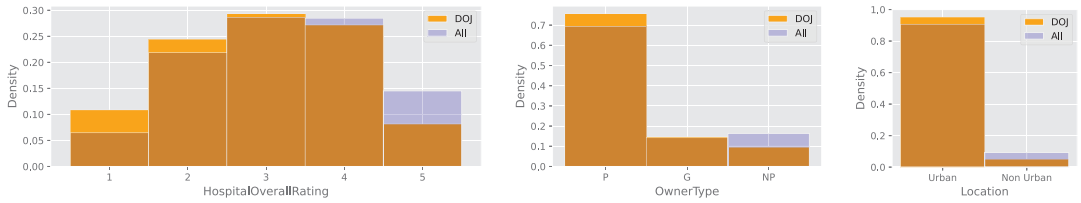


FIGURE 6 | Distributions over states for DOJ enforcement hospitals.



(a) Hospital rating (scale 1 to 5)

(b) Hospital ownership type: Private, Govt, Non-profit

(c) Hospital location

FIGURE 7 | Distributions over categorical covariates for DOJ enforcement hospitals.

extensive pharmaceutical fraud enforcement due to companies headquartered there.

Next, we consider the characteristics of hospitals targeted by the DOJ. Figure 7 presents these results, which are parallel to the characteristics of hospitals we find, shown in Online Appendix Figure 13.

Regarding ratings, as shown in Figure 7a, we see that the DOJ tends to target hospitals with low ranking, and largely does not enforce against hospitals with high rankings. In contrast, the ratings of our outlier distribution is much closer to the ratings distribution of all hospitals. Our method is slightly more likely to tag highly rated hospitals as outliers.

In terms of ownership, as shown in Figure 7b, the DOJ is more likely to name a for-profit hospital in a press release, but also slightly more likely to name a government-owned hospital (than their share of the hospital population). Nonprofits are less likely to be named. Our outlier detection method is even more likely to highlight for-profits, but somewhat less likely to tag government hospitals.

Finally, comparing urbanicity in Figure 7c, the DOJ is more likely to target urban hospitals, and somewhat less likely to target non-urban hospitals, though it is possible. Our metric almost exclusively picks up urban hospitals, showing even more concentration of potentially suspicious behavior in cities than the DOJ enforcement suggests.

Taken together, these results indicate that our methods are not just replicating DOJ enforcement, but rather capturing distinct dimensions and hospitals engaged in suspicious behavior.

10.3 | Comparison to PEPPER Variables

The Centers for Medicare and Medicaid uses a program called PEPPER, the Program for Evaluating Payment Patterns Electronic Report, to flag hospitals for potential improper payments, an indicator of potential fraud (Centers for Medicare and Medicaid Services 2023). We considered the metrics from the most recent available PEPPER analytics report, from FY2023 Q2. To compare the PEPPER variables of interest to our method, we developed a simple list of ICD codes that are most predictive of appearing at the top of our list. To do so, we regressed our final ranking on the frequency of ICD codes used.

In terms of statistical methodology, there are a number of important ways in which PEPPER differs from our own analysis. First, PEPPER nearly exclusively focuses on DRGs, and in particular the ratio of some DRGs, for example, those with complications and comorbidities relative to those without. In contrast, our analysis drills down to the ICD code fundamentals that underpin DRG categorization. Second, PEPPER uses very simple statistics, such as percentages within certain categories or a hospital's rank among all national hospitals. This does not compare hospitals to their appropriate peers, and therefore is less valuable as it does not distinguish between hospitals that treat a sicker patient population versus those that just upcode.

Appendix K in the Supporting Information compares the PEPPER variables to the ICD codes that our method highlights as suspicious. Our list differs from the PEPPER target variables, though reassuringly, Sepsis/Septicemia is both the #1 predictor for our analysis as well as one of the predictors in PEPPER. Other similarities include a focus on cardiovascular, gastroenterological, and musculoskeletal conditions.

11 | Conclusion

The unsupervised ensemble method introduced in this work provides a new data-driven approach to identifying health care fraud using massive claims data. Our approach uses different data modalities—including patient medical history, provider coding patterns, and provider spending—to detect anomalous behavior consistent with fraud and abuse. Besides detection, the methodology offers interpretability, model-specific explanations pinpoint specific ICD and DRG codes associated with excess spending at a provider. Finally, our method allows us to characterize the types of providers most likely to be ranked as suspicious, which may be useful for guiding anti-fraud policy more broadly.

Our method substantially outperforms baseline algorithms. We combine evidence from multiple unsupervised outlier detection algorithms that use different types of global and local analysis to create a final ranking of suspiciousness, based on Medicare data. While only 1 in 12 hospitals is named in our DOJ ground truth data as fraudulent, 21 of our top ranked 50 hospitals are in the same corpus, achieving a nearly 5-fold improvement in detection rate.

Medicare spends over a hundred billion dollars per year on hospitalizations, and the federal government has limited enforcement capacity. We believe our findings are per se interesting, because they help pinpoint fraud by private firms against the government in a way that could be used to improve public spending.

Our method has natural extensions beyond Medicare and beyond hospitalizations. We believe that the same method will prove useful in detecting fraud against private insurers, who face many of the same issues. Private insurers spend hundreds of billions of dollars per year on reimbursing care, 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. Our method also has a natural extension to Medicaid, the federal-state partnered low-income subsidy program, which spends an additional $400 Billion per year on health care. With health care spending at 19.7% of U.S. GDP Centers for Medicare and Medicaid Services (2022), tools for detecting health care fraud can find wide-ranging use.

Acknowledgments

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 thank the NBER for data access and support and Lowell Taylor for his contributions to earlier stages of the project. We benefited greatly from comments from Anna Zink and participants of the 2023 ASHEcon conference.

Endnotes

[1] Patient refers to a person receiving health care; beneficiary refers to a person covered by health insurance. Here, they are used interchangeably, as all of our data come from patients who are Medicare beneficiaries.

[2] Twenty percent samples are the largest available for physician office visits.

References

AAMC. 2022. Council of Teaching Hospitals and Health Systems (COTH). https://www.aamc.rg/career-development/affinity-groups/coth.

Aggarwal, C. C., and S. Sathe. 2017. *Outlier Ensembles: An Introduction*. Springer.

Annual Report. 2022. 2022 Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds. https://www.cms.gov/files/document/2022-medicare-trustees-report.pdf.

Bar-Yossef, Z., T. S. Jayram, R. Kumar, and D. Sivakumar. 2004. "An Information Statistics Approach to Data Stream and Communication Complexity." *Journal of Computer and System Sciences* 68, no. 4: 702–732. (Preliminary Version in *43rd FOCS*, 2002).

Bauder, R. A., and T. M. Khoshgoftaar. 2017. "Medicare Fraud Detection Using Machine Learning Methods." In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 858–865. IEEE.

Bauder, R. A., and T. M. Khoshgoftaar. 2018. "The Detection of Medicare Fraud Using Machine Learning Methods With Excluded Provider Labels." In The Thirty-First International Flairs Conference.

Bauder, R., and T. Khoshgoftaar. 2018. "Medicare Fraud Detection Using Random Forest With Class Imbalanced Big Data." In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 80–87. IEEE.

Bauder, R., T. M. Khoshgoftaar, and N. Seliya. 2017. "A Survey on the State of Healthcare Upcoding Fraud Analysis and Detection." *Health Services and Outcomes Research Methodology* 17, no. 1: 31–55.

Becker, D., D. Kessler, and M. McClellan. 2005. "Detecting Medicare Abuse." *Journal of Health Economics* 24, no. 1: 189–210.

Bekker, J., and J. Davis. 2020. "Learning From Positive and Unlabeled Data: A Survey." *Machine Learning* 109, no. 4: 719–760.

Brot, Z., S. Burn, T. Layton, and B. Vabson. 2022. "Rationing Medicine Through Bureaucracy: Authorization Restrictions in Medicare." Working Paper.

Brunt, C. S. 2011. "CPT Fee Differentials and Visit Upcoding Under Medicare Part B." *Health Economics* 20, no. 7: 831–841.

Centers for Medicare & Medicaid Services. 2022. NHE Fact Sheet. U.S. Department of Health and Human Services.

Centers for Medicare and Medicaid Services. 2023. Program for Evaluating Payment Patterns Electronic Report. U.S. Department of Health and Human Services.

Chandola, V., S. R. Sukumar, and J. C. Schryver. 2013. "Knowledge Discovery From Massive Healthcare Claims Sata." In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1312–1320. Association for Computing Machinery.

Chikina, M., A. Frieze, and W. Pegden. 2017. "Assessing Significance in a Markov Chain Without Mixing." *Proceedings of the National Academy of Sciences of the United States of America* 114, no. 11: 2860–2864.

Dafny, L. S. 2005. "How Do Hospitals Respond to Price Changes?" *American Economic Review* 95, no. 5: 1525–1547.

Department of Justice. 2020. Internet Archive: The United States Department of Justice Health Care Fraud Unit.

Ekin, T., G. Lakomski, and R. M. Musal. 2019. "An Unsupervised Bayesian Hierarchical Method for Medical Fraud Assessment." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 12, no. 2: 116–124.

Eliason, P. J., R. J. League, J. Leder-Luis, R. C. McDevitt, and J. W. Roberts. 2025. "Ambulance Taxis: The Impact of Regulation and Litigation on Health Care Fraud." *Journal of Political Economy* 133: 5.

Fang, H., and Q. Gong. 2017. "Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked." *American Economic Review* 107, no. 2: 562–591.

Franceschini, F., D. A. Maisano, and L. Mastrogiacomo. 2022. "Ranking Aggregation Techniques." In *Rankings and Decisions in Engineering*, 85–160. Springer.

Guha, S., N. Mishra, G. Roy, and O. Schrijvers. 2016. "Robust Random Cut Forest Based Anomaly Detection on Streams." In *International Conference on Machine Learning*, 2712–2721. PMLR.

Herland, M., T. M. Khoshgoftaar, and R. A. Bauder. 2018. "Big Data Fraud Detection Using Multiple Medicare Data Sources." *Journal of Big Data* 5, no. 1: 1–21.

Howard, D. 2020. "False Claims Act Liability for Overtreatment." *Journal of Health Politics, Policy and Law* 45, no. 3: 419–437.

Joudaki, H., et al. 2015. "Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature." *Global Journal of Health Science* 7, no. 1: 194.

Kriegel, H.-P., P. Kröger, E. Schubert, and A. Zimek. 2009. "Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data." In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 831–838. Springer.

Kumaraswamy, N., M. K. Markey, T. Ekin, J. C. Barner, and K. Rascati. 2022. "Healthcare Fraud Data Mining Methods: A Look Back and Look Ahead." *Perspectives in Health Information Management* 19, no. 1.

Kuncheva, L. I., and C. J. Whitaker. 2003. "Measures of Diversity in Classifier Ensembles and Their Relationship With the Ensemble Accuracy." *Machine Learning* 51, no. 2: 181–207.

Leder-Luis, J. 2025. "Can Whistleblowers Root Out Public Expenditure Fraud? Evidence From Medicare." *Review of Economics and Statistics* 107, no. 5: 1169–1186.

Liu, F. T., K. M. Ting, and Z.-H. Zhou. 2008. "Isolation Forest." In *2008 Eighth IEEE International Conference on Data Mining*, 413–422. IEEE.

Lundberg, S. M., et al. 2020. "From Local Explanations to Global Understanding With Explainable AI for Trees." *Nature Machine Intelligence* 2, no. 1: 56–67.

Lundberg, S. M., and S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems*, 30: 4765–4774.

Luo, W., and M. Gallagher. 2010. "Unsupervised DRG Upcoding Detection in Healthcare Databases." In *2010 IEEE International Conference on Data Mining Workshops*, 600–605. IEEE.

MedPAC. 2023. Hospital Acute Inpatient Services Payment System.

Nam, G., J. Yoon, Y. Lee, and J. Lee. 2021. "Diversity Matters When Learning From Ensembles." *Advances in Neural Information Processing Systems*, 34, 8367–8377.

Noridian Healthcare Solutions. 2022. Unified Program Integrity Contractor (UPIC).

Pevnỳ, T. 2016. "Loda: Lightweight On-Line Detector of Anomalies." *Machine Learning* 102, no. 2: 275–304.

Rosenberg, M. A., D. G. Fryback, and D. A. Katz. 2000. "A Statistical Model to Detect DRG Upcoding." *Health Services and Outcomes Research Methodology* 1, no. 3: 233–252.

Sathe, S., and C. C. Aggarwal. 2016. "Subspace Outlier Detection in Linear Time With Randomized Hashing." In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 459–468. IEEE.

Shi, M. 2022. Monitoring for Waste: Evidence From Medicare Audits. https://mshi311.github.io/website2/Shi_MedicareAudits_2022_09_15.pdf.

Silverman, E., and J. Skinner. 2004. "Medicare Upcoding and Hospital Ownership." *Journal of Health Economics* 23, no. 2: 369–389.

Suresh, N. C., J. De Traversay, H. Gollamudi, A. K. Pathria, and M. K. Tyler. 2014. "Detection of Upcoding and Code Gaming Fraud and Abuse in Prospective Payment Healthcare Systems." US Patent 8,666,757.

U.S. Department of Health and Human Services. 2022. Annual Report of the Departments of Health and Human Services and Justice. https://oig.hhs.gov/publications/docs/hcfac/FY2021-hcfac.pdf.

U.S. Government Accountability Office. 2020. Payment Integrity Federal Agencies' Estimates of FY 2019 Improper Payments. https://www.gao.gov/assets/gao-20-344.pdf.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Appendix S1:Internet Appendix.