

Explainable Medicare Payment Integrity Analytics using Rank-Aggregated Anomaly Ensembles and Conformal Cost Reference Bands

Nassamon Bootwisas

Rajamangala University of Technology Suvarnabhumi

Pasin Marupanthorn

quantfilab@gmail.com

Maejo University

Research Article

Keywords: Medicare payment integrity, explainable anomaly detection, conformal cost bands, public healthcare data

Posted Date: June 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10053303/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Explainable Medicare Payment Integrity Analytics using Rank-Aggregated Anomaly Ensembles and Conformal Cost Reference Bands

Nassamon Bootwisas¹ and Pasin Marupanthorn^{2*}

¹Division of Computational Science, Faculty of Science and Technology,
Rajamangala University of Technology Suvarnabhumi, Phra Nakhon Si
Ayutthaya, 13000, Thailand.

^{2*}Department of Mathematics, Faculty of Science, Maejo University, 63
Moo 4, Nong Han, San Sai, Chiang Mai, 50290, Thailand.

*Corresponding author(s). E-mail(s): quantfilab@gmail.com;

Abstract

Public Medicare provider-payment releases create an important setting for reproducible payment-integrity analytics, but existing studies often separate fraud classification, anomaly screening, cost prediction, and uncertainty reporting. This paper develops an explainable public-data workflow that links provider-service feature construction, peer-relative anomaly ranking, expected-payment modeling, and conformal cost-reference bands. The anomaly layer aggregates isolation-based structure, empirical tail behavior, peer deviations, and specialty-service rarity into a label-free triage score. The cost layer uses tree-ensemble learning to estimate standardized Medicare payment and attaches calibrated reference bands so that unusual payments can be interpreted together with model uncertainty. The empirical evidence shows that the highest-priority review queue is concentrated in interpretable service environments rather than being driven only by raw payment magnitude. It also shows that the cost model performs best for routine payment behavior but becomes stressed in the extreme right tail, where residual inflation and interval violations provide additional escalation evidence. The framework therefore supports audit triage, governance, and reproducible public-data monitoring while maintaining the distinction between statistical anomaly and proof of misconduct.

Keywords: Medicare payment integrity, explainable anomaly detection, conformal cost bands, public healthcare data

1 Introduction

Medicare payment integrity is no longer only an administrative auditing problem; it is a statistical-learning and governance problem in which early detection, evidentiary caution, and reproducibility must be held together. The scale of the task is substantial. The Centers for Medicare and Medicaid Services (CMS) reported a Fiscal Year 2025 Medicare Fee-for-Service improper-payment estimate of 6.55%, or \$28.83 billion, while also emphasizing that improper payments are not synonymous with fraud because documentation, coding, and eligibility issues can generate payment error without proving intent [1]. CMS program-integrity activities explicitly use predictive analytics to identify and prevent fraud, waste, and abuse [2]. These facts create a practical need for analytic systems that can screen unusual provider-service behavior, report transparent reasons for escalation, and avoid converting statistical abnormality into a legal conclusion. Public Medicare provider-payment releases are therefore a valuable research setting because they contain observable use, payment, charge, specialty, service, and National Provider Identifier information without requiring restricted claims access [3–5].

The machine-learning literature on Medicare and healthcare-insurance fraud detection has advanced rapidly, but it remains fragmented across supervised classification, data-centric preprocessing, unsupervised anomaly detection, and interpretability. Hancock et al. [6] show that explainable supervised feature selection can support Medicare fraud detection in highly imbalanced data, and the same research program later demonstrates that data reduction is a central modeling decision for large imbalanced Medicare files rather than a minor preprocessing step [7]. Johnson and Khoshgoftaar [8] frame healthcare-fraud detection as a data-centric AI problem, where data representation and feature quality can be as important as model choice. Recent Medicare-focused and healthcare-insurance studies reinforce this view: standard classifiers remain sensitive to imbalance and benchmarking choices [9]; robust ensemble pipelines can improve predictive performance while using SHAP, partial dependence, and local explanation tools to support transparency [10]; and broad reviews identify persistent weaknesses in label quality, class imbalance, benchmark comparability, dataset diversity, and operational interpretability [11, 12].

A purely supervised fraud-classification design is not sufficient for the present problem. Confirmed fraud labels are sparse, delayed, institution-specific, and usually unavailable in public provider-service files. This has motivated one-class learning, unsupervised label generation, and explainable unsupervised screening for severely imbalanced fraud settings [13, 14]. De Meulemeester et al. [15] develop an explainable unsupervised workflow for healthcare-insurance resource-use anomalies, while Shekhar et al. [16] study unsupervised and explainable Medicare overbilling detection using hospitalization data and external enforcement evidence. Tajrobehkar et al. [17] further illustrate current interest in Medicare utilization analysis and fraud screening. These studies justify label-free triage, but most available designs still stop at suspiciousness ranking or supervised classification. They rarely combine public provider-service reproducibility, peer-relative service economics, expected-payment modeling, calibrated prediction bands, and tail-specific model-stress diagnostics in one audit-oriented workflow.

The methodological design of this paper follows from the structure of the data. Medicare provider-service tables are mixed-type, high-cardinality, skewed, and heavy tailed. For tabular prediction, tree-based ensembles remain strong benchmarks and are often difficult for deep architectures to dominate on ordinary structured data [18–20]. Category-aware boosting methods are useful comparators for heterogeneous categorical tables [21], but the absence of reliable public fraud labels makes a single supervised classifier inappropriate. The anomaly layer therefore combines complementary unsupervised views: isolation-forest structure, empirical distributional tail evidence, peer-relative payment deviations, and specialty-service rarity. Isolation forests provide an efficient isolation-based detector for unusual observations [22], while empirical-CDF outlier scoring, as in ECOD, supplies a parameter-light and interpretable tail anchor for heavy-tailed variables [23]. The cost-reference layer serves a different role: it estimates expected standardized payment and attaches split-conformal prediction intervals, using distribution-free calibration to report uncertainty around black-box regression estimates [24–26].

The gap addressed here is deliberately focused. Existing work often follows separate paths: supervised fraud detection with enforcement proxies, unsupervised anomaly detection without an expected-cost reference, payment prediction without anomaly screening, or private-data pipelines whose reproducibility is limited. This study develops an explainable public-data workflow at their intersection. It ranks unusual provider-service combinations, estimates expected Medicare standardized payment bands, relies on peer-relative and relational features rather than raw charge magnitudes alone, and treats anomaly as an operational warning signal rather than as proof of fraud.

The paper makes four contributions. First, it constructs a reproducible CMS provider-service analytics pipeline that preserves the boundary between public payment-integrity evidence and restricted claims-based adjudication. Second, it proposes a rank-aggregated anomaly ensemble that joins isolation-based structure, empirical tail evidence, peer-relative payment deviations, and specialty-service rarity into a label-free triage score. Third, it couples anomaly screening with a provider-held-out tree-ensemble cost model and split-conformal expected-payment reference bands, so each flagged row can be interpreted through both suspiciousness and calibrated cost uncertainty. Fourth, it reports tail-stress diagnostics showing that the highest anomaly tail is also the region where residual magnitude and conformal interval violations concentrate, thereby linking unsupervised screening to operationally meaningful cost-model failure rather than presenting a standalone anomaly score.

The remainder of the manuscript is organized as follows. The next section describes the public CMS data architecture, reproducibility boundary, and limits of supplementary commercial-data enrichment. The following section presents the feature engineering, rank-aggregated anomaly ensemble, validation-tuned cost ensemble, and split-conformal interval construction. The empirical section reports anomaly-tail structure, benchmark cost performance, interval calibration, and the operational comparison between the proposed framework and standalone point predictors. The conclusion summarizes the contribution and outlines extensions needed for longitudinal and externally validated Medicare payment-integrity research.

2 Data Architecture and Reproducibility Boundary

2.1 Public-data unit and source hierarchy

The empirical object of this study is a public provider-service payment record rather than a beneficiary-level claim or an adjudicated fraud case. The central file is the 2023 CMS Medicare Physician and Other Practitioners by Provider and Service release, which reports utilization, submitted charges, allowed amounts, Medicare payments, standardized payments, HCPCS code, place of service, provider specialty, state, and National Provider Identifier (NPI) [4]. Each observation is therefore indexed as

$$i = (n_i, h_i, p_i), \quad (1)$$

where n_i denotes the NPI, h_i the HCPCS procedure or service code, and p_i the place-of-service indicator. This unit is appropriate for payment-integrity triage because it connects a provider, a billable service, and a payment environment without requiring restricted claim-line access.

Two provider-level aggregate files are used to add stable context to the detailed rows. The 2023 Medicare Physician and Other Practitioners by Provider file contributes provider totals, service breadth, beneficiary counts, average beneficiary age, dual-eligibility composition, and average beneficiary risk score [3]. The corresponding 2022 provider file contributes lagged provider totals for year-over-year context. These lagged variables are used only as provider-level background measures; the design does not claim a full longitudinal provider-service panel because the detailed NPI–HCPCS–place structure is observed only for 2023 in the main estimation set. The response variable for the cost-reference layer is the average Medicare standardized payment amount, denoted c_i , while the anomaly layer remains label-free because the public files contain no official fraud or improper-payment target.

Table 1 summarizes the source hierarchy. CMS SynPUF documentation is retained as methodological context for public Medicare data structure, but it is not used for calibration because the present object is provider-payment screening rather than beneficiary-level synthetic claims modeling [5]. The CMS Open Payments API is documented as a possible enrichment path, but it is not merged in the reported tables because manufacturer-to-provider entity resolution is less direct than the NPI-based CMS joins [27]. A local Refinitiv Workspace environment is also available for future validation around listed healthcare parent organizations; however, it is outside the main design because most NPI-level provider-service rows do not map cleanly to tradable parent entities. The reported results therefore remain fully grounded in reproducible public CMS payment files.

2.2 Extraction, joins, and reproducible sample construction

The pipeline follows a fixed sequence so that the empirical run can be audited. Raw CMS files are downloaded and retained before any transformation. The detailed provider-service file is read in chunks because the raw release is large, and the estimation set is then drawn as a reproducible 500,000-row sample using a fixed random seed. This sampling step is computational rather than inferential: it is used to make

Table 1 Public data sources, granularity, and analytical role.

Source	Granularity	Main fields used	Role in the empirical design
CMS Physician and Other Practitioners by Provider and Service, 2023	NPI-HCPCS-place of service	Utilization, submitted charge, allowed amount, Medicare payment, standardized payment, specialty, state, HCPCS, place of service	Main row-level backbone for anomaly ranking and cost-reference modeling
CMS Physician and Other Practitioners by Provider, 2023	NPI	Provider totals, beneficiary mix, average age, dual eligibility, risk score, service breadth	Provider context merged to detailed rows by NPI
CMS Physician and Other Practitioners by Provider, 2022	NPI	Prior-year provider totals for services, submitted charges, and Medicare payments	Lagged provider context for growth variables; not a detailed panel
CMS DE-SynPUF documentation	Synthetic beneficiary-claim documentation	Public Medicare metadata and file-structure context	Documentation only; excluded from model calibration
CMS Open Payments API	Provider-manufacturer records	Possible payment-to-provider enrichment fields	Future extension; excluded from reported empirical tables

repeated model fitting, grouped resampling, conformal calibration, and figure generation feasible on an ordinary research workstation. National expenditure totals or provider-population prevalence claims are not estimated from the sample.

After sampling, provider aggregate variables are merged by NPI. Aggregate fields are renamed before merging so that row-level service quantities are not confused with provider-level totals. The merge produces three layers of variables: detailed provider-service fields from 2023, contemporaneous provider context from 2023, and lagged provider context from 2022. When a provider has no matched lagged aggregate record, the missing value is interpreted as missing public-data context rather than as a zero economic quantity. The modeling layer therefore uses median or mode imputation after the split logic is defined, while ratio features use small numerical offsets in denominators to avoid unstable divisions around zero.

The reproducibility boundary is also a leakage boundary. Provider-held-out evaluation is used later for the cost model, so NPI membership is the grouping variable for train, calibration, and holdout partitions. Peer-cell and network features are computed from the sampled public table before model evaluation because they describe the public service environment in which each row appears. Consequently, the reported cost evaluation is provider-held-out but partially transductive with respect to public service frequencies. This is appropriate for an offline audit prototype, where the entire public release is available to construct peer norms, but it should be tightened in prospective deployment by estimating peer summaries from historical data only.

2.3 Sample profile and empirical scope

Table 2 reports the resulting estimation set. The sample contains 500,000 provider-service rows, 346,038 unique providers, 3,901 HCPCS codes, 103 provider specialties, and 60 state or territorial units. The mean standardized Medicare payment is \$82.38, while the mean submitted charge is \$417.79. This gap is analytically important: submitted charges, allowed amounts, Medicare payments, and standardized payments

Table 2 Sample summary for the 2023 provider-service estimation set.

Statistic	Value
Provider-service rows	500,000
Unique providers	346,038
Unique HCPCS codes	3,901
Provider specialties	103
States and territories	60
Mean standardized payment (\$)	82.38
Mean submitted charge (\$)	417.79

measure different institutional quantities. A high submitted charge is therefore not automatically anomalous once service code, provider specialty, and place of service are taken into account. The design accordingly emphasizes peer-relative and service-context features rather than raw charge magnitude alone.

The public CMS architecture supports two linked empirical tasks. First, it supports unsupervised anomaly ranking because each row contains enough information to compare payment magnitude, utilization intensity, provider-service concentration, specialty-service rarity, and peer-cell deviation. Second, it supports expected-payment modeling because standardized payment can be predicted conditional on observable service, provider, and peer context. The same architecture also imposes clear limits. The files do not contain patient-level sequences, medical necessity documentation, diagnosis histories, full claim adjudication narratives, or enforcement outcomes. Thus, the model can prioritize unusual provider-service rows for review, but it cannot determine intent, clinical appropriateness, or legal culpability.

These limits shape the interpretation throughout the paper. A row with a high anomaly score is a statistically unusual payment-integrity signal, not an allegation of fraud. A row outside a conformal expected-cost band is a payment observation that the model did not cover under its calibrated reference distribution, not proof of misconduct. The value of the dataset lies in the combination of public reproducibility and operational caution: the same public files can be reprocessed by independent researchers, while the absence of restricted adjudication labels prevents the analysis from overstating what public payment summaries can prove.

3 Feature Engineering and Multi-view Ensemble Design

3.1 Notation and payment-integrity feature map

The methodological layer is built to separate four forms of unusualness that are often confounded in raw Medicare payment tables: large payment magnitude, intensive utilization, deviation from clinically comparable peers, and rare provider-service relationships. For row i , let n_i denote the NPI, h_i the HCPCS code, s_i the provider specialty, p_i the place-of-service indicator, and c_i the average Medicare standardized payment amount used as the response in the cost-reference layer. Let m_i , a_i , and q_i

denote the observed average Medicare payment, average allowed amount, and average submitted charge, and let T_i , B_i , and D_i denote service count, beneficiary count, and beneficiary-day count when available. Define $M_i = m_i T_i$ as the row-level Medicare payment mass used for concentration features. All ratios use a small ridge $\varepsilon = 10^{-6}$ to avoid undefined values in rare near-zero payment rows:

$$R_i^{q/m} = \frac{q_i + \varepsilon}{m_i + \varepsilon}, \quad R_i^{a/q} = \frac{a_i + \varepsilon}{q_i + \varepsilon}, \quad R_i^{m/a} = \frac{m_i + \varepsilon}{a_i + \varepsilon}. \quad (2)$$

Utilization intensity is represented by services per beneficiary, beneficiary days per beneficiary, and logarithmic payment-scale transformations,

$$U_i^{T/B} = \frac{T_i}{\max\{B_i, 1\}}, \quad U_i^{D/B} = \frac{D_i}{\max\{B_i, 1\}}, \quad L_i(v) = \log(1 + v_i), \quad (3)$$

where v_i is a nonnegative payment, charge, service, or beneficiary quantity. Provider concentration is measured within the detailed 2023 provider-service file. If $\mathcal{I}(n_i) = \{j : n_j = n_i\}$ is the set of sampled rows attached to provider n_i , the service and payment shares are

$$H_i^T = \frac{T_i}{\sum_{j \in \mathcal{I}(n_i)} T_j + \varepsilon}, \quad H_i^M = \frac{M_i}{\sum_{j \in \mathcal{I}(n_i)} M_j + \varepsilon}. \quad (4)$$

These features prevent the score from treating every high-volume provider equally: a row is more informative when one service accounts for a large fraction of a provider's public service or payment profile. The provider-level 2023 aggregate file adds service breadth, total beneficiaries, total services, mean beneficiary age, dual-eligibility share, and average beneficiary risk score. The corresponding 2022 provider aggregate file is used only to form limited year-over-year background variables,

$$\Delta_i(v) = \log(1 + v_{n_i, 2023}) - \log(1 + v_{n_i, 2022}), \quad (5)$$

for services, submitted charges, and Medicare payments when both years are observed. Equation 5 gives temporal context without converting the design into a full longitudinal provider-service panel.

3.2 Peer-relative and relational features

Raw payment amounts are not directly comparable across specialties, procedures, and service environments. The peer layer therefore defines a comparison cell $g(i) = (s_i, h_i, p_i)$ and computes robust cell-relative deviations. For any scalar variable x , the peer score is

$$z_i(x) = 0.6745 \frac{x_i - \text{median}_{j \in g(i)}(x_j)}{\text{MAD}_{j \in g(i)}(x_j) + \varepsilon}. \quad (6)$$

The factor 0.6745 puts the median absolute deviation on the usual normal scale under ideal Gaussian conditions, but the scores are not interpreted as calibrated Gaussian residuals. In this application they are robust ranking features. The implementation uses Equation 6 for average Medicare payment, submitted charge, and the charge-to-payment ratio. Sparse cells and zero-MAD cells are retained with the ridge term rather

than discarded, because rare service settings are themselves part of the audit signal; however, the resulting peer features are evaluated jointly with other views rather than used as deterministic rules.

Relational structure is captured by a provider–procedure bipartite representation. Provider degree and service degree are

$$d_i^{NPI} = |\{h_j : j \in \mathcal{I}(n_i)\}|, \quad d_i^{HCPCS} = |\{n_j : h_j = h_i\}|. \quad (7)$$

The first quantity distinguishes narrow, concentrated public service profiles from broad provider profiles; the second distinguishes common HCPCS codes from procedures or service lines used by relatively few providers in the sample. The specialty–service rarity feature is a smoothed negative log conditional frequency,

$$r_i = -\log \left(\frac{N(s_i, h_i) + 1}{N(s_i) + 1} \right), \quad (8)$$

where $N(s_i, h_i)$ is the number of rows with specialty s_i and HCPCS code h_i , and $N(s_i)$ is the number of rows with specialty s_i . High r_i indicates that the service is unusual for the provider’s specialty in the public sample. Equations 6–8 are deliberately simple: they produce auditable quantities that can be inspected by a reviewer before any machine-learning score is considered.

3.3 Four anomaly views

Let $\mathbf{x}_i^A$ denote the anomaly-feature vector formed from the log payment and charge variables, ratios in Equation 2, utilization intensities in Equation 3, provider shares in Equation 4, provider aggregate variables, growth variables, peer scores, network degrees, and rarity. Continuous features are winsorized at the 1st and 99th percentiles and robustly scaled before model fitting. The anomaly score is not learned from fraud labels; instead, it is constructed from four complementary unsupervised views.

The first view is the isolation-forest score S_i^{IF} , fitted to $\mathbf{x}_i^A$ after robust scaling. Isolation forest ranks observations by how easily randomized partitions isolate them, making it suitable for mixed heavy-tailed anomaly screening without a parametric density model [22]. The second view is an empirical-CDF tail score S_i^{ECDF} . For feature k , let $\hat{F}_k$ be the empirical distribution function and define

$$E_{ik} = \max \left\{ -\log(\hat{F}_k(x_{ik}) + \varepsilon), -\log(1 - \hat{F}_k(x_{ik}) + \varepsilon) \right\}. \quad (9)$$

The ECDF view is then $S_i^{ECDF} = \sum_k E_{ik}$, following the logic of ECOD that extreme empirical ranks provide a parameter-light tail signal in high-dimensional tabular data [23]. The third view is the peer-deviation score

$$S_i^{Peer} = \frac{1}{3} \left(|z_i(m)| + |z_i(q)| + |z_i(R^{q/m})| \right), \quad (10)$$

which captures unusual behavior relative to specialty–HCPCS–place cells. The fourth view is a rarity-concentration score S_i^{Rare} , defined as the average of percentile-normalized specialty–service rarity, provider service share, provider payment-mass share, and charge-to-payment structure. This view is included because some audit-relevant rows are not globally extreme but are unusual because a rare service or concentrated billing pattern occurs inside a particular specialty profile.

Each raw view S_i^v , $v \in \{IF, ECDF, Peer, Rare\}$, is winsorized and converted to a percentile rank $\rho_i^v \in [0, 1]$, with larger values indicating more anomalous observations and average ranks used for ties. The final label-free anomaly score is the Borda-style aggregation

$$A_i = \frac{1}{4} (\rho_i^{IF} + \rho_i^{ECDF} + \rho_i^{Peer} + \rho_i^{Rare}). \quad (11)$$

The top anomaly set at tail probability τ is

$$\mathcal{T}_\tau = \{i : A_i \geq \hat{Q}_\tau(A)\}, \quad (12)$$

where $\hat{Q}_\tau(A)$ is the empirical τ -quantile of the aggregated score. The empirical section uses $\tau = 0.99$. Rank aggregation is used instead of optimized weights because the public file provides no adjudicated fraud outcome. A supervised weighting scheme would therefore optimize toward an arbitrary proxy and would overstate validation. Equation 11 keeps the ranking reproducible, scale-robust, and explainable through its four component ranks.

3.4 Expected-cost model and grouped ensemble design

The cost-reference layer answers a different question from the anomaly layer. Rather than asking whether a provider-service row is unusual, it estimates the standardized Medicare payment that would be expected given the observed service, provider, peer, and aggregate context. The target is modeled on the log scale,

$$y_i = \log(1 + c_i), \quad (13)$$

with predictions transformed back to the payment scale before reporting dollar errors. The feature vector for this layer, $\mathbf{x}_i^C$, includes categorical fields such as provider type, state, place of service, HCPCS code, and drug indicator, together with the numeric payment ratios, utilization variables, provider aggregates, growth measures, peer scores, network degrees, and rarity features. Categorical variables are ordinally encoded, numeric missing values are median-imputed within the training workflow, and all evaluation splits are grouped by NPI so that rows from the same provider do not appear simultaneously in training and holdout sets.

Two tree-based regressors are used because structured Medicare payment data are mixed, sparse, nonlinear, and highly skewed. Random forest reduces variance through bootstrap aggregation of decorrelated trees [18], while extra trees adds stronger randomization in the split-generation step [19]. Tree ensembles are also strong practical benchmarks for heterogeneous tabular data [20]. If $\hat{\mu}_i^{RF}$ and $\hat{\mu}_i^{ET}$ are the log-scale

predictions, the payment-scale predictions are

$$\hat{c}_i^{RF} = \exp(\hat{\mu}_i^{RF}) - 1, \quad \hat{c}_i^{ET} = \exp(\hat{\mu}_i^{ET}) - 1. \quad (14)$$

A grouped calibration split inside the training data selects a transparent convex blend,

$$\hat{c}_i = w\hat{c}_i^{RF} + (1 - w)\hat{c}_i^{ET}, \quad w \in \{0, 0.05, \dots, 1\}, \quad (15)$$

where w minimizes calibration mean absolute error. This outer blend is not boosting: no sequential residual fitting is performed, and no pseudo-label is created for fraud. Bagging-style aggregation occurs only inside the tree ensembles. The final blend is a validation-tuned cost reference whose main purpose is operational interpretability rather than universal point-prediction dominance.

The split design is conservative but not deployment-pure. Holding out providers prevents direct NPI memorization in the cost model. At the same time, peer-relative and network variables are computed on the public analytical sample before the split, so the evaluation is partially transductive with respect to the shared service environment. This is acceptable for a retrospective public-data benchmark, but a prospective production deployment should recompute peer and network summaries using only information available before the scoring date.

3.5 Conformal cost bands and diagnostic metrics

Point predictions alone are insufficient for audit triage because a large residual can reflect model uncertainty, rare case mix, coding heterogeneity, or truly unusual payment behavior. The framework therefore adds split-conformal payment bands around Equation 15. Let $\mathcal{C}_{cal}$ be the grouped calibration set and let

$$e_j = |c_j - \hat{c}_j|, \quad j \in \mathcal{C}_{cal}, \quad (16)$$

be the absolute calibration residuals on the payment scale. For nominal miscoverage α , define

$$q_{1-\alpha} = e_{(k)}, \quad k = \lceil (|\mathcal{C}_{cal}| + 1)(1 - \alpha) \rceil, \quad (17)$$

where $e_{(k)}$ is the k th order statistic. The conformal cost-reference band is

$$\mathcal{C}_{i,\alpha} = [\max\{0, \hat{c}_i - q_{1-\alpha}\}, \hat{c}_i + q_{1-\alpha}]. \quad (18)$$

For $\alpha = 0.10$, Equation 18 targets 90% marginal coverage under the usual exchangeability condition of split conformal prediction, without assuming Gaussian errors or homoskedastic residuals [24]. The guarantee is marginal rather than conditional on specialty, HCPCS code, or anomaly status; consequently, subgroup undercoverage is treated as an empirical diagnostic rather than as a contradiction of conformal validity.

Model performance is summarized by mean absolute error, root mean squared error, R^2 , median absolute percentage error, empirical interval coverage, interval

Table 3 Compact interpretation map for the empirical evidence.

Result block	Statistical and operational reading
Anomaly-score tail	The top 1% threshold is 0.9066 and selects 5,000 rows. This defines a reproducible review queue, not a fraud probability.
Peer and ratio diagnostics	The median charge-to-payment ratio rises from 3.729 to 14.058, while the median absolute peer-payment score rises from 0.653 to 1.408. Tail rows are therefore both ratio-extreme and peer-unusual.
Specialty concentration	Clinical laboratory has an 11.16% tail share, with ambulatory surgical center, oncology, pain-management, and advanced-practice categories also elevated. This is an audit-allocation signal, not specialty-level culpability.
Cost reference	Full-holdout MAE is \$12.25 and below-95th-percentile MAE is \$4.66. The cost layer is reliable for routine payments but compresses the extreme right tail.
Conformal stress	Overall 90% coverage is 90.07%, but anomaly-tail coverage is 79.92%. The most useful escalation signal is the combination of high anomaly score and cost-band violation.

width, out-of-band rate, and tail-specific residual inflation. For a tail set $\mathcal{T}_{0.99}$, the residual inflation factor is

$$\Gamma = \frac{|\mathcal{T}_{0.99}|^{-1} \sum_{i \in \mathcal{T}_{0.99}} |c_i - \hat{c}_i|}{|\mathcal{T}_{0.99}^c|^{-1} \sum_{i \in \mathcal{T}_{0.99}^c} |c_i - \hat{c}_i|}, \quad (19)$$

and the out-of-band ratio is defined analogously using the indicator $\mathbf{1}\{c_i \notin \mathcal{C}_{i,0.10}\}$. These quantities connect the two layers of the paper. A row is operationally important when it is unusual across the rank-aggregated anomaly views and also lies in a region where the expected-cost model is surprised. The empirical design therefore reports not only predictive accuracy, but also whether the anomaly tail concentrates residual stress and conformal interval violations across repeated provider-held-out resamples.

4 Empirical Results and Operational Interpretation

4.1 Reading the evidence as an audit workflow

The empirical results should be read as an integrated audit-triage experiment rather than as a conventional fraud-classification exercise. The public Medicare file does not reveal adjudicated fraud labels, so the analysis asks whether the proposed workflow can produce a reproducible and interpretable review queue. Four empirical questions organize the section. First, does the rank-aggregated anomaly score isolate a narrow tail with stronger peer-relative and ratio-based irregularity than the full sample? Second, is this tail concentrated in interpretable service environments rather than scattered randomly across the provider-service universe? Third, can the cost-reference model produce a useful expected-payment benchmark under provider-held-out evaluation? Fourth, do conformal intervals reveal where the cost model is surprised, especially for anomaly-tail rows? Figure 1 gives the graphical summary, while Tables 4–9 provide the numerical audit record.

Table 4 Anomaly-tail summary for the 2023 provider-service sample. Peer-based values are robust cell-relative scores rather than calibrated standard-normal quantities.

Statistic	Value
Top 1% anomaly threshold	0.9066
Top 1% anomaly rows	5,000
Median charge-to-payment ratio, full sample	3.729
Median charge-to-payment ratio, top 1% tail	14.058
Median absolute peer-payment score, full sample	0.653
Median absolute peer-payment score, top 1% tail	1.408

4.2 Anomaly-tail selection and peer-relative meaning

Panel A of Figure 1 shows the empirical distribution of the final rank-aggregated anomaly score. The 99th-percentile cutoff is 0.9066, which selects exactly 5,000 rows from the 500,000-row estimation sample. This threshold is useful because it converts a continuous unsupervised score into a concrete audit budget: the tail can be reviewed, stratified, or sampled without relying on an arbitrary manual cutoff. The threshold should not be interpreted as a probability of fraud. It is a percentile-based boundary defining the most unusual provider-service rows according to the combined feature views.

Table 4 shows that the selected tail is not merely a high-score artifact. The median charge-to-payment ratio increases from 3.729 in the full sample to 14.058 in the top 1% anomaly tail, a multiplicative increase of about 3.77. The median absolute peer-payment score rises from 0.653 to 1.408, approximately 2.16 times the full-sample median. These two shifts have different meanings. The charge-to-payment ratio captures the gap between submitted charges and Medicare payments, while the peer-payment score asks whether the row is unusual after conditioning on specialty, HCPCS code, and place of service. Their joint increase indicates that the anomaly tail contains rows that are both ratio-extreme and peer-unusual. A purely dollar-based screen would not provide this distinction.

The rank-aggregation design matters for interpretation. Each observation receives evidence from several views: multivariate isolation, empirical tail rarity, peer-relative deviation, and specialty-service rarity. A row enters the final top tail when it ranks persistently high across these views, not because one hand-tuned weight assigns it a large score. This is important in payment-integrity settings because extreme raw charges, sparse cells, rare HCPCS codes, and high utilization can each generate false emphasis when used alone. Rank aggregation dampens dependence on any single scale and gives reviewers a more stable list of candidate rows. Nevertheless, the score remains a statistical prioritization device. Sparse peer cells, zero-MAD cells, and near-zero payment denominators can still make ratio and peer features unstable, so the correct use is triage with contextual review rather than automatic accusation.

Table 5 Selected specialties with the highest anomaly-tail concentration.

Specialty	Rows	Top 1% anomaly share
Clinical Laboratory	5,716	0.1116
Ambulatory Surgical Center	3,094	0.0385
Medical Oncology	2,482	0.0379
Hematology-Oncology	7,646	0.0327
Pain Management	2,284	0.0201
Physician Assistant	28,157	0.0181
Nurse Practitioner	45,881	0.0161
Interventional Pain Management	1,381	0.0159

4.3 Specialty concentration and row-level audit interpretation

Panel B of Figure 1 and Table 5 show that the anomaly tail is concentrated in recognizable service environments. Clinical laboratory rows have the highest top-tail share, 11.16%, which is more than eleven times the 1% sample-wide tail rate. Ambulatory surgical center and oncology-related categories also show elevated shares: ambulatory surgical center rows have a 3.85% tail share, medical oncology 3.79%, and hematology-oncology 3.27%. Pain-management and advanced-practice categories are lower in absolute share but still above the population tail rate. These concentrations are substantively plausible because the corresponding environments often contain heterogeneous service intensity, specialized procedure mixes, and payment patterns that differ sharply across provider-service cells.

The specialty pattern must be interpreted carefully. A large specialty-level top-tail share does not imply specialty-level misconduct. It means that, within the public file and the chosen feature map, unusual combinations of payment ratios, utilization intensity, provider concentration, peer-cell deviation, and service rarity are more common in those service environments. For audit planning, this distinction is valuable. It suggests where review teams should demand richer clinical and administrative context, but it does not justify a conclusion about individual providers, organizations, or specialties without external documentation.

The exported row-level anomaly table is consistent with the same reading. The highest-ranked rows include vaccine-administration lines, laboratory diagnostics, radiology-guided procedures, infusion-related services, and office or outpatient management codes. The common element is not a single clinical family. It is the interaction between service code, provider type, provider concentration, service intensity, and peer-relative payment behavior. This is precisely the type of interaction that a one-dimensional rule, such as submitted charge above a fixed threshold or payment above a fixed percentile, would miss. The anomaly layer therefore functions as an explainable routing mechanism: it indicates why a row is statistically unusual and where a reviewer should begin asking questions.

Table 6 Repeated grouped-split benchmark comparison for average standardized payment. Results are mean $\pm$ standard deviation across three provider-held-out resamples.

Model	MAE	RMSE	R^2	MedAPE
Global median baseline	59.65 $\pm$ 0.21	249.13 $\pm$ 10.43	-0.013 $\pm$ 0.001	0.5813 $\pm$ 0.0007
HCPCS median baseline	19.26 $\pm$ 0.16	206.23 $\pm$ 13.28	0.306 $\pm$ 0.031	0.0473 $\pm$ 0.0001
Random forest	13.67 $\pm$ 0.04	125.58 $\pm$ 3.29	0.742 $\pm$ 0.021	0.0354 $\pm$ 0.0009
Extra trees	12.44 $\pm$ 0.28	135.31 $\pm$ 12.75	0.699 $\pm$ 0.057	0.0299 $\pm$ 0.0008
Validation-tuned ensemble	12.42 $\pm$ 0.26	133.42 $\pm$ 11.74	0.707 $\pm$ 0.054	0.0306 $\pm$ 0.0007

4.4 Expected-payment accuracy and benchmark comparison

The cost-reference layer evaluates whether a provider-service payment looks consistent with observable public-file covariates. Table 6 reports repeated provider-held-out results across three grouped resamples. The global median baseline has a mean absolute error of \$59.65, showing that an unconditional center is a poor description of standardized Medicare payment. The HCPCS median baseline is much stronger, with MAE \$19.26, because procedure code captures a large share of payment variation. This is the relevant simple benchmark: any useful learning model should improve on procedure-level payment medians.

All learned tree models improve on the HCPCS median baseline. The validation-tuned ensemble records the lowest repeated-split MAE, \$12.42 $\pm$ \$0.26, corresponding to an approximate 35.5% reduction in dollar error relative to the HCPCS median baseline and an approximate 79.2% reduction relative to the global median baseline. The margin over extra trees is very small, however: extra trees obtains \$12.44 $\pm$ \$0.28. Random forest has the strongest repeated-split R^2 , 0.742 $\pm$ 0.021, while extra trees has the lowest median absolute percentage error. The result should therefore be stated conservatively. The proposed cost layer is competitive and narrowly best on the main dollar-error metric, but it does not dominate every metric. This qualification strengthens the empirical claim because the contribution of the paper is the integrated anomaly-cost-uncertainty workflow, not universal point-prediction superiority.

Panel D of Figure 1 gives the same message visually. The learned models form a tight group far below the two median baselines, while the proposed ensemble and extra trees are nearly indistinguishable on MAE. This is operationally useful because MAE is expressed in the same dollar units as the target and is less dominated by extreme cases than RMSE. The RMSE and R^2 columns still matter: they show that the right tail is difficult and that models differ in how they trade routine accuracy against large-error behavior. For audit screening, the main objective is not to remove every large error. It is to identify when large errors occur in rows that are also statistically unusual.

4.5 Segment analysis and right-tail compression

Table 7 explains where the cost layer works well and where it fails. On the representative holdout split, the validation-tuned ensemble obtains full-holdout MAE \$12.25 and $R^2 = 0.722$. For the routine mass below the 95th percentile of actual payment, MAE falls to \$4.66 and R^2 rises to 0.932. The mean predicted-to-actual ratio in this

Table 7 Representative holdout cost accuracy by payment segment for the validation-tuned ensemble.

Segment	Rows	MAE	R^2	Mean pred./actual
Full holdout	99,773	12.25	0.722	0.905
Actual below 95th percentile	94,784	4.66	0.932	0.989
Actual top 5%	4,989	156.35	0.644	0.738
Actual top 1%	998	535.94	0.554	0.630
Anomaly top 1% tail	1,056	46.13	0.680	0.668

segment is 0.989, meaning that the model is nearly unbiased on average for routine standardized payments. This is the strongest positive result for the cost layer: it can serve as a stable expected-payment reference for the large majority of provider-service rows.

The same table also reveals the main weakness. In the actual top 5% of payments, MAE increases to \$156.35 and the mean predicted-to-actual ratio falls to 0.738. In the actual top 1%, MAE rises further to \$535.94 and the mean predicted-to-actual ratio falls to 0.630. Panel C of Figure 1 illustrates this right-tail compression: the highlighted highest-payment observations tend to lie above the diagonal, meaning that actual standardized payment exceeds the model prediction. This is not surprising in a highly skewed public payment file. Tree ensembles trained for average accuracy borrow strength from dense routine regions and tend to smooth sparse extreme outcomes toward less extreme conditional averages. The important point is that the paper does not hide this limitation. It uses the limitation as part of the audit diagnostic.

The anomaly top 1% segment in Table 7 should be read separately from the actual top-payment segment. Its MAE is \$46.13, much larger than the below-95th-percentile MAE but far below the MAE in the actual top 1% payment segment. This confirms that the anomaly score is not just a proxy for being one of the largest-payment rows. It captures unusualness across payment ratios, peer deviations, utilization intensity, provider concentration, and rarity. The mean predicted-to-actual ratio of 0.668 indicates that cost underprediction is still present in the anomaly tail. Thus, the anomaly layer and the cost layer are complementary: the anomaly layer identifies non-routine structure, while the cost layer quantifies how far the observed payment departs from expected payment behavior.

4.6 Conformal cost bands as calibrated stress evidence

Table 8 evaluates the uncertainty layer. The split-conformal band is designed for 90% marginal coverage, and the representative holdout coverage is 90.07%. This indicates that the conformal calibration step is working at the aggregate level. The mean interval width is \$32.52, which is informative for routine payments and intentionally modest relative to the extreme right tail. The interval is therefore not a substitute for tail-specialized modeling. It is a calibrated reference band for asking whether a row is unusually far from the expected-payment surface.

The subgroup diagnostics provide the most important operational signal. The top anomaly tail has 79.92% coverage, while all remaining rows have 90.17% coverage.

Table 8 Split-conformal interval diagnostics for the 90% cost band.

Metric	Value
Nominal coverage	0.90
Observed holdout coverage	0.9007
Mean interval width (\$)	32.52
Top 1% tail coverage	0.7992
Remaining-row coverage	0.9017
Top 1% out-of-band rate	0.2008
Remaining-row out-of-band rate	0.0983
Top-to-remaining out-of-band ratio	2.0431

Equivalently, 20.08% of top-anomaly rows fall outside the band, compared with 9.83% of non-tail rows. The top-to-remaining out-of-band ratio is 2.0431, so top-anomaly rows are about twice as likely to violate the cost band. This is not a contradiction of conformal prediction, because split conformal prediction guarantees marginal coverage, not equal conditional coverage in every subgroup. Instead, it is exactly the kind of subgroup stress diagnostic that is useful for payment-integrity review. A row that is both anomaly-tail and outside the conformal band deserves more attention than a row with only one of these properties.

Repeated grouped resampling supports the robustness of this relationship. Across resamples, the top anomaly tail has a residual inflation factor of 6.33 ± 1.10 relative to non-tail rows and an out-of-band ratio of 2.09 ± 0.13 . These repeated-split results reduce the chance that the anomaly-cost link is produced by a single lucky provider split. They also clarify the empirical contribution: the anomaly score is not validated as fraud, but it is validated as a way to locate observations where the cost-reference model is materially more likely to be stressed.

4.7 Integrated comparison with standalone predictors

Table 9 compares the full workflow with the benchmark predictors. The global median, HCPCS median, random forest, and extra trees models are useful cost-prediction baselines, but in this design they output only point predictions. They do not define an anomaly ranking, do not attach peer-relative explanatory context, do not produce a calibrated cost band, and do not quantify whether the anomaly tail has a higher interval-violation rate. The proposed framework offers all four outputs in a single reproducible pipeline.

This comparison gives the proper interpretation of superiority. The proposed framework is not superior because every point-prediction column favors it. It is superior for the payment-integrity task because it transforms public data into a richer decision object: a provider-service row receives an anomaly score, a peer context, an expected payment, a calibrated interval, and a flag indicating whether the observed payment falls outside that interval. For interpretability, the exported permutation-importance analysis ranks provider payment share, provider service share, HCPCS code, and

Table 9 Operational outputs produced by the proposed framework versus standalone benchmark predictors. The contribution rests on integrated outputs, not on universal point-prediction dominance.

Method	Repeated MAE	Anomaly ranking	Calibrated cost band	Additional operational output
Global median base-line	59.65±0.21	No	No	Point forecast only
HPCPS median base-line	19.26±0.16	No	No	Point forecast only
Random forest benchmark	13.67±0.04	No	No	Point forecast only; strongest repeated-split R^2 among point predictors
Extra trees benchmark	12.44±0.28	No	No	Point forecast only; lowest repeated-split MedAPE among point predictors
Proposed framework	12.42±0.26	Yes	Yes	Rank-aggregated top 1% anomaly threshold = 0.9066; 90.1% holdout coverage; top-tail out-of-band rate = 20.1% versus 9.8% for remaining rows; repeated residual inflation = 6.33 ± 1.10

service degree among the dominant cost predictors. This is consistent with the institutional structure of Medicare payment because expected payment is shaped by service economics and provider concentration, not by raw charges alone.

4.8 Synthesis, limitations, and deployment reading

The combined evidence supports three substantive conclusions. First, public CMS provider-service files are rich enough to sustain an explainable anomaly-screening prototype even without restricted beneficiary-level claims. Second, peer-relative and rarity-oriented features are necessary because they separate unusual provider-service behavior from ordinary high-volume or high-charge settings. Third, uncertainty information is not optional. The cost model is accurate for routine payments, but its interval failures and residual inflation identify precisely where the expected-payment surface is least reliable.

The appropriate deployment reading is therefore conservative and governance-oriented. A high anomaly score should trigger review, not punishment. A large predicted-payment residual should trigger explanation, not a finding. A conformal-band violation should be interpreted as evidence that the row is outside a calibrated public-data reference band, not as proof of improper billing. The strongest escalation signal occurs when these diagnostics agree: a row lies in the rank-aggregated anomaly tail, deviates from its peer cell, and falls outside the conformal cost band. Such rows are natural candidates for manual audit, documentation requests, or linkage to additional data sources.

The limitations remain material. The anomaly layer is unsupervised because the public file has no official fraud label. The empirical run uses a reproducible large sample rather than every row in the detailed file. Open Payments is documented as an extension rather than merged into the present analytical tables. Several peer-relative and network features are computed before the provider split, so the cost evaluation is provider-held-out but not fully prospective in a deployment-pure sense.

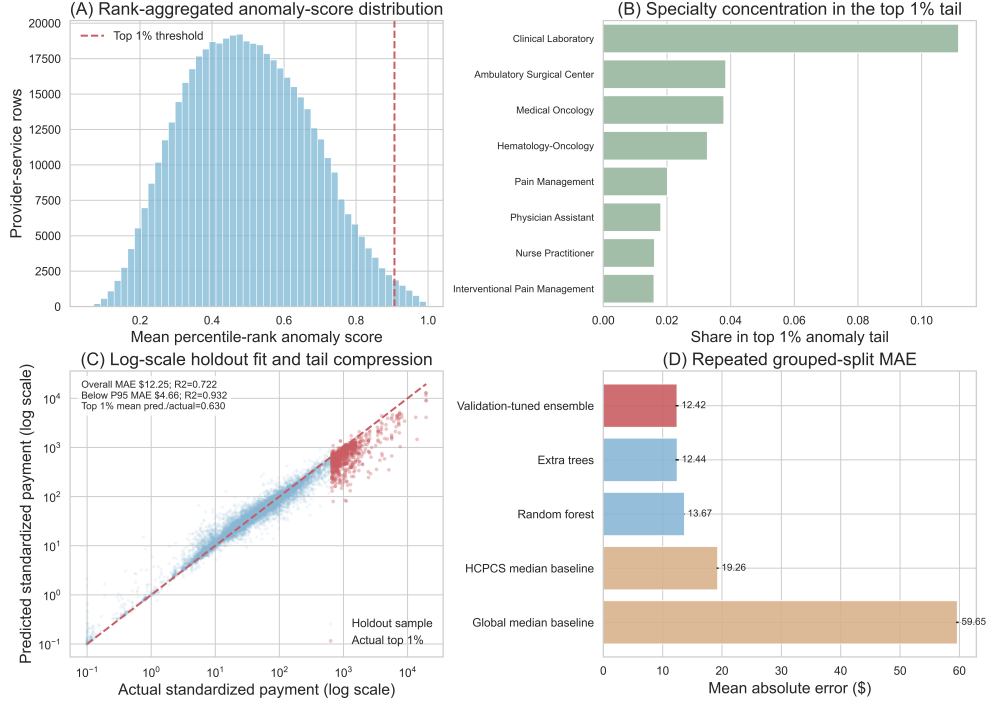


Fig. 1 Compact panel presentation of the main graphical results: (A) distribution of the rank-aggregated multi-view anomaly score with the top 1% threshold, (B) specialty concentration in the top anomaly tail, (C) provider-held-out cost prediction on log scales with the actual top 1% payment rows highlighted to reveal right-tail compression, and (D) repeated grouped-split mean absolute error with one-standard-deviation error bars across benchmark models.

The point predictor compresses the extreme right tail, as the highlighted top-1%-actual rows in Figure 1 show. Split-conformal prediction supplies marginal coverage but not subgroup-conditional validity. Finally, commercial or restricted data augmentation may improve validation, but any such extension would introduce linkage, selection, and governance issues that are outside the public-data reproducibility boundary of the present study. These limitations do not invalidate the workflow; they define the next empirical frontier.

5 Conclusion

This study developed a reproducible Medicare payment-integrity workflow that connects provider-service anomaly screening with expected-cost reference modeling. The framework is designed for public CMS data and therefore respects the evidentiary boundary between statistical triage and claims-based adjudication. Its central contribution is the integration of peer-relative features, specialty-service rarity, multi-view rank aggregation, tree-ensemble cost prediction, and split-conformal reference bands into a single explainable audit-support pipeline.

The empirical results show that the anomaly score identifies a concentrated and interpretable review tail rather than a simple list of high-payment rows. The flagged observations are shaped by payment ratios, peer-cell deviations, provider concentration, and service rarity, with elevated concentration in service environments where coding complexity, utilization intensity, and payment heterogeneity are operationally plausible. The cost-reference layer adds a complementary perspective: it gives a strong benchmark for routine standardized payments, while its residual inflation and conformal-band violations reveal where the expected-payment surface is least reliable. The most useful escalation signal is therefore not the anomaly score alone, but the agreement between anomaly ranking, peer-relative deviation, and out-of-band cost behavior.

The findings should be interpreted conservatively. Public provider-service files do not contain adjudicated fraud outcomes, so the workflow cannot and should not label misconduct. A high anomaly score indicates that a row deserves review; a large cost residual indicates that the observed payment is difficult to explain using the public reference model; and a conformal-band violation indicates that the row lies outside a calibrated public-data payment range. These diagnostics support prioritization, documentation review, and linkage to richer evidence, but they do not replace clinical, legal, or administrative judgment. The study also retains important limitations, including the absence of official fraud labels, the retrospective sampling design, partial transductive construction of peer and network features, and the marginal rather than subgroup-conditional nature of split-conformal coverage.

Future work should extend the design to multi-year provider-service panels, incorporate additional public sources such as Open Payments under explicit governance controls, and develop tail-specialist cost models for extreme payment behavior. Operational deployment should present reviewers with the full diagnostic bundle - anomaly rank, peer comparators, expected payment, conformal interval status, specialty context, and uncertainty - rather than a single opaque risk number. In this form, public Medicare data can support a transparent and policy-facing analytics prototype for payment-integrity monitoring while preserving the necessary distinction between anomaly detection and fraud determination.

Declarations

Funding. The authors received no specific funding for this work.

Data Availability Statement. The study uses publicly available Centers for Medicare and Medicaid Services Medicare Physician and Other Practitioners provider-service and provider aggregate files, all cited in the manuscript. The processed benchmark tables, figure outputs, and project scripts are available from the corresponding author upon reasonable request.

Conflict of interest. The authors declare that they have no competing interests.

Consent to Publish declaration: not applicable.

Consent to Participate declaration: not applicable.

Ethics declaration: not applicable.

References

- [1] Centers for Medicare & Medicaid Services: Fiscal Year 2025 Improper Payments Fact Sheet. CMS Newsroom fact sheet, 15 January 2026. Accessed 16 June 2026 (2026). <https://www.cms.gov/newsroom/fact-sheets/fiscal-year-2025-improper-payments-fact-sheet>
- [2] Centers for Medicare & Medicaid Services: Center for Program Integrity. CMS program-integrity page. Accessed 16 June 2026 (2026). <https://www.cms.gov/medicare/medicaid-coordination/center-program-integrity>
- [3] Centers for Medicare & Medicaid Services: Medicare Physician & Other Practitioners - by Provider. CMS Data. Accessed 28 May 2026 (2026). <https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-provider>
- [4] Centers for Medicare & Medicaid Services: Medicare Physician & Other Practitioners - by Provider and Service. CMS Data. Accessed 28 May 2026 (2026). <https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-provider-and-service>
- [5] Centers for Medicare & Medicaid Services: Medicare Claims Synthetic Public Use Files (SynPUFs). CMS overview page last modified 10 September 2024. Accessed 28 May 2026 (2024). <https://www.cms.gov/data-research/statistics-trends-and-reports/medicare-claims-synthetic-public-use-files>
- [6] Hancock, J.T., Bauder, R.A., Wang, H., Khoshgoftaar, T.M.: Explainable machine learning models for medicare fraud detection. *Journal of Big Data* **10**(1), 154 (2023) <https://doi.org/10.1186/s40537-023-00821-5>
- [7] Hancock, J.T., Wang, H., Khoshgoftaar, T.M., Liang, Q.: Data reduction techniques for highly imbalanced medicare big data. *Journal of Big Data* **11**(1), 8 (2024) <https://doi.org/10.1186/s40537-023-00869-3>
- [8] Johnson, J.M., Khoshgoftaar, T.M.: Data-centric ai for healthcare fraud detection. *SN Computer Science* **4**(4), 389 (2023) <https://doi.org/10.1007/s42979-023-01809-x>
- [9] Farahmandazad, D., Danesh, K., Abadi, H.F.N.: Application of standard machine learning models for medicare fraud detection with imbalanced data. *Risks* **13**(10), 198 (2025) <https://doi.org/10.3390/risks13100198>
- [10] Wang, Z., Chen, X., Wu, Y., Jiang, L., Lin, S., Qiu, G.: A robust and interpretable

- ensemble machine learning model for predicting healthcare insurance fraud. *Scientific Reports* **15**(1), 218 (2025) <https://doi.org/10.1038/s41598-024-82062-x>
- [11] Curtis, E.D., Billion-Polak, P., Khoshgoftaar, T.M., Furht, B.: A review of distinct machine learning classifiers for healthcare fraud detection. *Journal of Big Data* **12**(1), 238 (2025) <https://doi.org/10.1186/s40537-025-01295-3>
 - [12] Razzaq, K., Shah, M.: Next-generation machine learning in healthcare fraud detection: Current trends, challenges, and future research directions. *Information* **16**(9), 730 (2025) <https://doi.org/10.3390/info16090730>
 - [13] Leevy, J.L., Hancock, J.T., Khoshgoftaar, T.M., Abdollah Zadeh, A.: Investigating the effectiveness of one-class and binary classification for fraud detection. *Journal of Big Data* **10**(1), 157 (2023) <https://doi.org/10.1186/s40537-023-00825-1>
 - [14] Walauskis, M.A., Khoshgoftaar, T.M.: Unsupervised label generation for severely imbalanced fraud data. *Journal of Big Data* **12**(1), 63 (2025) <https://doi.org/10.1186/s40537-025-01120-x>
 - [15] De Meulemeester, H., De Smet, F., Dorst, J., Derroitte, E., De Moor, B.: Explainable unsupervised anomaly detection for healthcare insurance data. *BMC Medical Informatics and Decision Making* **25**(1), 14 (2025) <https://doi.org/10.1186/s12911-024-02823-6>
 - [16] Shekhar, S., Leder-Luis, J., Akoglu, L.: Can machine learning target health care fraud? evidence from medicare hospitalizations. *Journal of Policy Analysis and Management* **45**(1), 70078 (2026) <https://doi.org/10.1002/pam.70078>
 - [17] Tajrobehkar, M., Guo, X., Nguyen, D., Chigullapally, N., Shravah, V., Yee, S., Chozhan, A., Hoang, K., Chang, R.T., Gutierrez, C., Lee, S.: Utilization analysis and fraud detection in medicare via machine learning. *medRxiv* (2025) <https://doi.org/10.1101/2024.12.30.24319784> . Preprint
 - [18] Breiman, L.: Random forests. *Machine Learning* **45**(1), 5–32 (2001) <https://doi.org/10.1023/A:1010933404324>
 - [19] Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. *Machine Learning* **63**(1), 3–42 (2006) <https://doi.org/10.1007/s10994-006-6226-1>
 - [20] Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on typical tabular data? In: *Advances in Neural Information Processing Systems* 35 (2022). <https://arxiv.org/abs/2207.08815>
 - [21] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems* 31, pp. 6638–6648 (2018). <https://arxiv.org/abs/1706.09516>

- [22] Liu, F.T., Ting, K.M., Zhou, Z.-H.: Isolation forest. In: Proceedings of the 2008 IEEE International Conference on Data Mining, pp. 413–422 (2008). <https://doi.org/10.1109/ICDM.2008.17>
- [23] Li, Z., Zhao, Y., Hu, X., Botta, N., Ionescu, C., Chen, G.H.: Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering* **35**(12), 12181–12193 (2023) <https://doi.org/10.1109/TKDE.2022.3159580>
- [24] Lei, J., G’Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-free predictive inference for regression. *Journal of the American Statistical Association* **113**(523), 1094–1111 (2018) <https://doi.org/10.1080/01621459.2017.1307116>
- [25] Romano, Y., Patterson, E., Candès, E.J.: Conformalized quantile regression. In: *Advances in Neural Information Processing Systems* 32, pp. 3538–3548 (2019). <https://papers.nips.cc/paper/8613-conformalized-quantile-regression>
- [26] Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J.: Predictive inference with the jackknife+. *The Annals of Statistics* **49**(1), 486–507 (2021) <https://doi.org/10.1214/20-AOS1965>
- [27] Centers for Medicare & Medicaid Services: CMS Open Payments API. Accessed 28 May 2026 (2026). <https://openpaymentsdata.cms.gov/about/api>