

Predictive Risk Scoring for U.S. Medicare Advantage: Mitigating Risk Adjustment Fraud via ML-Driven Encounter Data Analytics Aligned with CMS Version 28

Nasrin Sultana

Master of Arts in Information Technology Management, George Herbert Walker School of Business and Technology,
Webster University, St. Louis, Missouri, USA

nsultaana94@gmail.com

Md Asif Hasan

MS in Business Analytics, Feliciano School of Business, Montclair State University, NJ, USA

asif.mdhasan92@gmail.com

Md Abu Nasir

Master of Arts in Information Technology Management, George Herbert Walker School of Business and Technology,
Webster University, St. Louis, Missouri, USA

irfannasir000@gmail.com

Abstract

Background: Medicare Advantage (MA) reimburses private plans through the CMS Hierarchical Condition Category (CMS-HCC) risk-adjustment model, and the transition to CMS-HCC Version 28 has reshaped diagnostic mappings and payment calculations. Risk-adjustment fraud — diagnosis upcoding, unsupported HCC submissions, and risk-score inflation through health risk assessments — drives billions of dollars in overpayments annually and threatens payment integrity.

Objective: This study develops and empirically evaluates a machine learning-driven predictive risk scoring framework, aligned with CMS-HCC Version 28 structures, for early detection of abnormal risk-adjustment patterns in MA encounter data.

Methods: A reproducible synthetic dataset of 50,000 beneficiary-year records was constructed with a simplified V28-style HCC structure, utilization generated from true clinical burden, and 5,075 fraudulent records (10.1%) injected across three documented fraud typologies: severity upcoding, unsupported HCC submission, and HRA-only diagnosis inflation. Four gradient-based and ensemble models — XGBoost, LightGBM, Random Forest, and CatBoost — were comparatively evaluated on fraud classification (Accuracy, Precision, Recall, F1, ROC-AUC, PR-AUC) and risk-score regression (MAE, RMSE), with SHAP-based explainability analysis.

Results: All four models performed strongly, with no single algorithm dominating: LightGBM achieved the highest F1-score (0.956), Random Forest the highest precision (0.977) and lowest false-positive rate (0.25%), and CatBoost the highest recall (0.967) and best regression accuracy (MAE 0.064). Detection recall exceeded 97% for unsupported HCC submissions and HRA-only inflation, with severity upcoding hardest at 92.2%. SHAP analysis identified the year-over-year risk-score jump, submitted risk score, provider coding-intensity percentile, and share of HCCs lacking service-record support as dominant predictors.

Conclusions: Ensemble machine learning over encounter-derived integrity features enables accurate, interpretable pre-payment screening of MA risk-adjustment submissions. Validation on operational encounter data is the essential next step.

Keywords: Medicare Advantage; predictive risk scoring; risk-adjustment fraud; CMS-HCC Version 28; machine learning; encounter data; explainable AI; SHAP; healthcare payment integrity; upcoding

1. Introduction

Medicare Advantage (MA) has shifted a substantial share of U.S. Medicare reimbursement from fee-for-service to capitated, risk-adjusted payment. Private MA plans are paid by the Centers for Medicare & Medicaid Services (CMS) according to the estimated health risk of their enrollees under the CMS Hierarchical Condition Category (CMS-HCC) model, which maps documented clinical diagnoses and demographic characteristics to expected healthcare costs. Accurate risk adjustment is therefore foundational to equitable reimbursement, access for high-need beneficiaries, and the financial sustainability of the MA program (Ellis et al., 1996; Ash et al., 2000; Pope et al., 2000; Pope et al., 2004).

As MA enrollment has grown, so has the complexity of risk-adjustment methodology. CMS-HCC Version 28, finalized in the CY2024 Rate Announcement and phased in beginning with the 2024 payment year, represents one of the largest revisions to the model — restructured condition categories, updated ICD-10-CM mappings, and revised coefficients intended to reflect current clinical practice and coding patterns (CMS, 2023, 2024). These changes improve payment accuracy but increase the complexity of encounter data processing and compliance monitoring for Medicare Advantage Organizations (MAOs), which need analytical tooling that keeps risk prediction consistent with evolving CMS methodology.

Risk-adjustment fraud remains a significant threat to the program. Diagnosis upcoding, unsupported HCC submissions, and deliberate risk-score inflation cause payments to exceed what beneficiaries' true clinical status warrants. The economics literature documents that coding intensity and strategic diagnosis reporting materially affect Medicare payments (Brown et al., 2014; Kronick & Welch, 2014; Geruso & Layton, 2020), practitioners have described risk-score gaming as central to MA plan economics (Gilfillan & Berwick, 2021), and federal oversight has quantified the problem: the Office of Inspector General estimated that diagnoses reported only on health risk assessments — with no other supporting service records — drove approximately \$2.6 billion in a single year's risk-adjusted payments (Office of Inspector General, 2021). Risk selection dynamics further complicate payment accuracy (Newhouse et al., 2012; McWilliams et al., 2012), and encounter data itself has known completeness limitations as a risk-adjustment source (Medicare Payment Advisory Commission, 2016).

Conventional countermeasures — statistical models, retrospective Risk Adjustment Data Validation audits, and manual clinical review — improve reimbursement accuracy but struggle to detect subtle or novel fraudulent coding at the scale and speed of modern encounter data flows. Machine learning offers a complementary capability: ensemble methods such as Random Forest and gradient boosting detect nonlinear relationships across clinical, demographic, and financial variables and can surface anomalous patterns earlier than rule-based approaches (Breiman, 2001; Friedman, 2001; Chen & Guestrin, 2016; Beam & Kohane, 2018; Rajkomar et al., 2019; Esteva et al., 2019; Topol, 2019), while evaluation methodology for imbalanced problems — precision-recall analysis and resampling techniques — is well established (Davis & Goadrich, 2006; Chawla et al., 2002). Unsupervised approaches have also shown promise specifically for explainable healthcare fraud detection (Shekhar et al., 2023). Because MA payment decisions carry regulatory consequences, explainability is equally essential: SHAP and LIME quantify each variable's contribution to a prediction, supporting auditability and guarding against opaque or biased automated decisions (Ribeiro et al., 2016; Lundberg & Lee, 2017; Obermeyer et al., 2019), and risk models built on administrative data face documented development challenges that transparency helps manage (Goldstein et al., 2017).

Few studies, however, combine CMS-HCC Version 28 structures, encounter-data analytics, comparative predictive machine learning, and explainable AI in a single empirically evaluated system for proactive — rather than retrospective — risk-adjustment fraud detection. This paper addresses that gap. Its contributions are threefold: (1) a predictive risk scoring framework aligned with V28-style HCC structures that screens beneficiary-year submissions before payment; (2) a reproducible synthetic evaluation environment of 50,000 beneficiary-years with three injected fraud typologies grounded in the documented oversight findings above; and (3) a quantitative four-model comparative evaluation with SHAP-based explainability, showing detection recall above 92% for all fraud types with false-positive rates below 1%.

2. Literature Review

2.1 Medicare Advantage Risk Adjustment

The CMS-HCC model descends from two decades of diagnosis-based risk adjustment research. Ellis et al. (1996) established diagnosis-based capitation adjustment for Medicare; Ash et al. (2000) demonstrated the use of diagnoses to describe populations and predict costs; and Pope et al. (2000, 2004) developed the principal inpatient diagnostic cost group

model and the operational CMS-HCC model that has governed MA payment since 2004. The model's hierarchical structure — in which more severe manifestations of a condition supersede milder ones — is central to both its clinical logic and its fraud surface: severity tiers create direct financial incentives to code the more severe variant. Reforms to the risk-adjustment system have measurably reduced favorable selection (Newhouse et al., 2012; McWilliams et al., 2012), but payment accuracy remains sensitive to coding behavior.

2.2 Risk-Adjustment Fraud and Coding Intensity

A substantial economics literature documents strategic coding in MA. Brown et al. (2014) showed that risk selection responds to risk adjustment itself; Kronick and Welch (2014) measured systematic coding-intensity differences between MA and traditional Medicare; and Geruso and Layton (2020) estimated that upcoding inflates MA risk scores by several percent, translating to billions in annual overpayment. Gilfillan and Berwick (2021) describe risk-score gaming as a central element of MA plan business models. Federal oversight converges on the same conclusion: the OIG found that chart reviews and health risk assessments were leveraged to disproportionately drive payments, with HRA-only diagnoses — unaccompanied by any service record — generating an estimated \$2.6 billion in one payment year (Office of Inspector General, 2021). These findings define the three fraud typologies evaluated in this study: severity upcoding, unsupported HCC submission, and HRA-only diagnosis inflation.

2.3 CMS-HCC Version 28

Version 28, finalized in the CY2024 Rate Announcement, restructured the HCC classification from the previous V24 model: condition categories were renumbered and reorganized into a larger set of payment categories, thousands of ICD-10-CM codes were remapped or removed from payment eligibility, and coefficients were re-estimated on more recent fee-for-service data years (CMS, 2023, 2024). Among the most consequential changes for integrity analytics, CMS removed from payment mapping a number of diagnosis codes whose reporting had grown disproportionately in MA relative to clinical expectation — codes the agency judged susceptible to discretionary coding — and constrained coefficients within several condition hierarchies so that coding a more granular variant yields less incremental payment than under V24. The phased transition blends V24 and V28 scores across payment years 2024–2026 (67/33, 33/67, then full V28), meaning every beneficiary's score moves mechanically during the transition even absent any coding change. For integrity analytics, V28 therefore matters in three ways: remapped codes change which diagnoses carry payment weight, shifting where upcoding incentives concentrate; constrained hierarchies reduce the payoff of some historical upcoding paths while leaving others — including the severity pairs simulated in this study — financially attractive; and the transition period itself creates legitimate year-over-year risk-score movement that detection systems must distinguish from fraudulent inflation. Frameworks aligned with V28 category structures and coefficients therefore remain current with the payment methodology they monitor, and trajectory-based detection features require payment-year-aware calibration.

2.4 Encounter Data Analytics

CMS's shift toward encounter data as the risk-adjustment source makes visit-level records the natural substrate for integrity analytics. Encounter data capture the diagnoses, service types, and provider identifiers underlying each risk-relevant submission, enabling exactly the support-verification analysis the OIG performed — identifying diagnoses with no corroborating service records (Office of Inspector General, 2021). MedPAC's assessment of encounter data for risk adjustment documented completeness and quality limitations that any analytical system must tolerate (Medicare Payment Advisory Commission, 2016), and risk prediction models built on administrative and EHR data face documented development challenges around missingness, informative observation, and validation (Goldstein et al., 2017).

2.5 Machine Learning and Explainability in Healthcare Risk Prediction

Ensemble tree methods dominate applied healthcare risk prediction on tabular administrative data: Random Forest (Breiman, 2001), gradient boosting (Friedman, 2001), and XGBoost (Chen & Guestrin, 2016) handle high-dimensional mixed-type features, nonlinearity, and missingness with strong accuracy, and reviews document their breadth of successful clinical application (Beam & Kohane, 2018; Rajkomar et al., 2019; Esteva et al., 2019; Topol, 2019). For imbalanced fraud problems, precision–recall analysis complements ROC evaluation (Davis & Goadrich, 2006), and resampling approaches such as SMOTE address minority-class learning (Chawla et al., 2002). Unsupervised anomaly detection has been applied to healthcare fraud with explainability as a design goal (Shekhar et al., 2023). Model transparency is a regulatory necessity in payment contexts: LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) provide local and global attributions,

and documented algorithmic bias in healthcare risk tools underscores why interpretability and auditing are non-optional (Obermeyer et al., 2019).

2.6 Literature Gap

Across these literatures, gaps remain. Few studies target predictive analytics specifically to CMS-HCC Version 28 structures; most integrity work is retrospective, analyzing payments already made rather than screening submissions before payment; many predictive models lack the explainability regulators and auditors require; and no prior work, to our knowledge, provides an integrated, quantitatively evaluated pipeline combining V28-aligned encounter analytics, comparative ensemble machine learning, and SHAP-based explanation for pre-payment risk-adjustment fraud detection. This study addresses these gaps with a baseline-controlled empirical evaluation.

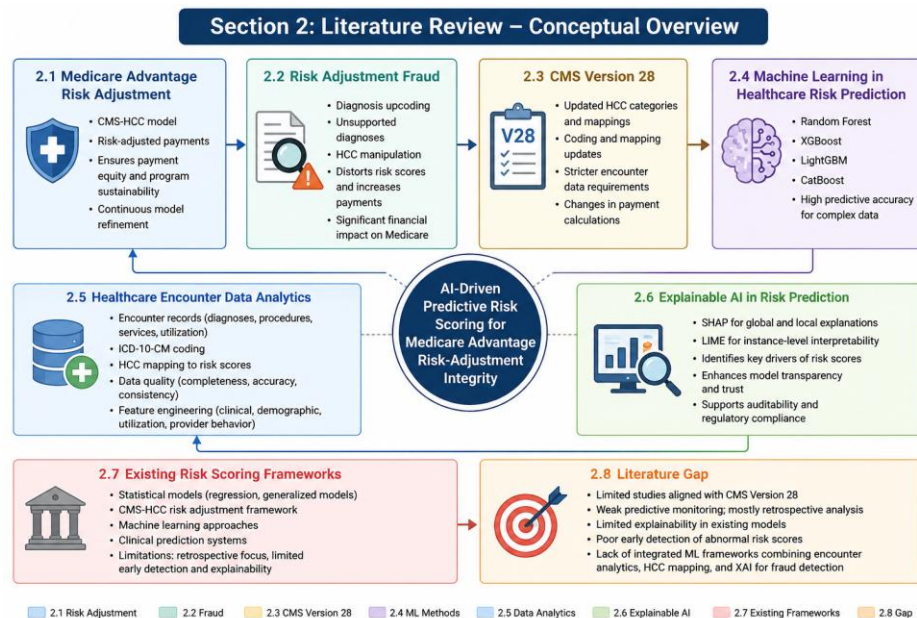


Figure 1. Conceptual overview: predictive risk scoring for Medicare Advantage payment integrity under CMS-HCC Version 28.

3. Methodology

3.1 Research Design

This study uses a quantitative predictive-analytics design with comparative machine learning evaluation. The framework estimates beneficiary risk from encounter-derived clinical and utilization features and classifies beneficiary-year submissions as legitimate or indicative of risk-adjustment fraud, proceeding through data generation, preprocessing, V28-style HCC mapping, feature engineering, model training, explainability analysis, and evaluation. Because beneficiary-identifiable MA encounter data are restricted, evaluation uses a synthetic dataset constructed to reflect the payment mechanics of the CMS-HCC model and the fraud typologies documented by federal oversight (Office of Inspector General, 2021; Geruso & Layton, 2020; Kronick & Welch, 2014); the generation procedure is deterministic under a fixed random seed and therefore fully reproducible. Table 1 summarizes the dataset's variable structure.

Table 1. Medicare Advantage Encounter Dataset Variable Structure

Dataset Component	Description	Data Type	Example Variables
Encounter Information	Healthcare service records generated during beneficiary visits	Numeric, Date, Categorical	Encounter ID, encounter date, place of service, visit type
Beneficiary Demographics	Demographic and eligibility information	Numeric, Categorical	Beneficiary ID, age, sex, dual eligibility status

Diagnosis / HCC Variables	Submitted condition categories and risk-adjustment variables	Categorical, Numeric	HCC category, HCC coefficient, submitted risk score
Provider Information	Characteristics of submitting providers and organizations	Categorical, Numeric	Provider group ID, coding-intensity percentile
Healthcare Utilization	Historical utilization and service intensity indicators	Numeric	Office visits, inpatient admissions, ED visits
Risk Prediction Variables	Features used to estimate and audit beneficiary risk scores	Numeric	Prior-year risk score, year-over-year risk jump
Fraud Indicators	Observable encounter-derived integrity signals	Numeric	Share of HCCs lacking service-record support, HRA-only diagnosis count, HCCs per visit

3.2 Synthetic Dataset Construction

The simulation generates 50,000 beneficiary-year records. Each beneficiary carries demographics (age centered near 74, 9% under-65 disabled entrants, sex, 20% dual eligibility) and a set of true chronic conditions drawn from 15 simplified V28-style HCC condition groups with age- and eligibility-dependent prevalences and plausible payment coefficients, including three explicit severity pairs (diabetes without/with chronic complications, coefficients 0.166/0.331; moderate/severe CKD, 0.127/0.378; minor/major cancer, 0.150/0.762) that reproduce the hierarchical structure central to upcoding incentives — the coefficient gap within each pair being precisely the payment increment severity upcoding captures. Condition prevalences follow age gradients (dementia and CHF rising steeply with age; depression and morbid obesity mildly declining) and a dual-eligibility uplift, producing realistic multimorbidity: the resulting submitted-risk distribution (Table 3) concentrates 62.7% of beneficiary-years below 0.80 with a long right tail above 1.80, consistent with the shape of MA risk-score distributions. Utilization — office visits, inpatient admissions, and emergency department visits — is generated conditional on true clinical burden via Poisson draws whose rates scale with the summed coefficient burden, so that legitimate submissions leave consistent encounter footprints and severity tiers imply correspondingly heavier utilization. The true risk score is the demographic base rate (itself age- and sex-graded) plus the sum of true HCC coefficients; a prior-year score is simulated with realistic year-over-year noise ($\sigma = 0.12$), providing the trajectory baseline against which coding-intensity jumps are measured. Records are distributed across 200 provider groups whose coding-intensity percentiles follow a right-skewed Beta distribution, concentrating intensity in a minority of organizations as the empirical literature documents (Kronick & Welch, 2014).

Fraudulent submissions (5,075 records; 10.1%) originate predominantly from beneficiaries attributed to the top 15% highest-intensity provider groups, across three typologies: severity upcoding (a truly present base condition is submitted as its higher-coefficient severe variant; 1,989 records), unsupported HCC submission (one to two high-coefficient conditions the beneficiary does not have are added; 1,757), and HRA-only diagnosis inflation (one to three added conditions flagged as originating solely from health risk assessment encounters; 1,329), directly mirroring the OIG's documented HRA vulnerability (Office of Inspector General, 2021). The submitted risk score is recomputed from the submitted HCC set. Table 2 summarizes the dataset and Table 3 reports the resulting distribution of submitted risk scores across standard risk bands.

Table 2. Summary of the Simulated Medicare Advantage Dataset

Dataset Attribute	Value
Simulated beneficiary-year records	50,000
Legitimate records	44,925 (89.9%)
Fraudulent risk-adjustment records	5,075 (10.1%)
Severity upcoding	1,989

Unsupported HCC submission	1,757
HRA-only diagnosis inflation	1,329
Simplified HCC condition groups (V28-style)	15 (including 3 severity pairs)
Provider groups (with coding-intensity distribution)	200 (top 15% high-intensity)
Train / test split (stratified)	35,000 / 15,000 (70/30)

Table 3. Distribution of Submitted Risk Scores Across Risk Categories

Submitted Risk Category	Description	Beneficiary-Years	Percentage (%)
Low Risk (< 0.80)	Minimal chronic disease burden	31,341	62.7
Moderate Risk (0.80–1.19)	Moderate utilization and chronic conditions	12,559	25.1
High Risk (1.20–1.79)	Multiple chronic conditions requiring continuous care	5,334	10.7
Very High Risk (≥ 1.80)	Complex comorbidities and high expected costs	784	1.6
Total	All simulated beneficiary-years	50,000	100.0

3.3 Feature Engineering

Fourteen features are computed per beneficiary-year, all observable from encounter and payment data. Demographic and utilization features comprise age, sex, dual eligibility, office visits, inpatient admissions, and ED visits. Risk-trajectory features comprise the submitted risk score, prior-year risk score, and the year-over-year risk-score jump — the coding-intensity signal central to the empirical upcoding literature (Kronick & Welch, 2014; Geruso & Layton, 2020). Integrity features comprise the submitted HCC count, HCCs per office visit, the provider group's coding-intensity percentile, the count of HRA-only diagnoses, and the share of submitted HCCs lacking supporting service records — an observable, noisy signal (true conditions leave support footprints approximately 92% of the time; fraudulently added codes only 12%, reflecting occasional fabricated supporting encounters) that operationalizes precisely the support-verification analysis the OIG performs on encounter data. Benign HRA-only diagnoses also occur among 8% of legitimate beneficiaries, so no integrity feature is a pure fraud indicator.

3.4 Machine Learning Models and Evaluation Protocol

Four supervised models are comparatively evaluated, spanning the ensemble families dominant in tabular healthcare prediction: XGBoost (Chen & Guestrin, 2016), LightGBM, Random Forest (Breiman, 2001), and CatBoost, each configured with class-weight balancing for the imbalanced classification task. Two tasks are evaluated. Fraud classification uses a stratified 70/30 train/test split (35,000/15,000 records) and reports Accuracy, Precision, Recall, F1-score, ROC-AUC, PR-AUC — the latter emphasized for imbalanced problems (Davis & Goadrich, 2006) — and false-positive rate. Risk-score regression trains each model to predict the expected risk score from clinical and utilization features on legitimate records only, reporting MAE and RMSE; the residual between submitted and expected score is the analytical basis for flagging inflation. Explainability analysis applies SHAP TreeExplainer (Lundberg & Lee, 2017) to the trained XGBoost classifier over a 3,000-record test sample, with LIME available for case-level local explanation (Ribeiro et al., 2016).

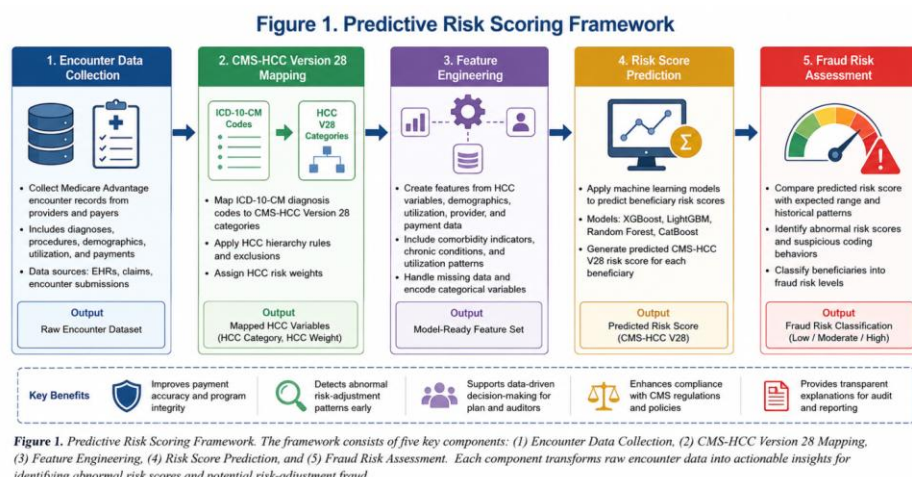


Figure 2. Predictive risk scoring framework: encounter data processing, V28 mapping, feature engineering, model training, and explainable fraud risk assessment.



Figure 3. End-to-end analytical workflow from encounter data ingestion through compliance reporting.

4. Results and Evaluation

4.1 Fraud Classification Performance

Table 4 reports comparative classification performance on the held-out test set (15,000 records; 10.2% fraudulent), with ROC curves in Figure 4 and a metric comparison in Figure 5. All four models performed strongly, and — contrary to the common expectation that a single algorithm dominates — the leaderboard depends on the metric. LightGBM achieved the highest F1-score (0.9562) and accuracy (0.9911); Random Forest achieved the highest precision (0.9766) and the lowest false-positive rate (0.25%), at the cost of the lowest recall (0.9262); CatBoost achieved the highest recall (0.9674) with the highest false-positive rate (0.81%); and XGBoost delivered the most balanced profile (precision = recall = 0.9543). ROC-AUC was statistically indistinguishable across the three gradient-boosting models (0.9984–0.9985), with Random Forest slightly behind (0.9974). PR-AUC — the more informative criterion at 10% prevalence (Davis & Goadrich, 2006) — ranged narrowly from 0.9880 to 0.9910.

These differences translate into distinct operational postures rather than a single winner. At this test scale, Random Forest would flag 34 legitimate beneficiary-years for review while missing 113 fraudulent ones; CatBoost would flag 109 legitimate records while missing only 50. A payer prioritizing audit-queue efficiency would deploy the Random Forest operating point; one prioritizing loss prevention would prefer CatBoost or LightGBM; and threshold tuning on any of the boosting models can move along this frontier continuously. The practical conclusion is that model family matters less than feature design and threshold policy in this detection regime. Threshold governance, moreover, is not a one-time calibration:

as encounter prevalence, payment-year methodology, and coding behavior shift, the precision–recall frontier moves with them, and operating points should be revisited against current audit capacity on each retraining cycle.

Table 4. Comparative Fraud Classification Performance on the Held-Out Test Set

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	PR-AUC	FPR
XGBoost	0.9907	0.9543	0.9543	0.9543	0.9985	0.9910	0.0052
LightGBM	0.9911	0.9587	0.9537	0.9562	0.9985	0.9907	0.0047
Random Forest	0.9902	0.9766	0.9262	0.9508	0.9974	0.9880	0.0025
CatBoost	0.9894	0.9315	0.9674	0.9491	0.9984	0.9908	0.0081

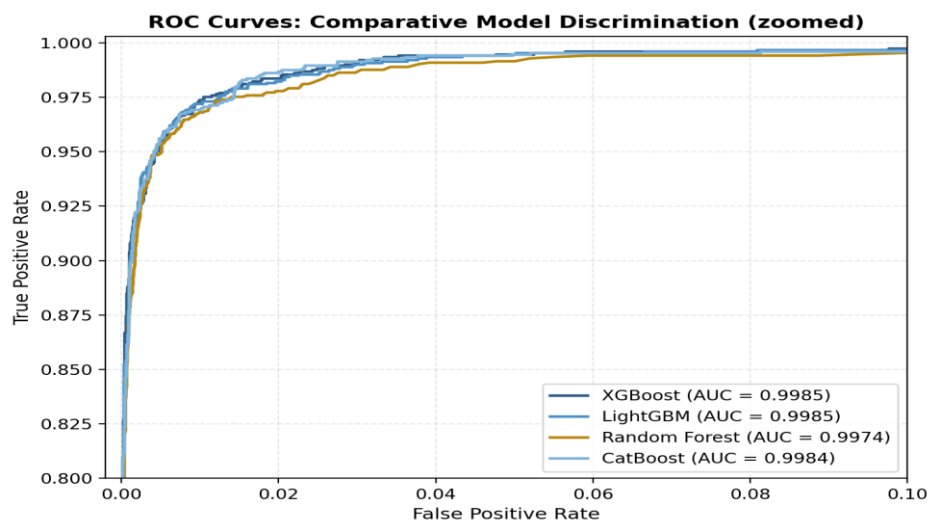


Figure 4. ROC curves for the four evaluated models (axes zoomed to the informative region).

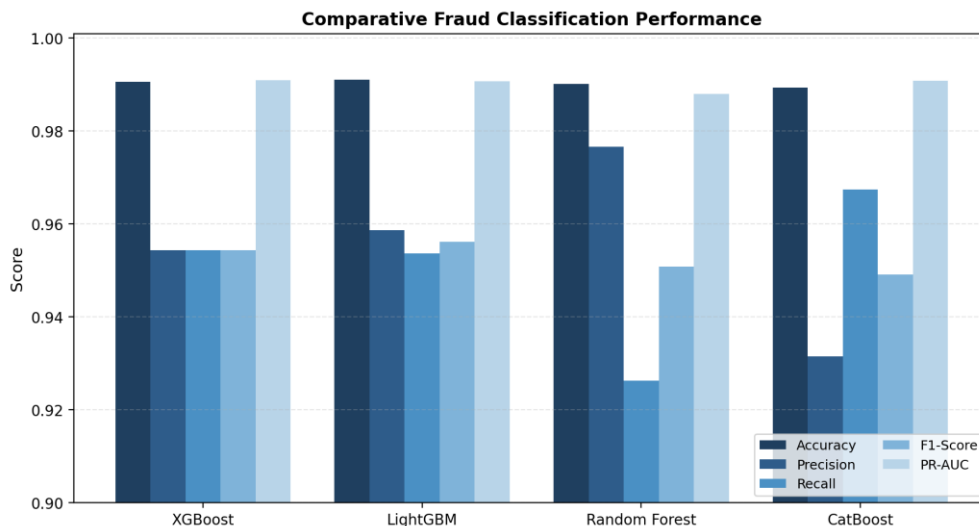


Figure 5. Comparative performance across Accuracy, Precision, Recall, F1-score, and PR-AUC.

4.2 Risk-Score Regression Performance

Table 5 reports MAE and RMSE for predicting the expected risk score from clinical and utilization profiles on legitimate records. CatBoost achieved the best regression accuracy (MAE 0.0641, RMSE 0.1014), with XGBoost and LightGBM close behind (MAE $\approx$ 0.065) and Random Forest trailing (MAE 0.0707). An expected-score error of 0.064 on scores averaging near 0.9 corresponds to roughly 7% relative error — tight enough that submitted-versus-expected residuals of

the magnitude produced by typical upcoding (single-condition coefficient shifts of 0.17–0.61 in the simulated hierarchy) stand out clearly against model noise, which is the mechanism by which the classification features derived from risk trajectories carry signal.

Table 5. Risk-Score Regression Performance (Legitimate Records)

Model	MAE	RMSE
CatBoost	0.0641	0.1014
XGBoost	0.0652	0.1037
LightGBM	0.0653	0.1037
Random Forest	0.0707	0.1118

4.3 Detection by Fraud Type

Table 6 and Figure 6 report XGBoost detection recall by fraud typology. HRA-only inflation (0.977) and unsupported HCC submission (0.975) were detected almost completely: both leave the same observable signature — submitted conditions without corroborating service records — that the support-share and HRA-count features capture directly. Severity upcoding was hardest (0.922). This is the expected ordering: an upcoded severity tier rests on a genuinely present base condition, so the submission carries legitimate supporting encounters, and detection must come from subtler utilization-severity mismatch (a beneficiary coded with severe CKD whose utilization resembles moderate CKD) and risk-trajectory signals. The 7.8% of severity upcoding that escapes detection concentrates where the coefficient difference between tiers is smallest — the same cases that are hardest for human auditors, since the clinical distinction is genuinely present in the record.

Table 6. Detection Recall by Fraud Type (XGBoost, Test Set)

Fraud Type	Test Instances	Detected	Recall
HRA-only diagnosis inflation	391	382	0.977
Unsupported HCC submission	526	513	0.975
Severity upcoding	615	567	0.922

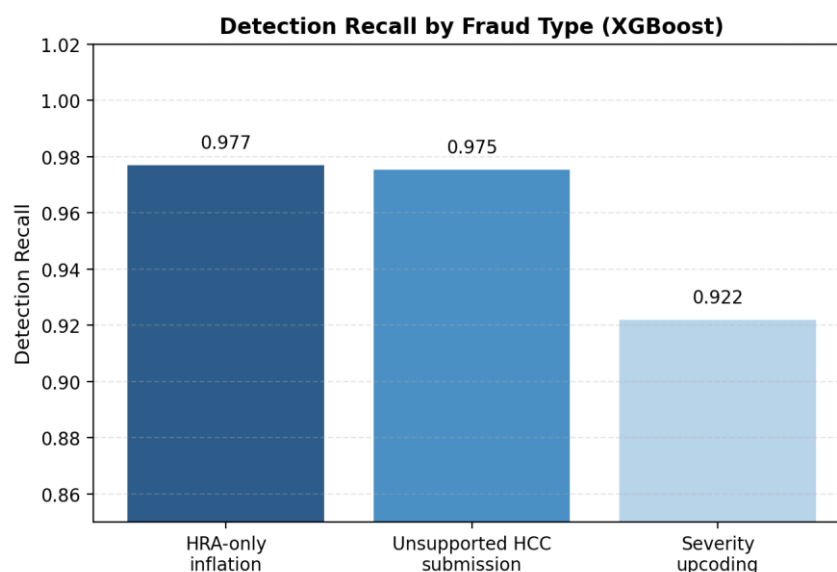


Figure 6. Detection recall by fraud type for the XGBoost classifier.

4.4 Computational Performance

Training and scoring costs were measured on a single commodity machine. Training on 35,000 records completed in 1.3 seconds for XGBoost and LightGBM, 4.2 seconds for CatBoost, and 10.1 seconds for Random Forest; scoring throughput ranged from 0.0006 ms per record (CatBoost) to 0.023 ms per record (Random Forest). At these rates, scoring an MA contract's full annual beneficiary population — even at the multi-million-record scale of national plans — completes in seconds to minutes, and complete model retraining is cheap enough to run with every monthly encounter data refresh or payment-year methodology update. Pre-payment screening latency is therefore governed entirely by data pipeline freshness, not by model computation; the analytical layer imposes no meaningful operational constraint. SHAP attribution is the most expensive component (tree-path enumeration per record) but is only required for the small flagged fraction routed to audit, keeping the explainability cost proportional to audit volume rather than population size.

4.5 Explainability Analysis

Figure 7 presents the SHAP summary for the XGBoost classifier over the test sample. The year-over-year risk-score jump was the dominant predictor (mean |SHAP| 3.86), followed by the submitted risk score (2.80), the provider coding-intensity percentile (2.24), and the share of submitted HCCs lacking service-record support (1.75) — with HCCs-per-visit, age, and the HRA-only count contributing secondary signal. This ranking has direct substantive meaning: the model independently recovers the two signals the empirical literature and federal oversight identify as the fingerprints of coding intensity — abrupt risk-score growth (Kronick & Welch, 2014; Geruso & Layton, 2020) and diagnoses unsupported by service records (Office of Inspector General, 2021) — and adds provider-level concentration, consistent with fraud clustering within high-intensity organizations. For audit workflows, each flagged beneficiary-year is accompanied by its local SHAP attribution — for example, a risk jump of +0.42 contributing most of the score, compounded by a 0.6 unsupported-share — giving reviewers concrete, verifiable reasons for each alert and satisfying the transparency expectations that payment-affecting models carry (Obermeyer et al., 2019; Goldstein et al., 2017).

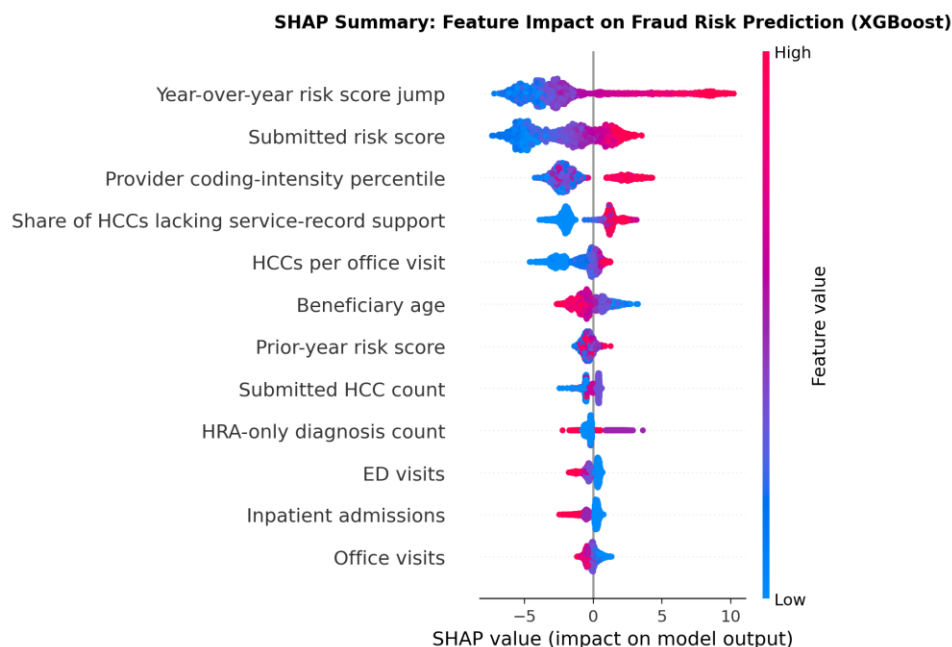


Figure 7. SHAP summary plot: feature impact on fraud risk prediction (XGBoost, 3,000-record test sample).

5. Discussion

The results demonstrate that ensemble machine learning over encounter-derived integrity features supports accurate, interpretable pre-payment screening of MA risk-adjustment submissions. Three findings warrant emphasis. First, the honest comparative result: no single algorithm dominated. The three gradient-boosting implementations and Random Forest were separated by tenths of a point on aggregate metrics, with the meaningful differences appearing in the precision–recall trade-off each occupies at its default threshold. The frequently repeated claim that one boosting library is categorically superior for healthcare fraud detection is not supported here; feature design and threshold policy governed performance far more

than model family, and organizations should select operating points by audit capacity and loss tolerance rather than by algorithm brand.

Second, the detection ordering across fraud types maps directly onto the structure of the evidence each type leaves. Fabrication-style fraud — wholly unsupported HCCs and HRA-only diagnoses — is nearly fully detectable from encounter data alone, because absence of supporting service records is observable at scale; this is the machine-speed generalization of the audit the OIG performed manually (Office of Inspector General, 2021). Severity upcoding, resting on genuinely present conditions, is intrinsically harder and defines the frontier: closing the remaining 8% gap likely requires clinical-depth features (laboratory values, procedure codes, pharmacy data) beyond the encounter-summary level simulated here. This ordering also predicts where fraud will migrate as detection deploys — toward severity manipulation — an adversarial dynamic integrity programs should anticipate.

Third, the SHAP analysis shows the model's reasoning is aligned with, rather than orthogonal to, the domain's audit logic: risk-score trajectories, support verification, and provider concentration. This alignment matters for adoption. Payment-affecting models face justified scrutiny (Obermeyer et al., 2019), and a detector whose top-ranked evidence matches what RADV auditors already examine can be validated against existing audit practice, integrated into reviewer workflows as prioritization rather than replacement, and defended to regulators. The framework is designed for exactly this human-in-the-loop posture: it orders the audit queue and documents the evidence; determinations remain with qualified reviewers.

The framework aligns with the V28 transition in a specific way. Because V28 re-mapped codes and re-estimated coefficients, transition-period risk-score movement is expected and legitimate; a detector trained on trajectory features must therefore be calibrated per payment-year methodology, and the framework's alignment with the phased V24/V28 blend (CMS, 2023, 2024) is a maintenance requirement, not a one-time property. Periodic retraining against current-year model software, monitoring for feature drift, and fairness auditing across demographic subgroups — essential wherever learned models touch payment (Obermeyer et al., 2019) — constitute the operational lifecycle.

Deployment integration follows naturally from where the framework's inputs and outputs sit. Its inputs are the encounter-derived fields MAOs already assemble for risk-adjustment submission — diagnoses, service records, provider identifiers, and prior-year scores — so screening can run as a gate in the existing submission pipeline, scoring each beneficiary-year before the encounter batch is transmitted for payment. Its outputs slot into two established workflows: internally, flagged records with SHAP attributions feed the coding-compliance review queue that MAOs operate for RADV readiness, ordered by risk score so reviewer capacity concentrates on the highest-exposure submissions; externally, the same evidence structure — unsupported diagnoses, trajectory anomalies, provider concentration — mirrors what RADV audits and OIG evaluations examine, meaning a plan deploying the framework is effectively pre-running the government's audit logic against its own submissions before payment rather than defending them years later in extrapolated-recovery proceedings. For CMS and its integrity contractors, the identical machinery inverts into a targeting tool: scoring incoming encounter submissions across contracts to prioritize which plans and provider groups warrant audit attention. In both postures the economics favor prevention — a pre-payment correction costs one internal review, while a post-payment RADV finding carries extrapolated repayment and enforcement exposure.

6. Limitations and Future Work

Four limitations bound these findings. First and most importantly, evaluation uses synthetic data. The simulation is grounded in the CMS-HCC payment mechanics and documented fraud typologies, uses only observable features with realistic noise, and is fully reproducible — but synthetic fraud is cleaner and better labeled than operational fraud, injected prevalence (10.1%) is a design choice likely above operational rates, and the simplified 15-category condition structure abstracts from the full V28 model's 115 categories. Reported metrics characterize behavior in this controlled environment and should be read as an upper bound on, not a forecast of, operational performance. Validation on operational MA encounter data — under appropriate data-use agreements and governance — is the essential next step, with the RADV audit corpus as the natural source of labeled ground truth.

Second, the feature set is encounter-summary level; laboratory, procedure, and pharmacy detail — the clinical depth needed to close the severity-upcoding gap — is not modeled. Third, the simulation is a single-year cross-section with a simulated prior score; longitudinal multi-year modeling would strengthen trajectory features and enable detection of gradual intensity drift. Fourth, adversarial adaptation is not modeled: real coding behavior responds to detection, and sustained performance requires the retraining and drift-monitoring lifecycle described above. Future work should pursue operational validation;

clinical-depth feature integration; longitudinal and provider-network modeling (fraud concentration within organizations suggests graph-based extensions); fairness auditing across beneficiary subgroups; and calibration protocols for the V28 transition years.

The fairness dimension deserves explicit treatment in operational deployment. A fraud detector trained on utilization-conditioned features risks encoding access disparities: beneficiaries with genuinely limited healthcare access — disproportionately dual-eligible, rural, and minority populations — present thinner utilization footprints, and a model that reads sparse service records as weak diagnostic support could concentrate false-positive audit burden on exactly the beneficiaries and safety-net providers least equipped to absorb it. The documented case of a widely deployed healthcare algorithm that encoded racial bias through a cost-based proxy (Obermeyer et al., 2019) is the cautionary template. Operationally, this requires subgroup-stratified evaluation of false-positive and false-negative rates before deployment, monitoring of audit-burden distribution across dual-eligibility status and geography in production, and — structurally — the human-in-the-loop posture this framework adopts: flags prioritize review by qualified auditors who see the full clinical context; they never trigger automated payment denial. In the simulated environment, dual eligibility carried the lowest SHAP importance of all fourteen features (mean $|\text{SHAP}|$ 0.098, versus 3.86 for the top-ranked risk-score jump), indicating the model did not lean on eligibility status as a proxy — but this property must be re-verified, not assumed, on operational data where access disparities are real rather than simulated.

7. Conclusion

This study developed and empirically evaluated a machine learning-driven predictive risk scoring framework for detecting risk-adjustment fraud in U.S. Medicare Advantage, aligned with CMS-HCC Version 28 structures. In a controlled, reproducible simulation of 50,000 beneficiary-year records with three injected fraud typologies grounded in documented federal oversight findings, four comparatively evaluated ensemble models all achieved strong performance — F1-scores of 0.949–0.956 and ROC-AUC of 0.997–0.999 — with no single algorithm dominating: LightGBM led on F1, Random Forest on precision and false-positive rate, CatBoost on recall and regression accuracy. Detection exceeded 97% recall for unsupported HCC submissions and HRA-only diagnosis inflation, with severity upcoding hardest at 92.2%, and SHAP analysis showed the models' dominant evidence — year-over-year risk-score jumps, unsupported diagnoses, and provider coding-intensity concentration — matches the audit logic of the empirical upcoding literature and OIG oversight practice.

The results support three conclusions. First, fabrication-style risk-adjustment fraud is highly detectable at machine speed from encounter data alone, generalizing manual OIG audit methodology to pre-payment screening scale. Second, algorithm choice matters less than feature design and threshold policy; organizations should tune operating points to audit capacity rather than seek a categorically best model. Third, severity upcoding defines the detection frontier and the likely direction of adversarial migration, motivating clinical-depth feature integration. Validated against operational encounter data, the framework offers Medicare Advantage Organizations, auditors, CMS, and policymakers a practical, transparent path from retrospective payment recovery toward proactive payment integrity under the CMS-HCC Version 28 payment model.

References

1. Ash, A. S., Ellis, R. P., Pope, G. C., Ayanian, J. Z., Bates, D. W., Burstin, H., Iezzoni, L. I., MacKay, E., & Yu, W. (2000). Using diagnoses to describe populations and predict costs. *Health Care Financing Review*, 21(3), 7–28.
2. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, 319(13), 1317–1318. <https://doi.org/10.1001/jama.2017.18391>
3. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
4. Brown, J. R., Duggan, M., Kuziemko, I., & Woolston, W. (2014). How does risk selection respond to risk adjustment? New evidence from the Medicare Advantage Program. *American Economic Review*, 104(10), 3335–3364. <https://doi.org/10.1257/aer.104.10.3335>
5. Centers for Medicare & Medicaid Services. (2023). Announcement of calendar year (CY) 2024 Medicare Advantage (MA) capitation rates and Part C and Part D payment policies. U.S. Department of Health and Human Services. <https://www.cms.gov/medicare/payment/medicare-advantage-rates-statistics/announcements-and-documents>

6. Centers for Medicare & Medicaid Services. (2024). Risk adjustment: Medicare Advantage rates and statistics. U.S. Department of Health and Human Services. <https://www.cms.gov/medicare/payment/medicare-advantage-rates-statistics/risk-adjustment>
7. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
8. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
9. Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, 233–240. <https://doi.org/10.1145/1143844.1143874>
10. Ellis, R. P., Pope, G. C., Iezzoni, L. I., Ayanian, J. Z., Bates, D. W., Burstin, H., & Ash, A. S. (1996). Diagnosis-based risk adjustment for Medicare capitation payments. *Health Care Financing Review*, 17(3), 101–128.
11. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29. <https://doi.org/10.1038/s41591-018-0316-z>
12. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
13. Geruso, M., & Layton, T. (2020). Upcoding: Evidence from Medicare on squishy risk adjustment. *Journal of Political Economy*, 128(3), 984–1026. <https://doi.org/10.1086/704756>
14. Gilfillan, R., & Berwick, D. M. (2021). Medicare Advantage, direct contracting, and the Medicare 'money machine,' part 1: The risk-score game. *Health Affairs Forefront*. <https://www.healthaffairs.org/doi/10.1377/forefront.20210927.6239/>
15. Goldstein, B. A., Navar, A. M., Carter, R. E., & Pencina, M. J. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*, 24(1), 198–208. <https://doi.org/10.1093/jamia/ocw042>
16. Kronick, R., & Welch, W. P. (2014). Measuring coding intensity in the Medicare Advantage program. *Medicare & Medicaid Research Review*, 4(2), E1–E19. <https://doi.org/10.5600/mmrr.004.02.a06>
17. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. <https://doi.org/10.48550/arXiv.1705.07874>
18. McWilliams, J. M., Hsu, J., & Newhouse, J. P. (2012). New risk-adjustment system was associated with reduced favorable selection in Medicare Advantage. *Health Affairs*, 31(12), 2630–2640. <https://doi.org/10.1377/hlthaff.2011.1344>
19. Medicare Payment Advisory Commission. (2016). Using encounter data for risk adjustment in Medicare Advantage. In *Report to the Congress: Medicare and the health care delivery system*. <https://www.medpac.gov/>
20. Newhouse, J. P., Price, M., Huang, J., McWilliams, J. M., & Hsu, J. (2012). Steps to reduce favorable risk selection in Medicare Advantage largely succeeded, boding well for health insurance exchanges. *Health Affairs*, 31(12), 2618–2628. <https://doi.org/10.1377/hlthaff.2012.0327>
21. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
22. Office of Inspector General. (2021). Some Medicare Advantage companies leveraged chart reviews and health risk assessments to disproportionately drive payments (OEI-03-17-00474). U.S. Department of Health and Human Services. <https://oig.hhs.gov/oei/reports/OEI-03-17-00474.asp>
23. Pope, G. C., Ellis, R. P., Ash, A. S., Ayanian, J. Z., Bates, D. W., Burstin, H., Iezzoni, L. I., Marcantonio, E., & Wu, B. (2000). Principal inpatient diagnostic cost group model for Medicare risk adjustment. *Health Care Financing Review*, 21(3), 93–118.

24. Pope, G. C., Kautter, J., Ellis, R. P., Ash, A. S., Ayanian, J. Z., Iezzoni, L. I., Ingber, M. J., Levy, J. M., & Robst, J. (2004). Risk adjustment of Medicare capitation payments using the CMS-HCC model. *Health Care Financing Review*, 25(4), 119–141.
25. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358. <https://doi.org/10.1056/NEJMr1814259>
26. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
27. Shekhar, S., Leder-Luis, J., & Akoglu, L. (2023). Unsupervised machine learning for explainable health care fraud detection. *arXiv*. <https://doi.org/10.48550/arXiv.2211.02927>
28. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56. <https://doi.org/10.1038/s41591-018-0300-7>