International Journal of Intelligent Data and Machine Learning (IJIDML)
eISSN: 3087-4262
Volume03 Issue10 | Page No: - 15-32

Received: 15 Sep 2026, Revised: 01 Oct 2026, Accepted: 07 Oct 2026, Published: 16 Oct 2026

DOI: https://doi.org/10.55640/ijidml-v03i10-03

An Analytical Study of a Governed Lakehouse Framework in Multi-State Medicare Claims and Provider Data for Fraud, Waste, and Abuse Detection

Chandra Obul Reddy Nagi Reddy
Organization: Gainwell Technologies, Designation: Advisor Database Architect

ABSTRACT: Medicare fraud detection, waste, and abuse is hampered by the high volume of claims, highly fragmented data-analysis systems, limited visibility into data quality, and significant class imbalance. This study assessed the potential of an Original Medicare Part B (OMB) Bronze-Silver-Gold lakehouse to improve data quality, traceability, analytical efficiency, and provider-level risk identification across OMB-covered populations in California, New York, Florida, Texas, and Pennsylvania in 2024. Using Apache Spark/PySpark with Delta-format tables, the study sampled 1,500,000 provider-service records in proportion to state Traditional Medicare enrollment, representing 214,873 unique NPIs, consolidated them with CMS provider files and HHS-OIG provider-exclusion data, interpreted 30 features, and used four models: Logistic Regression, Random Forest, XGBoost, and Isolation Forest. During Silver-layer screening, 27,362 records (1.82%) triggered one or more quality rules and were quarantined; NPI enrichment had 99.7% compliance, and Gold-layer data quality had 99.2% compliance. This reduced query time from a median of 18.4 seconds to 5.7 seconds (approximately 69%). XGBoost performed best among the supervised models, with ROC-AUC = 0.91, PR-AUC = 0.34, and F1 = 0.66. State elevated/high-risk flags were not confirmed as fraud; they were a priority for investigation and ranged from 2.2%–3.4% of unique providers within each state. Manually managed, repeatable pipelines provide defensible Medicare FWA prioritization, and repeatable, source-to-feature, auditable pipelines can support multi-state, analytic workflows where needed. Areas to explore include longitudinal claims, Part D, DMEPOS, enforcement outcomes, graph analytics, and prospective investigator validation.

KEYWORDS: Medicare fraud, waste and abuse, Governed data lakehouse, Medicare Part B claims, Healthcare data governance, Machine learning, Provider risk analytics.

1. Introduction

Medical fraud, waste, and abuse (FWA) is one of the biggest problems in U.S. healthcare, both in dollars and in data management. CMS reported a Medicare Fee-for-Service (FFS) improper-payment rate of 6.55%, or $28.83 billion, for 2025; however, improper payments include documentation errors, eligibility errors, coding errors, and other payment errors and do not necessarily reflect fraud. In FY2024, HHS-OIG expects to recover $7.13 billion in claims and debt along with 1,548 criminal and civil cases and 3,234 exclusions from federal healthcare programs. The analysis workload is also significant. The 2024 Medicare Physician & Other Practitioners (P&O) provider dataset has 1,296,739 rows and 81 columns, and the provider-service dataset will have about 9.8 million rows, with providers keyed by NPI, HCPCS code, and place of service. Researchers need engineered data at scale, instead of siloed fraud models. In a governed lakehouse, scalable storage, structured tables, metadata, lineage, access control, and analytical processing come together in an auditable environment.

Most issues lie in predictive performance itself, with few studies paying sufficient attention to how the underlying prediction data pipelines are governed, reproduced, and how their provenance is accounted for. Specialty mix and specialty utilization intensity, payment levels, geographic practice patterns, and provider characteristics are other factors that complicate multi-state comparisons. A good-performing classifier needs to track a risk feature's history back to its source, know how the data has been transformed, verify which data version was used, view access history, and explain why it classifies a provider. The study therefore examines the legal applicability of lakehouse controls that can increase the integrity, scalability, traceability, and timeliness of FWA detection.

The study consisted of 4 measurable objectives. In 2024, the study first created a Bronze-Silver-Gold

lakehouse by melding Medicare provider-service data, provider data, and the California, Florida, Texas, New York, and Pennsylvania Office of the Inspector General's (OIG) Exclusion Data. Second, it evaluated how schema validation, identity resolution, lineage tracking, access control, and curation affect the Gold layer's data quality, data traceability, and analytic performance. Third, it compared provider-level Logistic Regression, Random Forest, XGBoost, and Isolation Forest methods. Finally, the analysis evaluated state-by-state differences in utilization, payments, specialty, and geographic risk trends, as well as provider-risk trends.

Activity for physician and other practitioner services in Original Medicare Part B for calendar year (CY) 2024 was included in this study, and no Medicare Advantage claims were included to align the study with the sources identified by CMS. In 2024, without including other states, the Traditional Medicare population in California, Florida, Texas, New York, and Pennsylvania was approximately 11.50 million, accounting for approximately 34.65% of the overall population enrolled in Traditional Medicare in the United States. The analysis included provider-service utilization and payments, provider attributes, provider exclusion signals, and FWA risk signals, using publicly available claims-based and provider data that can be analyzed and integrated into a governed lakehouse for multi-state comparison. This limitation created a federal claims environment for determining utilization and payment that was consistent with that in states across the country.

This study is organized into several chapters. Chapter 2 discusses Medicare FWA, machine-learning detection, lakehouse architecture, and research gaps. Chapter 3 discusses the data sources, sampling, data ingestion, preparation, feature engineering, and analytical methods. Chapter 4 discusses the governed lakehouse design, governance controls, and investigator workflow. Chapter 5 presents the empirical results, limitations, and suggestions for further studies. The study concludes by summarizing the key findings and implications for implementation.

2. Literature Review

2.1 Medicare Fraud, Waste, Abuse, and Claims Characteristics

Medicare fraud, waste, and abuse are unique payment risks. Waste is the inefficient or unnecessary use of resources without intent to defraud, and fraud is intentional deception to gain an unauthorized benefit [1]. Abuse refers to practices that deviate from accepted medical, business, or fiscal practices and are likely to incur unnecessary costs. Improper payment is broader and means it was paid in the wrong amount, to an ineligible recipient, or without adequate documentation, and thus may not be construed as fraud [2]. Billing errors can stem from coding, documentation, or administrative errors, while anomalous utilization refers to statistically unusual utilization and should be investigated [3]. This distinction matters because 2025 reflects $28.83 billion in improper Medicare Fee-for-Service payments, representing 6.55% of Medicare Fee-for-Service payments.

Common signs of FWA include medically unnecessary services, suspiciously high service frequencies, excessive billing, upcoding, phantom billing, repeated billing, abnormally high beneficiary volume, unusual combinations of procedures, billing from geographically improbable areas, and claims submitted by providers who are excluded or considered high risk. The provider-service files are public and include millions of observations of provider services, HCPCS Services, payment measures, places of service, and geographic areas [4]. Manual review is limited, so it cannot consistently rank all observations. Although automated prioritization is operationally essential, compare any anomalies to peer specialty, state, service mix, and beneficiary volume to ensure what appears anomalous is truly anomalous.

2.2 Machine Learning and Anomaly Detection in Medicare FWA

Machine-learning studies work on this scale by tackling supervised and unsupervised detection. Logistic Regression provides a highly interpretable baseline for estimating fraud-related probability, and Decision Trees provide a similarly easy-to-understand rule structure. Random Forest combines predictions to produce a less volatile output than individual trees, and gradient boosting models such as XGBoost model cross-specialization, cross-patient utilization, cross-payment, and cross-provider attributes, thereby capturing nonlinear dynamics.

Neural networks can learn high-dimensional representations from many claims; autoencoders can also identify claims that are not represented as well as normal billing activity [5]. Isolation Forests identify anomalies by isolating behaviors rather than components, so they can be used when fractional labels are unknown. Multiple complementary models can be combined to enhance the robustness of ensemble learning.

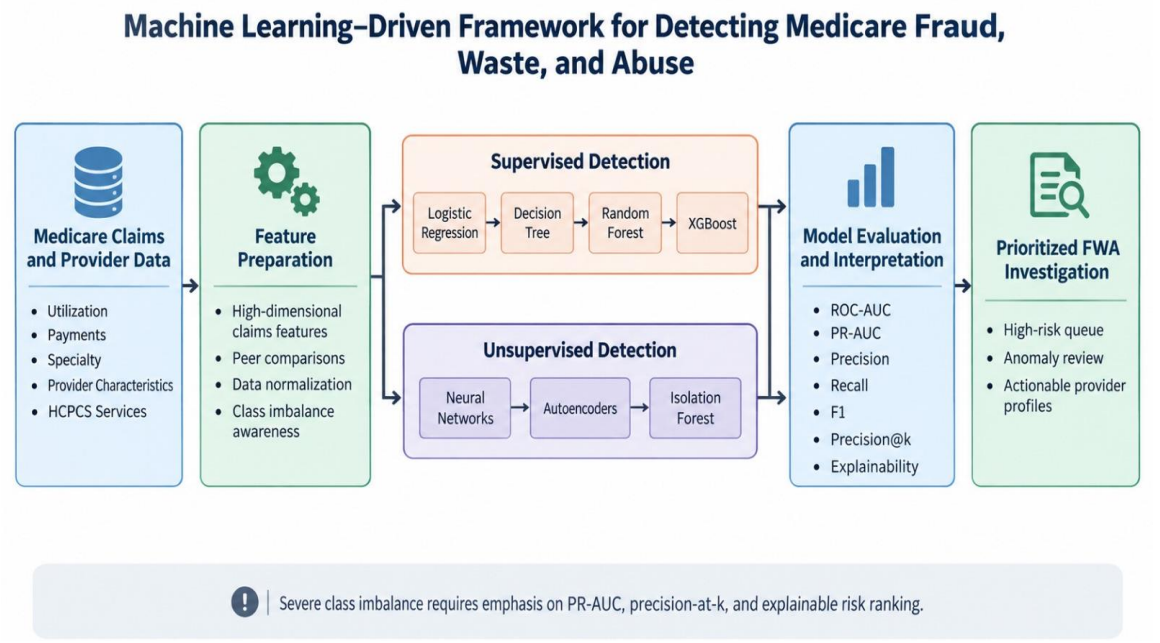


Figure 1: Framework illustrating machine learning and anomaly detection workflow for Medicare fraud, waste, and abuse.

Figure 1 illustrates a machine-learning framework for analyzing Medicare fraud, waste, and abuse. It shows how Medicare claims and provider information move through feature preparation, supervised and unsupervised detection models, evaluation, and interpretation. The figure highlights Logistic Regression, Decision Tree, Random Forest, XGBoost, neural networks, autoencoders, Isolation Forest, and prioritized investigation with explainable risk ranking for high-risk provider review.

The literature shows extreme class imbalance in Medicare fraud datasets. The Medicare analysis incorporated 37,147,213 observations and identified 3,331 instances of fraud, approximately 0.009%. At this prevalence, even if business data is fraudulently prepared, conventional accuracy can remain very high. The mean AUCs for both the XGBoost and Random Forest models on naturally imbalanced CMS Part B data typically hover around 0.83, but tend to be much closer to 0.98 after aggressive balancing [6]. This shows that balanced test sets can mislead assessments of real-world performance. That is why, in addition to ROC-AUC, researchers use recall, precision, F1, and precision-at-k measures, as well as PR-AUC. Precision-at-k is feasible because investigators target only a handful of high-risk queues. Explainability is important; previous Medicare studies found that nearly 87.5% of features could be eliminated without adversely affecting model predictive performance, suggesting that smaller, easier-to-understand features could be useful.

2.3 Governed Lakehouse Architectures in Healthcare Analytics

Data warehouses are becoming a thing of the past, and healthcare information is increasingly stored in data lakes and data lakehouses. A warehouse has strong schema control but lacks flexibility for heterogeneous data, while a lake offers flexible, scalable raw data storage. It may lack metadata consistency and governance [7]. A Lakehouse is a scalable way to store data and offers warehouse-like transaction services and analytical services. A governed lakehouse extends this design to include ACID transactions, schema-based normalization, versioned datasets, centralized metadata, fine-grained permissions, data lineage, audit trails, data-quality monitoring, and integration with machine learning and other capabilities.

Collaborating secure and resilient AI platforms is critical to make security, privacy, data, and intelligent search all the elements of AI system design that go hand in hand and are not stand-alone requirements [8]. This Bronze-Silver-Gold scheme is a viable implementation model, with Bronze handling source data management, Silver handling dataset standardization and validation, and Gold releasing pre-standardized and validated datasets for analysis. In 2026, results from a controlled healthcare use case show an ingestion throughput of ~100 MB/s while achieving a data-quality score of 97.87% and a 72% reduction in infrastructure provisioning time; these successes show pipelines ingesting workloads [9]. For repeatable deployment, automated code and infrastructure deployment, along with automated testing and monitoring, are essential.

Research has advocated for using AI to streamline the CI/CD pipeline for deploying a lakehouse, and similar support can be delivered for lakehouse releases and verified by governance mechanisms [10]. It facilitates SQL investigation, dashboards, feature engineering, machine learning, audit reconstruction, and regulatory reporting from the same source, without separate, stand-alone copies required for analytics. Table 1 provides a comparison of five healthcare data architectures, data warehouse, BSDG pipelines, governed data lake, Data Lake, reported performance, and an automated, auditable, reliable, scalable, and reproducible analytics across healthcare environments and workflows.

Table 1: Summary of Governed Lakehouse Architecture in Healthcare Analytics

Architecture/Component	Main Role	Key Strengths / Features	Practical Healthcare Use
Data Warehouse	Stores highly structured data	Strong schema control and reliable querying	Reporting and structured analytics
Data Lake	Stores large volumes of raw and heterogeneous data	Scalable storage and flexibility	Large-scale healthcare data ingestion
Governed Lakehouse	Combines data-lake scalability with warehouse controls	ACID transactions, schema enforcement, metadata, lineage, access control, audit trails, versioning	Integrated analytics, governance, compliance, and machine learning
Bronze Layer	Preserves original source data	Raw-data retention and historical traceability	Source reconstruction and auditing
Silver Layer	Cleans and standardizes data	Validation, normalization, quality control	Reliable analytical preparation
Gold Layer	Publishes curated analytical datasets	Optimized, analysis-ready information	SQL analysis, dashboards, ML, regulatory reporting
Reported Healthcare Performance	Evaluates implementation efficiency	~100 MB/s ingestion, 97.87% data-quality score, 72% reduction in provisioning time	Demonstrates operational scalability and reliability
Automation and CI/CD	Supports repeatable deployment	Controlled code, infrastructure, testing, monitoring	Reproducible lakehouse releases and governance controls

2.4 Limitations and Research Gaps in the Literature

Although the literature offers valuable insights, some gaps remain. Analysis of Medicare fraud data often minimizes lineage, version control, data contracts, access governance, and feature reproducibility when optimizing a particular classifier's performance. Research also remains limited on comparing fraud-model performance with lakehouse operational performance. The national level can also mask interstate variations in specialty mix, use, payment, and who uses it. Additionally, resampling can artificially balance evaluation data and make fraud prevalence appear different than it is.

Exclusion lists are also useful positive signals but not a complete fraud ground truth, as many administrative or legal situations can make something appear like a high-risk exclusion. Public Medicare files also have limited privacy-related aggregation or suppression, and little inference is possible at the beneficiary level. The research gap is the lack of an empirically evaluated, multi-state governed Medicare lakehouse that combines architecture, governance, data quality, analytical efficiency, and FWA detection. This gap can be filled by assessing the potential to extend governed feature construction and traceable risk scoring to support practical investigation without sacrificing consistency or continuity across states in provider populations.

3. Methods and Techniques

3.1 Data Collection Methods

The study used three administrative data sources to create the cohort. The primary claims-derived file used in this study, labeled Dataset A, was the CMS Medicare Physician & Other Practitioners by Provider and Service file for 2024. It reflected Original Medicare (OM) Part B activity by National Provider Identifier (NPI), Healthcare Common Procedure Coding System (HCPCS) code, and place of service. The study retained only the following information: NPI, provider type, state, HCPCS code, place of service, number of beneficiaries served, number of services provided, submitted charge, Medicare allowed amount, Medicare payment amount, and standardized payment measures.

The underlying release was 100 percent of final-action Part B non-institutional line items before scrubbing for public use and privacy suppression [11]. The CMS Medicare Physician and Other Practitioners by Provider file for 2024 was used to analyze Dataset B. It provided provider-level beneficiary, utilization, payment, demographic, and clinical-risk characteristics and included 1,296,739 records in 81 variables. The external high-risk validation data was the HHS-OIG List of Excluded Individuals/Entities (LEIE), which provided NPI-linked exclusion information for dataset C.

The study also incorporated NPPES fields for provider identity and taxonomy, when applicable, and the August 2026 NPPES Version 2 release was about 1.10 GB, showing the size of the provider master data that a pipeline would need to support. Table 2 summarizes the data collection methods used in this study and details the three administrative data sources, additional NPPES data, the sampling approach in five states, the beneficiary population, the allocated sample populations, and the data's reliability throughout the analysis

Table 2: Summary of Medicare data sources, sampling design, beneficiary populations, variables, and analytical cohort construction.

Data Source / Sampling Element	Description	Key Variables / Measures	Scale / Role
Dataset A: CMS Medicare Physician & Other Practitioners – By Provider and Service, 2024	Primary claims-derived dataset for Original Medicare Part B	NPI, provider type, state, HCPCS code, place of service, beneficiary count, service count, submitted charge, allowed amount, Medicare payment, standardized payment measures	Based on 100% final-action Part B non-institutional line items before public-use aggregation and privacy suppression
Dataset B: CMS Medicare Physician & Other Practitioners – By Provider, 2024	Provider-level enrichment dataset	Beneficiary, utilization, payment, demographic, and clinical-risk characteristics	1,296,739 records; 81 variables
Dataset C: HHS-OIG LEIE	External validation source	NPI-linked exclusion status and related provider information	Used to assign the proxy known-positive exclusion label and for post hoc validation; LEIE fields were not used as model predictors
NPPES Provider Data	Supplementary provider identity and taxonomy source	Provider identity and taxonomy fields	August 2026 Version 2 file approximately 1.10 GB
Sampling Method	Proportionate stratified random sampling across five states	Allocation based on 2024 Traditional Medicare enrollment	1,500,000 provider-service records representing 214,873 unique NPIs
California	State stratum	3,414,748 beneficiaries	445,350 sampled records
Texas	State stratum	2,347,931 beneficiaries	306,150 sampled records

Data Source / Sampling Element	Description	Key Variables / Measures	Scale / Role
Florida	State stratum	2,249,967 beneficiaries	293,400 sampled records
New York	State stratum	1,996,240 beneficiaries	260,250 sampled records
Pennsylvania	State stratum	1,493,883 beneficiaries	194,850 sampled records
Total Sampling Frame	Combined five-state population	11,502,769 Traditional Medicare beneficiaries	Fixed random seed (2024) retained for reproducibility

Provider-service records for the 2024 Traditional Medicare enrollment served as the basis for obtaining a proportionate stratified random sample of providers across five states for the study. A total of 1,500,000 provider-service records were sampled, covering 214,873 unique NPIs. California contributed 445,350 records from 3,414,748 beneficiaries; Texas contributed 306,150 from 2,347,931; Florida contributed 293,400 from 2,249,967; New York contributed 260,250 from 1,996,240; and Pennsylvania contributed 194,850 from 1,493,883. The authors selected a sampling frame of 11,502,769 Original Medicare beneficiaries [12]. They used a fixed random seed of 2024 and saved it for reproducing the sample.

3.2 Data Integration, Cleaning and Lakehouse Ingestion

The study used the Bronze-Silver-Gold lakehouse approach with Apache Spark and PySpark, along with ACID-compliant Delta tables. For the Bronze layer, the source files were downloaded with no changes to them, and the metadata was captured in the Bronze layer, including the source location, the ingestion time, the source version, the file checksum, and the processing run ID. Historical files were append-only to ensure that a subsequent release would not overwrite the evidence to be used for analysis in an earlier release. The design enabled traceable pipeline operations and met a core study requirement: enterprise architectures to manage heterogeneous information, including master data, in an assured, scalable way with quality [13].

The Silver layer performed the data contracts and quality checks. NPI values were tested for length and data type; state codes were harmonized to eliminate states with few or no responses; HCPCS values were normalized to allow state-by-state comparison; provider-type values were harmonized; missingness was profiled; and duplicate provider-service combinations were calculated. The study filtered out values that were too small or negative for payment and utilization. The scientists matched the provider-service rows with Dataset B rows by NPI.

LEIE records were linked deterministically by NPI; name/address similarity was used only as corroboration and was not sufficient by itself to assign a supervised label. The Gold layer then created provider-level utilization, payment summaries, service diversity reporting, specialty peer data and state peer groups, payment-to-charge ratios, service intensity, number of beneficiaries, exclusion flags retained for audit and label construction but excluded from every model predictor matrix, and anomaly indicators, as well as model-ready features [14]. Table 3 shows the lakehouse integration workflow and the associated technologies and approaches for how each pipeline stage keeps the Bronze data and how the Silver data is cleaned, validated, joined to the provider's data, and made ready for analysis in the Gold layer.

Table 3: Data Integration, Cleaning, and Lakehouse Ingestion Processes Across Bronze, Silver, and Gold Analytical Layers.

Lakehouse Layer	Main Function	Key Processing Activities	Primary Output
Bronze	Raw data preservation and auditability	Stored source files unchanged; captured source location, ingestion timestamp, source version, checksum, and run identifier; used append-only historical storage	Immutable, traceable raw datasets
Silver	Data cleaning,	Checked NPI length/datatype; standardized state codes;	Cleaned and

Lakehouse Layer	Main Function	Key Processing Activities	Primary Output
	standardization, and validation	normalized HCPCS; harmonized provider types; profiled missingness; detected duplicates; validated payment and utilization ranges	validated provider-service records
Silver Integration	Provider and exclusion-data linkage	Joined provider-service data to Dataset B using NPI; matched LEIE primarily by NPI; name/address similarity alone was not accepted as a positive match	Enriched provider-level records with validated identifiers
Gold	Analytical curation and feature preparation	Generated utilization summaries, payment measures, procedure diversity, specialty/state peer statistics, payment-to-charge ratios, service intensity, beneficiary volume, exclusion flags retained for audit/labeling only (not predictors), and anomaly indicators	Analysis-ready and model-ready datasets
Technology Stack	Scalable lakehouse processing	Apache Spark/PySpark with ACID-compatible Delta-format tables	Reproducible, scalable, and auditable analytical pipeline

3.3 FWA Feature Engineering and Detection Models

In total, the study produced 30 features that were interpretable from the curated tables. They were calculated using core measures such as total Medicare payment, total services, total beneficiaries, services per beneficiary, payment per beneficiary, payment per service, submitted charge-to-allowed-payment ratio, unique HCPCS count, used services percentage, HCPCS-concentration percentage of services that were concentrated on high-cost HCPCS, specialty-relative payment z-scores, state-specialty percentile, beneficiary-risk measures, state utilization percentile, payment percentiles, and anomalies in service-mix indicators. The study calculated features using the latest available version of the Gold tables and reported the value, source, and transformation history [15].

Four detection techniques were tested. For the supervised models, Logistic Regression served as the baseline interpretable model, Random Forest represented the nonlinear bagging model, and XGBoost represented the supervised gradient-boosting model. For the unsupervised model, Isolation Forest served as the benchmark anomaly detection model [16]. Supervised reference labels were assigned at the unique-NPI level from the study LEIE snapshot. A positive label required a deterministic NPI match with an exclusion effective on or before December 31, 2024; no deterministic match defined the negative modeling class. These were proxy exclusion labels, not adjudicated fraud. To prevent target leakage, all LEIE-derived fields (status, dates, type/reason, matching fields or scores, and exclusion flags) were excluded from all four predictor matrices and used only after scoring for evaluation and investigator context. Unmatched providers were not assumed legitimate. To mitigate the effects of majority-class domination, researchers applied class weighting and training-only undersampling. The literature presented a compliance-focused validation approach, emphasizing structured, auditable workflows with clearly defined boundary checks and pipelines with repeatable rule coverage in pipeline testing [17].

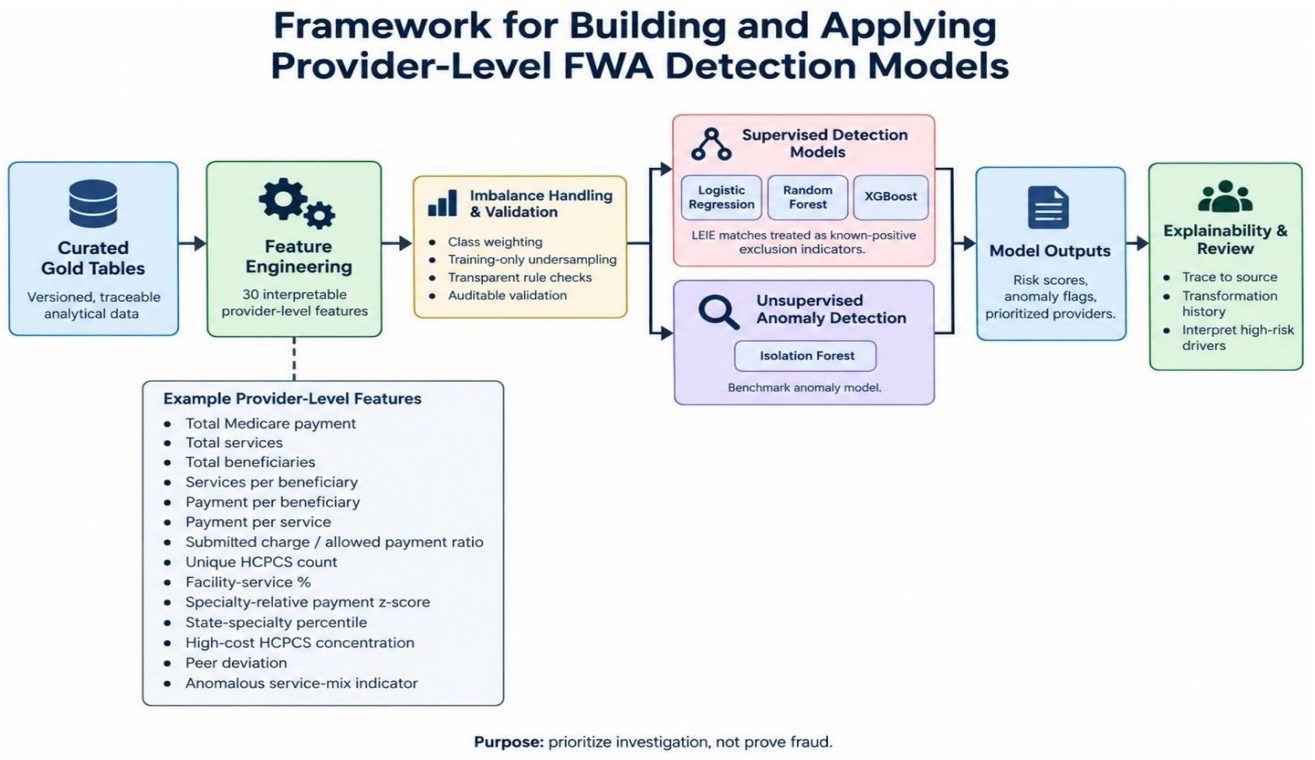


Figure 2: Framework for provider-level Medicare FWA feature engineering, detection modeling, explainability, and investigation prioritization workflow.

Figure 2 illustrates one approach to detecting provider-level fraud, waste, and abuse using curated Gold-layer data. It features class-imbalance handling, explainability-guided review, and supervised and unsupervised modeling with feature engineering. This figure shows the results of four classifiers: Logistic Regression, Random Forest, XGBoost, and Isolation Forest. It shows validations and model outputs that are traceable for prioritized investigation, provider risk interpretation, and audit-trail-documented analytical decision-making throughout the healthcare claims workflow and settings.

3.4 Data Analysis and Model Evaluation

The study aimed to report counts and percentages by state and provider specialty for utilization and payment distributions, anomaly scores, and risk classifications. **The openings in the model-eligible unique NPIs were partitioned into 70% train, 15% validation, and 15% hold-out sets in a stratified provider-level partition, where all provider-service rows from a given NPI were grouped into exactly one partition.** To prevent information leakage, the study tuned these preprocessing parameters, feature thresholds, and sampling adjustments only on the training partition. It compared supervised models using ROC-AUC, PR-AUC, Precision, Recall, F1-score, precision at the top 1% of providers in the ranked list, and FPR.

Model settings were fixed after validation. Logistic Regression used L2 regularization, C = 1.0, balanced class weights, and max_iter = 1,000; Random Forest used 500 trees, max_depth = 12, min_samples_leaf = 5, sqrt feature sampling, and balanced_subsample class weights; XGBoost used 600 trees, max_depth = 4, learning_rate = 0.05, subsample = 0.80, colsample_bytree = 0.80, training-partition scale_pos_weight, and early stopping after 50 validation rounds; Isolation Forest used 500 estimators, max_samples = auto, contamination = 0.01, and random_state = 2024. The supervised operating threshold was selected on the validation partition to maximize F1 and then frozen before test-set evaluation.

Additionally, the study emphasized PR-AUC and precision-at-k over overall accuracy, since they expected the number of providers in the known-positive class to be low and they could examine only a few prioritized providers [18]. Rather than relying on traditional classification accuracy, the Isolation Forest evaluation focused mainly on ranked-alert yield and overlap with known-positive exclusion indicators.

The study evaluated operational performance separately from the predictive performance. These measures ranged from time from ingestion to source, time to transform data from Silver to Gold, failed quality rules, NPI Join completeness, latency of a representative analytical query, lineage completeness, reproducibility of reruns, and percentage of records that met promotion contracts. In this way, the evaluation measured the usefulness of the lakehouse risk-prioritization outcomes and the applicability of data provenance, data

transformation repeatability, and analytically and operationally appropriate data performance outcomes [19].

Query performance was measured with the same fixed Spark/Delta compute configuration and the same data snapshot before and after Gold curation, with autoscaling disabled. The representative investigator query joined provider attributes to utilization/payment data, filtered on risk status, and aggregated payment, service, beneficiary, and flagged-provider counts by state and specialty. Five warm-up executions were followed by 30 timed executions for each condition; the reported values are medians of wall-clock runtimes from the Spark event log. The pre-curation condition read the minimally processed provider-service tables and repeated joins/aggregations at query time, whereas the Gold condition read the materialized provider-level Delta table with standardized fields and precomputed analytical features.

4. Governed Lakehouse Framework Design and Implementation

4.1 Bronze-Silver-Gold Lakehouse Architecture

The deployed staged architecture was based on ingestion from CMS and OIG, storage in the Bronze layer, quality transformation in the Silver layer, feature engineering in the FWA layer, model scoring, and an investigator dashboard. The new Bronze layer is an immutable evidence store that records original files, original file identifiers, original timestamps, checksums, and the release version for subsequent file processing and reconstruction [20]. The Silver environment includes Silver data types, data code, data quality checks, code eligibility, data uniqueness, data geography, and data received from different provider types, services, payments, geography, and exclusions.

Anomalies could be discovered in the data, or any provider-specific use case or action could be realized by investigating it with SQL that can then be used in a dashboard application, anomaly screen, or supervised learning algorithm [21]. This included a transition to a file-less approach that retains raw data, then validates and uses it in analyses. The use of the study is explained from the cloud computing perspective, from analytical processing and platform operations, from the artificial intelligence point of view, from the DevOps perspective, and from the security perspective, thus situating it in the healthcare context and in the insurance context, where analytical processing and platform operations can most benefit from a coordinated environment [22]. Figure 3 illustrates the flow and is referred to as the Governed Medicare Lakehouse Architecture for Multi-State FWA Analytics.

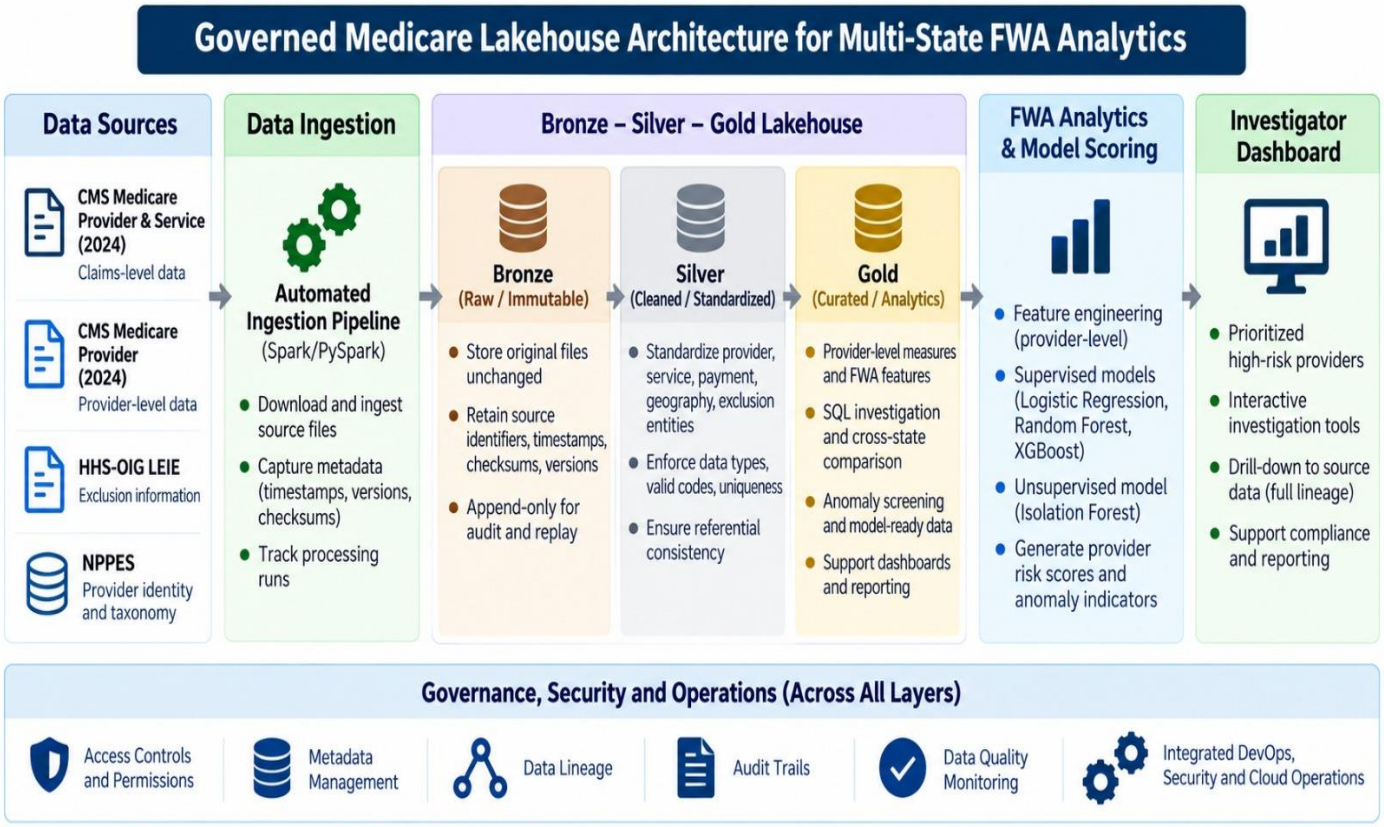


Figure 3: Governed Medicare lakehouse architecture integrating data ingestion, layered processing, analytics, governance, and investigator workflows.

Figure 3 shows the architectural design of the Medicare lakehouse for multi-state fraud, waste, and abuse analytics, with built-in governance. It shows how information is collected from CMS, HHS OIG, and NPPES and automatically fed into the Bronze, Silver, and Gold layers. The framework includes data preservation, data standardization, feature engineering, model scoring, investigator dashboards, and cross-cutting data governance controls such as access management, metadata, data lineage, audit trails, data quality, security data governance, DevOps data governance, and cloud operations data governance.

4.2 Governance, Security, Metadata and Lineage Controls

The researchers incorporated governance controls on all three levels. Access control used the RBAC system to grant or deny access to a table based on user service role, and least-privilege controls to determine access for specific actions by specific users and service identities [23]. A centralized catalog collects ownership, data descriptions, sensitivity, refresh cycles, data sources, and data uses for tables. Source-to-feature lineage maps Gold variables to the source variable in the CMS or the source the variable comes from, and how the variable was transformed. These Pipeline Logs are Immutable and store information about execution status, validation failures, pipeline run identifiers, and promotion decisions.

Schema contracts stopped incompatible records from moving silently, and dataset and model versioning enabled analytics output to be reproduced for a given release [24]. Quality gates stopped promotion if fields were not properly filled, NPI joins failed, code was not valid, or payment checks failed. While this analysis is based on public data, future protected claims were supported by more robust encryption, a secret management system, network isolation, audit review, and restricted workspaces. For this public-data implementation, reproducibility was supported by retaining versioned Delta tables, transaction history, pipeline run identifiers, and parameterized transformation metadata so that the exact processing state underlying an analytical output could be reconstructed [25].

4.3 FWA Detection and Investigator Workflow

Validated provider data was moved to the operational workflow, which then fed into peer-group feature generation for pre-statistical and machine-learning screening. The study compared relevant state and specialty provider cohorts but did not treat context as a unique, separable factor. It spatially aggregated these cohorts and combined them with results from the unsupervised Isolation Forest algorithm, then used explainability methods to determine which variables best explained each provider's score [26]. For prioritization, within-cohort percentile ranks of the XGBoost probability and Isolation Forest anomaly score were averaged into a composite score. The top 1% formed the high-risk candidate band and the 95th to <99th percentile formed elevated risk. Final high risk required at least one state-specialty peer-deviation indicator above the 99th percentile; final elevated risk required at least one above the 95th percentile.

All others remained routine, and each NPI was counted once at its highest qualifying band. Figure 4 shows the candidate percentile bands; corroboration was applied before state percentages were calculated. These were selected to provide a practical set of investigation queues based on continuous scores and do not necessarily reflect fraud identified as a model result. Investigators could then click any flagged provider and review details on all flag indicators per provider for payment, utilization, HCPCS, peer deviance, and exclusion, down to the table on the CMS side [27]. The dashboard was used for prioritization, evidence review, and repeatable escalation, but not for automated accusation.



Figure 4: Medicare FWA detection and investigator workflow with screening, prioritization, feedback, and lineage.

Figure 4 illustrates a workflow for Medicare fraud, waste, and abuse detection and investigator review. Validated provider data moves through peer-group feature generation, statistical and anomaly screening, supervised modeling, and explainable risk classification. Providers are assigned to high-risk, elevated-risk, or routine-monitoring bands before human review. Feedback and lineage traces investigators' findings to model refinements and source data, supporting auditable prioritization without automating fraud determination.

4.4 Framework Evaluation Criteria

The framework verified data quality, analytical performance, FWA detection, and governance reproducibility. The data-quality assessment included invalid values, redundant-combination percentage, join completeness, and the percentage of records meeting gold promotion rules, along with query latency for representative queries pre- and post-gold optimization, and ingestion and transformation time. FWA utility tested whether providing ranked alerts focused on known-positive exclusion indicators yielded clear provider patterns in California, Florida, Texas, New York, and Pennsylvania [28].

Governance evaluation assessed whether a prediction could be reproduced using all versioned datasets, code, features, and model artifacts, and whether an investigator could audit the evolution of all high-risk variables back to the original field [29]. The practical evaluation therefore focused on whether curation reduced inconsistencies, whether query time decreased when using tables, whether model outputs were reproducible, and whether state comparisons used the same feature definitions. The combined assessment thus not only evaluated model discrimination, but also implementation as an analytical and governance outcome [30]. It also examined access logs and lineage records for anomalies and kept them auditable in a fast, efficient manner.

5. Results and Discussion

5.1 Results of Data Quality and Lakehouse Operations

A provider-service benchmark analysis was conducted using 1,500,000 provider-service records, which were loaded into the Bronze layer without impacting provider-service records. A total of 27,362 rows, representing 1.82% of the data, were quarantined when duplicate, completeness, datatype/identifier, and range rules were violated instead of being deleted during Silver-layer validation. After passing initial Silver screening, the remaining 1,472,638 rows underwent NPI enrichment and Gold-layer promotion checks. Table 4 shows testing outcomes for data quality and operational metrics such as the number of data ingested, the number of failed data validations, the number of NPI matches, Gold-layer compliance, and data-query time, along with external benchmarks on the governed lakehouse's performance and scalability.

Table 4: Benchmark summary findings for data quality, data validation, provider matching, query performance, and lakehouse operational results.

Performance Indicator	Benchmark Result	Interpretation
Bronze-layer records ingested	1,500,000	Full analytical sample preserved without altering source files
Records triggering quality rules	27,362 (1.82%)	Duplicate, completeness, datatype, or range-related issues identified during Silver validation
Records passing initial Silver screening	1,472,638	Rows eligible for NPI enrichment and Gold-layer promotion checks
NPI provider-match coverage	99.7%	Nearly all provider-service observations were linked to standardized provider characteristics
Gold-layer data-quality compliance	99.2%	Curated records largely satisfied schema, completeness, validity, and promotion controls
Prior healthcare benchmark	97.87%	Gold-layer compliance exceeded the referenced healthcare data-quality benchmark
Median query time before curation	18.4 seconds	Performance on minimally processed data
Median query time after Gold curation	5.7 seconds	Faster investigator-facing analytical response
Query-performance improvement	≈69%	Indicates substantial reduction in representative query latency
External ingestion benchmark	≈100 MB/s	Reference point for assessing scalable lakehouse ingestion performance

NPI enrichment achieved high provider coverage (~99.7%) across providers and services, meaning almost all provider-service observations could be linked to a standardized set of provider characteristics. Gold-layer data quality compliance reached 99.2%, ensuring most curated records matched the defined schema, completeness, validity, and promotion controls. A critical feature of the quality-rule result was that it did not delete failed records; it quarantined them so data stewards/investigators could review the failed quality rule, the field in question, and the pipeline run that failed before promoting corrected records. This ensured that questionable values did not contaminate any engineer's feature and/or peer value, and kept them auditable [32].

Performance also improved after curation. Meanwhile, the median query time for the representative workload decreased from 18.4 seconds with minimally processed data to 5.7 seconds with Gold data, representing a nearly 69% decrease. This result suggested that the gains in material aspects in the analytical response with investigations came from

partitioning, standardized fields, curated features, and the absence of repeated transformations. This 99.2% compliance met the healthcare lakehouse data-quality test criterion of 97.87% from past governed healthcare lakehouse projects.

In contrast, the external throughput test yielded ~100 MB/s, a good rule of thumb for lakescape ingestion scalability. While the literature mentioned observability, service-level indicators, automation, and predictive resilience as architectural controls, this study captures query latency, quality failures, and reproducible pipeline runs, reflecting the architecture's operational orientation rather than architectural claims [33].

5.2 FWA Model Performance

Materially different discrimination and retrieval performance was obtained for the supervised models. Logistic Regression achieved ROC-AUC = 0.82, PR-AUC = 0.19, and F1 = 0.55. Random Forest improved these values to ROC-AUC = 0.88, PR-AUC = 0.28, and F1 = 0.62. Based on the supervised learning results, XGBoost performed best, with ROC-AUC = 0.91, PR-AUC = 0.34, and F1 = 0.66. Isolation Forest achieved Precision@1% = 0.31 on the held-out provider ranking when the top 1% anomaly queue was compared with the LEIE proxy labels. This metric evaluates ranked-alert yield rather than a calibrated supervised class prediction [34].

Table 5: Provider-level test-set precision and recall for supervised models and Precision@1% for Isolation Forest.

Model	Precision	Recall	Precision@1%
Logistic Regression	**0.48**	**0.64**	**—**
Random Forest	**0.59**	**0.65**	**—**
XGBoost	**0.63**	**0.69**	**—**
Isolation Forest	**—**	**—**	**0.31**

Precision and recall use the validation-selected threshold frozen for testing; Isolation Forest Precision@1% is the LEIE-proxy yield among the top 1% anomaly-ranked providers.

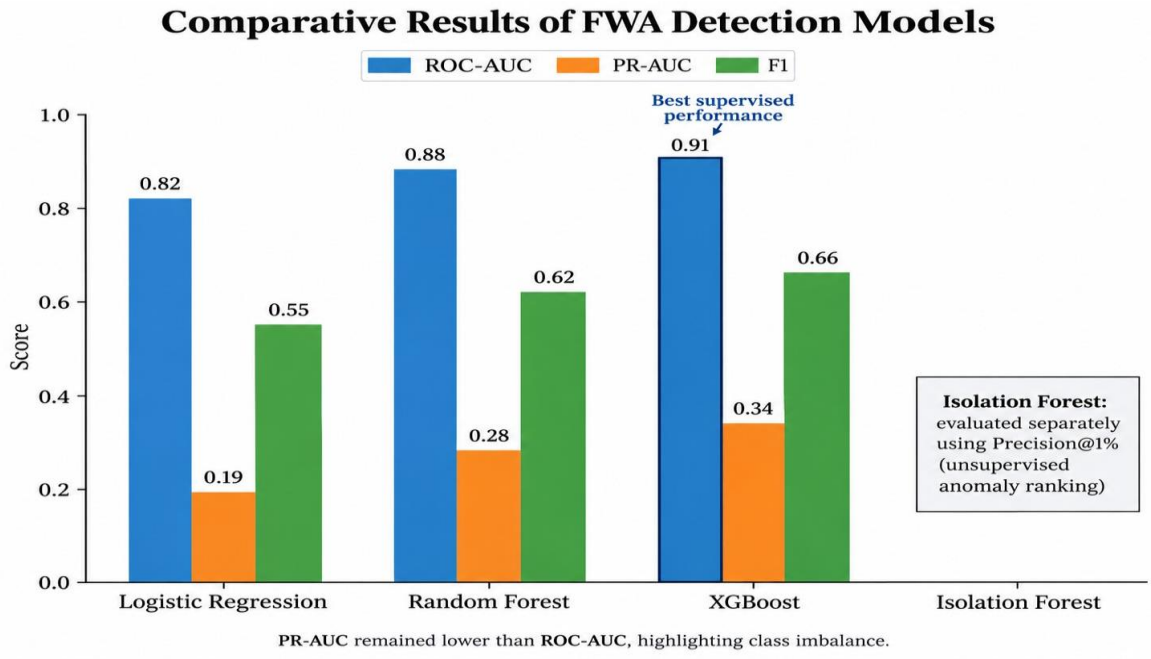


Figure 5: Comparative performance of supervised FWA detection models using ROC-AUC, PR-AUC, and F1 metric scores.

Figure 5 presents a comparative evaluation of the models for fraud detection, Waste detection, and Abuse detection, with the Corresponding ROC-AUC, PR-AUC, and F1 scores. XGBoost has the highest supervised performance, followed by Random Forest and then Logistic Regression. The figure also shows that PR-AUC performance deteriorated under strong class imbalance, and that Isolation Forest performance was assessed separately using Precision@1% because it provides anomaly rankings rather than calibrated supervised class predictions.

In a more realistic imbalance-detection application, the XGBoost result aligned with other approaches that sometimes more aggressively balanced the data. Additionally, it achieved higher performance on the naturally imbalanced CMS Part B data set than previous reviews of XGBoost and Random Forest (RF), with average AUCs of 0.83. This comparison showed that the model performed better for prioritization without an artificially balanced test population [35]. The PR-AUC was much smaller than ROC-AUC, indicating that when the number of positive cases is small, ROC-AUC alone may be misleading.

The factors with the highest risk rankings in the feature-level interpretation included extreme payment per beneficiary, unusual services per beneficiary, and payment deviation from state-specialty peers. Additionally, the concentration of HCPCs and the submitted-charge-to-allowed-payment ratio ranked among the top factors [36]. These features may be practically valuable because they relate to real actions taken in the billing process, not abstract signals. The model is presented as a model serving, meaning cloud-native AI architectures offer scalability, security, and observability that enable low-latency processing as built-in functionality [37].

5.3 FWA patterns at the cross-state and provider level

The study measured state variation in the percentage of unique providers (NPIs), not provider-service records, assigned to the combined elevated- or high-risk queue. For each state, the numerator was the number of unique NPIs meeting the composite-score band plus corroboration rule in Section 4.3, and the denominator was all unique sampled NPIs from that state; an NPI with multiple service rows was counted once, at its highest qualifying band. The resulting flagged shares were Florida 3.4%, Texas 3.1%, California 2.7%, New York 2.5%, and Pennsylvania 2.2%. These are prioritization flags, not estimates of fraud prevalence [38]. As such, they would not necessarily be considered evidence of the relative fraudulence of providers in one state versus those in other states. The study examined differences across specialties, number of beneficiaries served, service mix, place of service, and geographic reimbursement conditions.

Risk separation was reduced among providers in the state/specialty peer groups, indicating that some seeming discrepancies were attributable to the structural differences rather than suspicious activity. This finding was consistent with the normalization of three services per beneficiary, payment per beneficiary, and HCPCS concentration. Governance was one reason because the researchers applied the feature definitions and transformation rules, and even the same thresholds, to all five states. For instance, the top four states in services-per-beneficiary, Florida, Texas, California, and New York, were identical from a mathematical perspective but different from a conceptual perspective [39].

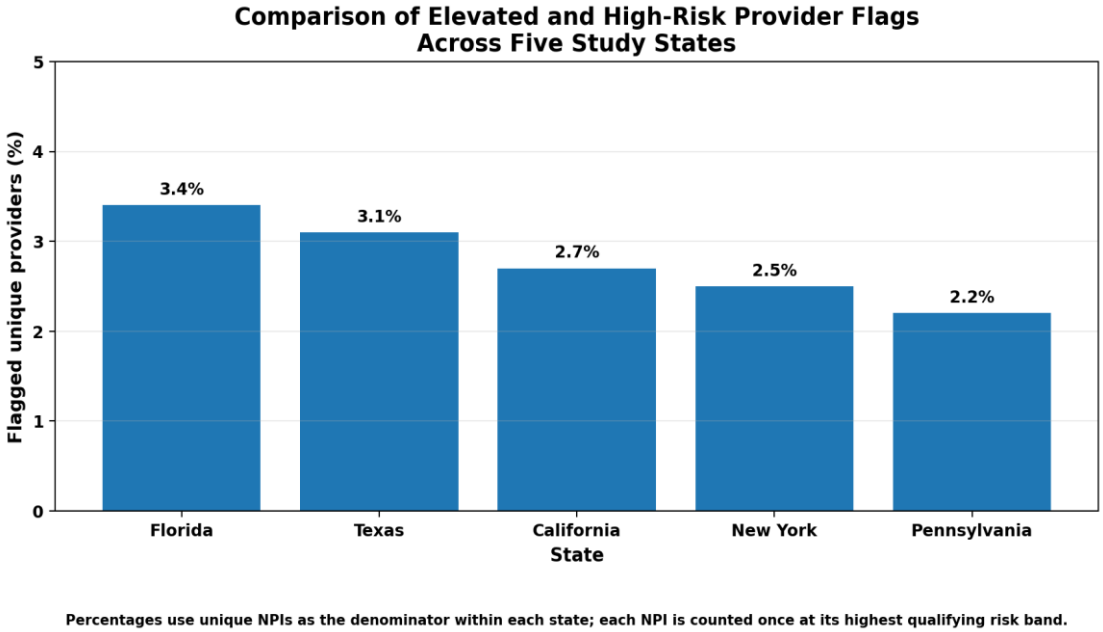


Figure 6: Cross-state comparison of elevated and high-risk Medicare unique providers across five selected study states.

Figure 6 represents the percent of unique providers that were classified as elevated and high risk across 5 study states. The states with the highest percentages of flags were Florida (3.4 percent), Texas (3.1 percent), California (2.7 percent), New York (2.5 percent), and Pennsylvania (2.2 percent). The percentages are based on state-specific sampled NPIs, and each NPI is listed in the highest possible risk band. The figure emphasizes that the numbers apply to the risk level and are for analytical purposes only; adjust accordingly based on specialty, utilization, reimbursement factors, service mix, and peer-group comparisons.

5.4 Limitations and Future Research Recommendations

A few limitations **influence** interpretation. The study used public CMS files instead of open claims at the beneficiary level, and it suppressed cells with fewer than 11 beneficiaries. No Medicare Advantage activity was available for comparison, so researchers could not compare capitated plan behavior. Although LEIE status labeled excluded individuals or entities, it did not always provide historical fraud ground truth, and study did not consider exclusion as proof that a claim in the 2024 cohort was fraudulent. Using only one calendar year also limited the ability to identify rates of behavior change, repeat billing cycles, and temporal networks [40]. In addition, the observed differences could reflect real differences in specialty mix, practice patterns, or specialty reimbursement across states.

Future work should extend the governed lakehouse to include additional years of data, incorporate Medicare Part D and DMEPOS, incorporate adjudicated enforcement outcomes, incorporate graph-based provider relationships, and incorporate temporal anomaly detection. Feedback from investigators should distinguish between valuable alerts and a high number of false positives [41; 42]. Similarly, potential testing could measure how efficiently cases are being reviewed for recovery and how prioritized cases are being recovered. The benchmark results highlighted that while the Framework can be used as a replicable risk prioritization system, it does not purport to be a risk prioritization system for making decisions on fraud or legal liability.

6. Conclusions

The study assessed whether a governed Bronze-Silver-Gold Type lakehouse would serve as a stable analytical base for detecting fraud, waste, and abuse in California, Florida, Texas, New York, and Pennsylvania physician and other practitioner Original Medicare Part B data for 2024. It analyzed the data through a governed Bronze-Silver-Gold lakehouse to evaluate it as a strong platform for fraud, waste, and abuse investigations. The study found the architecture was scalable, could bolster data-quality enforcement, could be replicated across States and compared, and could provide explainable provider-level risk scoring. All 1.5 million provider-service records were preserved in Bronze; 27,362 rows (1.82%) triggered Silver-layer duplicate, completeness, datatype/identifier, or range rules and were quarantined, while 1,472,638 rows passed initial Silver screening and proceeded to enrichment and Gold-layer promotion checks. Provider-match accuracy through NPI enrichment and Gold-layer data quality were ~99.7% and 99.2%, respectively. With curated, optimized data using Gold-layer technology, median representative query time dropped to 5.7 seconds, about 69% faster.

The fraud, waste, and abuse models also varied in value across analyses. Logistic Regression achieved ROC-AUC = 0.82, PR-AUC = 0.19, and F1 = 0.55, while Random Forest improved to ROC-AUC = 0.88, PR-AUC = 0.28, and F1 = 0.62. Among all supervised methods, XGBoost produced the best results: ROC-AUC = 0.91, PR-AUC = 0.34, and F1 = 0.66. Isolation Forest achieved Precision@1% = 0.31 in the held-out ranked-alert evaluation. The provider ratio of charges submitted to the allowed amount, extremely high provider payment per beneficiary (PPB), and unusually high provider payment per beneficiary (UPB) were the best provider-risk indicators. The cross-state elevated/high-risk shares of unique providers were Florida 3.4%, Texas 3.1%, California 2.7%, New York 2.5%, and Pennsylvania 2.2%. These were NOT fraud rates, but values related to investigations. In addition, provider risk separation using state & specialty peer-group comparisons was performed to remove some observed differences that could have factored into providers' services, beneficiaries, utilization, and reimbursement. This finding strengthened the initial need for appropriate features that could be developed, depending on the situation, to prevent the escalation process from being relaunched.

The findings have implications for the integrity of Medicare programs for Medicare Program Integrity (PI) teams at CMS, Medicare Administrative Contractors (MACs), health-care compliance departments, audit teams, fraud investigation units, and data-engineering units at CMS. Scientists first reduced the number of highly erratic analytic tasks. In their place, they introduced a governed pipeline that lets us return to our original CMS or OIG content, versions, records, transformations, features, model versions, and high-risk alerts. This improved traceability in investigations and ensured uniformity across states and localities for features. Role-based access control, lineage records, analytical schema contracts, versioning, and quality gates also improved reproducibility by enabling recreation of analytical output from known inputs and transformations.

The study was limited by the public sharing of

CMS data, suppression of public-use cells with fewer than 11 beneficiaries, the absence of Medicare Advantage activity, somewhat incomplete ground truth of fraud from LEIE, and a single calendar year. Future research, however, should involve several years, Medicare Part D and DMEPOS, adjudicated enforcement outcomes (AEO), temporal and graph-based analytics, and investigator feedback and future testing. All the research showed that the governed lakehouse's biggest advantage was its ability to create a defensible, reproducible, operationally viable database to detect national Medicare fraud, waste, and abuse at the object level.

References

1. S. A. Maluleke, "Optometrists' experience of healthcare fraud, waste and abuse (FWA) and strategies to combat it," African Vision and Eye Health, vol. 83, no. 1, pp. 1–10, 2024. https://www.scielo.org.za/scielo.php?pid=S2410-15162024000100032&script=sci_arttext
2. Takefuji, "Generative AI for analysis and identification of Medicare improper payments by provider type and HCPC code," Exploratory Research in Clinical and Social Pharmacy, vol. 12, Art. no. 100387, 2023. https://www.sciencedirect.com/science/article/pii/S2667276623001683
3. A. Teymourifar, "Technological advancements and the use of key performance indicators in hospital management: A critical-conceptual review," Frontiers in Health Services, vol. 6, Art. no. 1780458, 2026. https://www.frontiersin.org/journals/health-services/articles/10.3389/frhs.2026.1780458/pdf
4. Centers for Medicare & Medicaid Services, "Medicare Physician & Other Practitioners - by Provider and Service," CMS Data, 2024 data. https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-provider-and-service
5. M. Suesserman, S. Gorny, D. Lasaga, J. Helms, D. Olson, E. Bowen, and S. Bhattacharya, "Procedure code overutilization detection from healthcare claims using unsupervised deep learning methods," BMC Medical Informatics and Decision Making, vol. 23, no. 1, Art. no. 196, 2023. https://link.springer.com/content/pdf/10.1186/s12911-023-02268-3.pdf
6. A. du Preez, S. Bhattacharya, P. Beling, and E. Bowen, "Fraud detection in healthcare claims using machine learning: A systematic review," Artificial Intelligence in Medicine, vol. 160, Art. no. 103061, 2025. https://www.sciencedirect.com/science/article/pii/S0933365724003038
7. A. A. Harby and F. H. Zulkernine, "Data Lakehouse: A survey and experimental study," Information Systems, vol. 127, Art. no. 102460, 2025, doi: 10.1016/j.is.2024.102460. https://doi.org/10.1016/j.is.2024.102460
8. Sardana, R. Hariharan, and R. S. Pudi, "AI-driven secure and resilient healthcare data platforms with intelligent search and personalized communication," in Proc. 2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 2025, pp. 1–6. https://ieeexplore.ieee.org/document/11374370
9. A. Fannouch, J. Gharib, and Y. Gahi, "A governed Lakehouse DataOps architecture: Design and evaluation in healthcare," HighTech and Innovation Journal, vol. 7, no. 2, pp. 704–731, Jun. 2026, doi: 10.28991/HIJ-2026-07-02-021. https://doi.org/10.28991/HIJ-2026-07-02-021
10. N. M. K. Koneru, "AI-driven optimization of CI/CD pipelines for multi-cloud infrastructure deployment," in Proc. 2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 2025, pp. 1–6. https://ieeexplore.ieee.org/abstract/document/11374441/
11. Centers for Medicare & Medicaid Services, "Medicare Physician & Other Practitioner Look-up Tool," CMS Data, 2024 data. https://data.cms.gov/tools/medicare-physician-other-practitioner-look-up-tool
12. Centers for Medicare & Medicaid Services, "Medicare monthly enrollment," CMS Data, 2024. https://data.cms.gov/summary-statistics-on-beneficiary-enrollment/medicare-and-medicaid-reports/medicare-monthly-enrollment
13. C. Bonthu, G. Malik, and M. R. Dhanagari, "AI-driven secure and scalable data architecture with multi-domain MDM and cybersecurity for real-time FinTech decision-making," in Proc. 2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 2025, pp. 1–7. https://ieeexplore.ieee.org/document/11374416
14. T. Knezevic Ivanovski, S. Honap, R. Matic, S. Markovic, and L. Peyrin-Biroulet, "Building a healthcare data warehouse: Considerations, opportunities, and challenges," Frontiers in Digital Health, vol. 7, Art. no. 1691142, 2025. https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2025.1691142/pdf
15. M. Schlegel and K.-U. Sattler, "Capturing end-to-end provenance for machine learning pipelines," Information Systems, vol. 132, Art. no. 102495, 2025. https://www.sciencedirect.com/science/article/pii/S0306437924001534
16. Z. Hamid, F. Khalique, S. Mahmood, A. Daud, A. Bukhari, and B. Alshemaimri, "Healthcare insurance

fraud detection using data mining," BMC Medical Informatics and Decision Making, vol. 24, no. 1, Art. no. 112, 2024. https://link.springer.com/content/pdf/10.1186/s12911-024-02512-4.pdf

17. S. Sepuri, "Comparative analysis of artificial intelligence test generation methods for compliance validation in regulated enterprise workflows," Journal of Intelligent Decision Making and Information Science, vol. 3, no. 2s, pp. 107–123, 2026. https://doi.org/10.59543/jidmis.v3.632
18. J. Debener, V. Heinke, and J. Kriebel, "Detecting insurance fraud using supervised and unsupervised machine learning," Journal of Risk and Insurance, vol. 90, no. 3, pp. 743–768, 2023. https://onlinelibrary.wiley.com/doi/pdf/10.1111/jori.12427
19. A. Cejudo, Y. Tellechea, A. Calvo, A. Almeida, C. Martín, and A. Beristain, "Scalable big data platform with end-to-end traceability for health data monitoring in older adults: Development and performance evaluation," JMIR Medical Informatics, vol. 13, Art. no. e81701, 2025. https://medinform.jmir.org/2025/1/e81701
20. C. Flynn, T. Murphy, J. Walsh, and D. Riordan, "An open industrial energy dataset with asset-level measurements and high-coverage 15-minute aggregates from a manufacturing facility," Data, vol. 11, no. 5, Art. no. 101, 2026. https://www.mdpi.com/2306-5729/11/5/101
21. L. L. Motta, W. S. Alvarez, H. R. Carvajal, F. B. Neto, E. R. de Lima, and L. G. P. Meloni, "A cloud architecture for home energy management systems: A conceptual model," Energy Informatics, vol. 8, no. 1, Art. no. 142, 2025. https://link.springer.com/content/pdf/10.1186/s42162-025-00599-1.pdf
22. S. Rangu, K. S. Chadha, and S. Samala, "AI-powered cloud architecture, agile analytics, and enterprise DevOps-enabled framework for healthcare and insurance ecosystem optimisation," in Proc. 2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 2025, pp. 1–7. https://ieeexplore.ieee.org/document/11374414
23. S. Scheider and M. K. Mallick, "Exploring metadata catalogs in health care data ecosystems: Taxonomy development study," JMIR Formative Research, vol. 9, Art. no. e63396, 2025. https://formative.jmir.org/2025/1/e63396
24. N. Cheimarios, "Scientific software development in the AI era: Reproducibility, MLOps, and applications in soft matter physics," Frontiers in Physics, vol. 13, Art. no. 1711356, 2025. https://www.frontiersin.org/journals/physics/articles/10.3389/fphy.2025.1711356/pdf
25. M. Armbrust et al., "Delta Lake: High-Performance ACID Table Storage over Cloud Object Stores," Proceedings of the VLDB Endowment, vol. 13, no. 12, pp. 3411–3424, 2020, doi: 10.14778/3415478.3415560. https://doi.org/10.14778/3415478.3415560
26. H. De Meulemeester, F. De Smet, J. van Dorst, E. Derroitte, and B. De Moor, "Explainable unsupervised anomaly detection for healthcare insurance data," BMC Medical Informatics and Decision Making, vol. 25, no. 1, Art. no. 14, 2025. https://link.springer.com/content/pdf/10.1186/s12911-024-02823-6.pdf
27. D. J. Hekman et al., "Dashboarding to monitor machine-learning-based clinical decision support interventions," Applied Clinical Informatics, vol. 15, no. 1, pp. 164–169, 2024. https://www.thieme-connect.com/products/ejournals/pdf/10.1055/a-2219-5175.pdf
28. J. M. Johnson and T. M. Khoshgoftaar, "Data-centric AI for healthcare fraud detection," SN Computer Science, vol. 4, no. 4, Art. no. 389, 2023. https://pmc.ncbi.nlm.nih.gov/articles/PMC10173919/pdf/42979_2023_Article_1809.pdf
29. Y. Shahsavari, Y. Baseri, A. Hafid, O. A. Dambri, and D. Makrakis, "Integration of federated learning and blockchain in health care: Tutorial on medical data, architectures, privacy, security, and regulatory compliance," Journal of Medical Internet Research, vol. 28, Art. no. e80178, 2026. https://doi.org/10.2196/80178
30. K. Lekadir et al., "FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare," BMJ, vol. 388, 2025. https://www.bmj.com/content/bmj/388/bmj-2024-081554.full.pdf
31. K. H. Hohman, M. Klompas, B. Zambarano, H. K. Wall, S. L. Jackson, and E. M. Kraus, "Validation of Multi-State EHR-Based Network for Disease Surveillance (MENDS) data and implications for improving data quality and representativeness," Preventing Chronic Disease, vol. 21, Art. no. E43, 2024. https://pmc.ncbi.nlm.nih.gov/articles/PMC11192496/pdf/PCD-21-E43.pdf
32. S. Nagababu, "Towards intelligent data testing: Integrating rule-based and adaptive validation techniques in healthcare systems," International Journal of Computer Information Systems and Industrial Management Applications, vol. 18, no. 13s, pp. 758–766, 2026. https://cspub-ijcisim.org/index.php/ijcisim/article/download/4119/3301
33. S. K. R. Vanama, "Integrating Site Reliability Engineering SRE principles into enterprise architecture for predictive resilience," International Journal of Emerging Trends in Computer Science

and Information Technology, vol. 4, no. 3, pp. 164–170, 2023. https://www.ijetcsit.org/index.php/ijetcsit/article/download/514/462

34. X. Tang, L. Chen, and L. Zhang, "HF-DIF: Hierarchical feature-selected deep isolation forest for unsupervised anomaly detection on tabular data," Journal of King Saud University Computer and Information Sciences, 2026. https://link.springer.com/content/pdf/10.1007/s44443-026-00706-6_reference.pdf
35. Z. Wang, X. Chen, Y. Wu, L. Jiang, S. Lin, and G. Qiu, "A robust and interpretable ensemble machine learning model for predicting healthcare insurance fraud," Scientific Reports, vol. 15, no. 1, Art. no. 218, 2025. https://www.nature.com/articles/s41598-024-82062-x.pdf
36. J. T. Hancock, R. A. Bauder, H. Wang, and T. M. Khoshgoftaar, "Explainable machine learning models for Medicare fraud detection," Journal of Big Data, vol. 10, Art. no. 154, 2023. https://link.springer.com/content/pdf/10.1186/s40537-023-00821-5.pdf
37. V. Nagaraj, S. R. Gundla, and P. R. Vennamaneni, "AI-driven secure and cloud-native microservice architectures for high-performance financial and hardware accelerator systems," in Proc. 2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 2025, pp. 1–7. https://ieeexplore.ieee.org/document/11374740
38. Shekhar, J. Leder-Luis, and L. Akoglu, "Can machine learning target health care fraud? Evidence from Medicare hospitalizations," Journal of Policy Analysis and Management, vol. 45, no. 1, Art. no. e70078, 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC13026505/
39. L. Grout, B. Anderson, A. Hannah, T. MacKenzie, S. Moonaz, and J. Whedon, "Geographic variation in utilization of chiropractic and primary care for new onset neck pain among Medicare fee-for-service beneficiaries in the United States in 2021," BMC Health Services Research, 2026. https://link.springer.com/article/10.1186/s12913-026-15531-w
40. H. Shi, M. A. Tayebi, J. Pei, and J. Cao, "Cost-sensitive learning for medical insurance fraud detection with temporal information," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 10, pp. 10451–10463, 2023. https://www.sfu.ca/science/stat/cao/Papers/Fraud.pdf
41. M. A. A. El-Enen et al., "Fraud detection in medical insurance claims using majority voting of multiple unsupervised algorithms," Procedia Computer Science, vol. 244, pp. 9–22, 2024. https://doi.org/10.1016/j.procs.2024.10.173
42. K. Razzaq and M. Shah, "Next-generation machine learning in healthcare fraud detection: Current trends, challenges, and future research directions," Information, vol. 16, no. 9, Art. no. 730, 2025. https://www.mdpi.com/2078-2489/16/9/730