Medicare Fraud Detection
A Data-Driven Strategy

Authors –

Chandana Yadav Datti

Pragati Burada

Radhika Srinivasan Iyengar

Mihira Teja Narvaneni

Contents

1. Executive Summary

Medicare, the federal health insurance program serving over 60 million Americans, is a cornerstone of healthcare delivery for seniors and individuals with disabilities. However, the system faces significant financial leakage due to fraud, waste, and abuse—particularly within Medicare Part D, which covers prescription drugs. In 2024 alone, improper payments were estimated at over $50 billion, with Part D contributing a disproportionate share. Common issues such as fraudulent claims, over-prescription, and phantom billing point to a legacy system lacking proactive oversight.

This report proposes a data-driven transformation for the Centers for Medicare & Medicaid Services (CMS). Leveraging publicly available Medicare datasets, we present a strategy that integrates descriptive analytics, predictive modeling, and interactive dashboards to detect and flag suspicious activity. The proposed solution includes a fraud scoring system, machine learning-based classification, and regional mapping tools to help CMS identify and act on anomalies early. This approach aligns with CMS's goals of improving accountability, transparency, and the overall efficiency of healthcare delivery.

2. Strategy

CMS's strategy is grounded in building a culture of proactive fraud prevention through advanced data analytics. The long-term vision is to evolve into a fully data-driven agency capable of screening, detecting, and responding to fraud signals in near real time. Our mission is to support this evolution by integrating machine learning models, interactive dashboards, and structured use cases that drive clarity and enable actionable decisions.

Strategic priorities include improving fraud detection accuracy, automating risk-based screening, reducing claim review timelines, and optimizing resource allocation across geographic regions and provider categories. Decision-making is simplified through real-time dashboards that present alerts, trends, and high-risk indicators.

This strategy fosters accountability by mapping fraud risk flags to provider IDs, enabling justified audits of repeat offenders. It also reinforces CMS's commitment to transparency and ethical oversight by publishing fraud insights with internal stakeholders and beneficiaries.

3. Competitive Landscape Analysis

Medicare is administered by CMS and provides baseline coverage to tens of millions of Americans. However, private organizations such as UnitedHealth Group, Humana, and CVS Health (via Aetna) offer Medicare Advantage (Part C) and Part D plans.

These private insurers manage substantial claim volumes and rely heavily on digital platforms, giving them operational flexibility, and speed. CMS's centralized structure and traditional oversight methods present a gap in proactive fraud detection. Unlike private competitors, which use advanced analytics to control costs and reduce risk exposure, CMS often relies on retrospective audits. This results in higher response times and inefficiencies in identifying fraud trends. By implementing a similar fraud-scoring mechanism and dashboarding tools, CMS can regain competitive ground, reduce reliance on audits, and enhance integrity in public spending.

Key Competitors and Differentiators

Major private insurers such as UnitedHealth Group, CVS Health (via Aetna), and Humana offer Medicare Advantage plans that compete with CMS's original Medicare by providing more personalized and benefit-rich coverage. These competitors differentiate themselves through:

- Customized plans tailored to specific healthcare needs
- Additional benefits such as dental, vision, and fitness programs
- Better integration with retail pharmacy chains and local healthcare networks
- Lower out-of-pocket costs and flexible pricing models

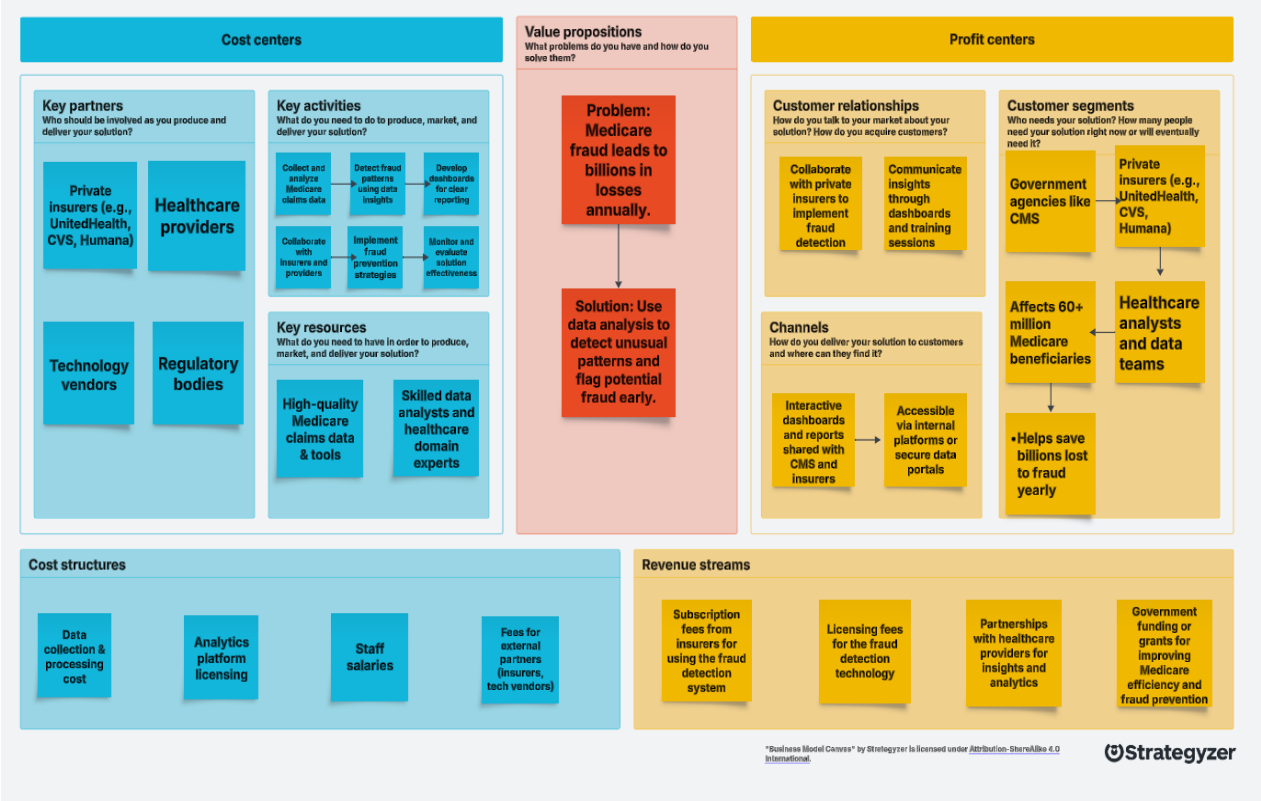
CMS, by comparison, maintains a one-size-fits-all model with limited personalization, which positions these private players as competitive threats for Medicare enrollment.

4. Balanced scorecard

The balanced scorecard for this initiative outlines strategic goals and metrics across four key perspectives: Financial, Customer, Internal Processes, and Learning & Growth. Each area is aligned with the overarching goal of transforming CMS into a data-driven agency capable of detecting and preventing Medicare fraud effectively.

- **Financial:** The primary objective is to reduce fraud-related financial losses. With enhanced detection tools and automation, CMS aims to recover or prevent over $350 million in improper payments annually. Improved audit targeting also enhances cost efficiency, reducing investigation costs per case.
- **Customer:** Transparency and trust are vital in public healthcare. Dashboards help ensure claims are evaluated equitably, while clear flagging and feedback mechanisms give both providers and beneficiaries confidence in the system.
- **Internal Processes:** Automation accelerates fraud detection by flagging claims in real time. Risk scoring, rule-based triggers, and reason codes streamline audit workflows, reducing manual review timelines and increasing screening coverage.
- **Learning & Growth:** CMS is fostering a data-literate culture by training staff in fraud analytics and promoting dashboard usage. Regular workshops and insight tracking help ensure the system evolves through feedback and supports continuous improvement.

5. Business Model Canvas



- **Key Partners** - CMS, state audit contractors, fraud analytics vendors, ML infrastructure providers.
- **Key Activities** - Data ingestion from claim repositories, machine learning model training and deployment, dashboard-based reporting, incorporating audit cycle feedback into the system
- **Key Resources-** Cleaned and labeled Medicare claim dataset,Python-based machine learning tools, Power BI dashboards, CMS internal audit workforce
- **Value Propositions** - Increased detection of fraudulent activity, reduced manual investigation workload, faster response to provider misconduct, Improved transparency for internal and external stakeholders
- **Customer Relationships** - Investigators receive risk-prioritized case queues, providers receive audit feedback with justification, CMS management receives dashboard-based fraud insights
- **Channels**- power BI internal dashboards, audit management systems, CMS data warehouses
- **Customer Segments**- CMS fraud investigators, audit planning staff, external policy analysts and reviewers

- **Cost Structure-** Investment in analytics infrastructure, model retraining and maintenance, staff training and upskilling, dashboard development and maintenance
- **Revenue Streams-** Cost savings from reduced improper payments, efficiency gains through reallocated audit resources

6. Recommended OKRs

Objective 1: Improve Fraud Detection Accuracy

1. **Key Result 1**: Increase fraud model precision from 75% to 90% within 6 months by incorporating updated feature engineering techniques and retraining cycles.
2. **Key Result 2**: Reduce the false-positive rate from 12% to below 5% to minimize unnecessary audits and improve provider trust in the system.

Objective 2: Reduce Improper Payments

1. **Key Result 1:** Lower total estimated fraud costs in Medicare Part D by $350 million through early detection and intervention tools.
2. **Key Result 2:** Flag 95% of high-risk claims before payment disbursement using machine learning prediction pipelines.

Objective 3: Enable Real-Time Screening

1. **Key Result 1:** Reduce average time to detect suspicious claims from 7 days to under 24 hours with real-time flagging integration.
2. **Key Result 2:** Implement automated screening for 100% of incoming Medicare Part D claims by integrating ML model outputs into the CMS pipeline.

Objective 4: Promote a Data-Driven Culture

1. **Key Result 1**: Train 100% of CMS fraud investigation staff on the new analytics platform within the next 3 months.
2. **Key Result 2**: Use insights from dashboards to update 10 fraud detection policies or processes quarterly.

7. Role Summary & Team Chart

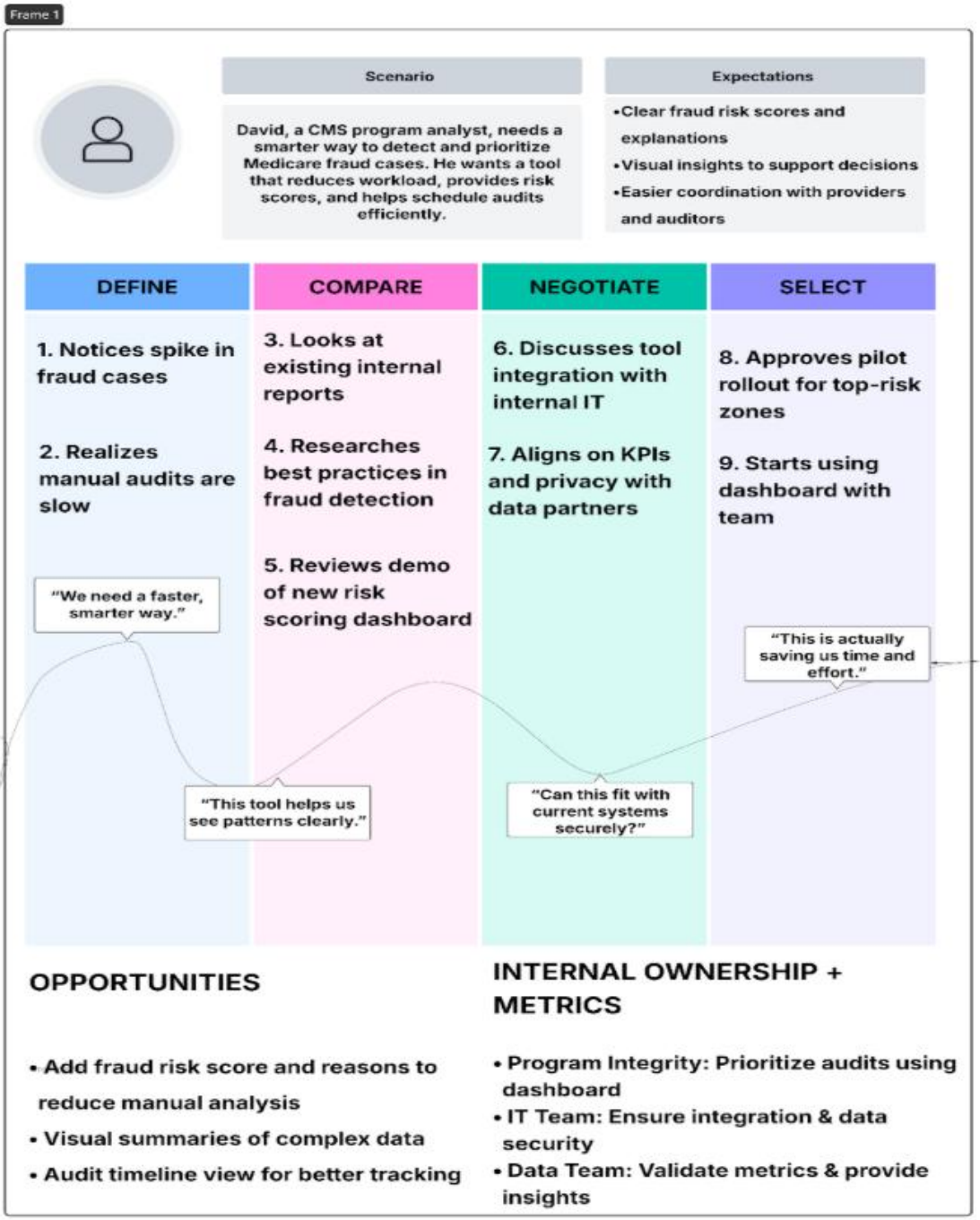
- **Project Manager**
  Oversees overall delivery timelines, ensures milestone completion, manages team coordination, and serves as liaison with CMS leadership.
- **Data Analyst**
  Cleans and analyzes Medicare claim datasets, identifies patterns and outliers, and supports metric development for dashboards.
- **Data Scientist**
  Designs and implements the machine learning fraud detection model, including feature selection, model validation, and retraining.
- **Technical Lead**
  Manages the integration of data pipelines, model outputs, and visualization tools like Power BI to ensure end-to-end solution performance.
- **Domain Expert (Healthcare/Medicare SME)**
  Provides contextual expertise on claim codes, Medicare billing, and common fraud scenarios to guide model development and dashboard logic.
- **Quality Assurance Lead**
  Ensures data accuracy, validates model outputs, and verifies dashboard reliability through ongoing QA processes.
- **Policy Advisor**
  Translates analytics insights into actionable audit policies and supports transparency reporting to external stakeholders.

8. User Personas

To better understand the needs and frustrations of those involved in fraud detection, we created the following personas. These users represent three key stakeholders in the CMS fraud investigation process: a data analyst, a fraud suspect profile, and a provider or billing entity.

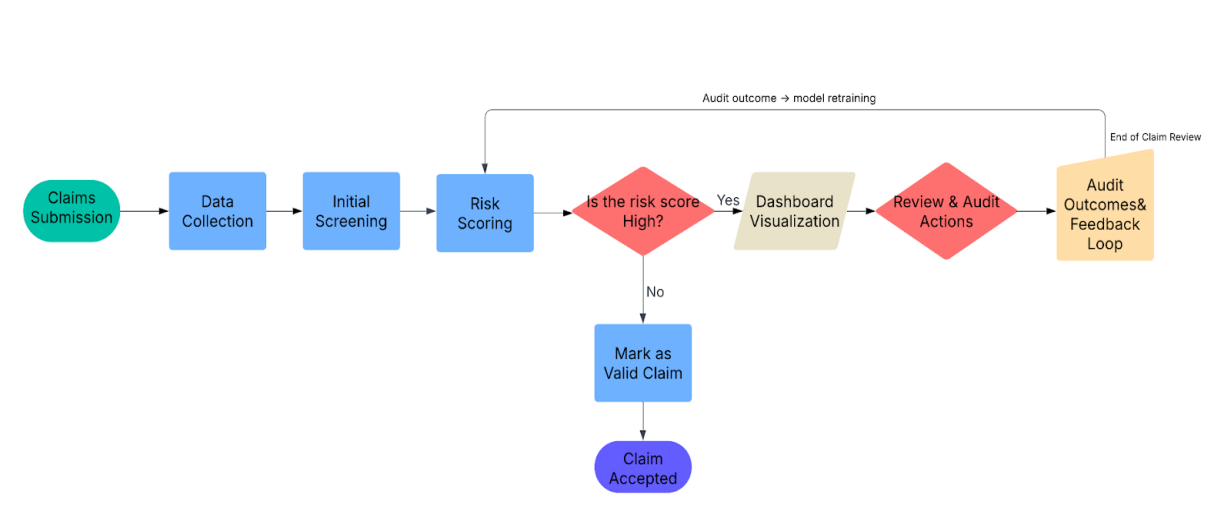
CMS Data Analyst	CMS Fraud Investigator	Healthcare/Billing Provider
Role: Supports model development, dashboard insights, and data validation.	**Role:** Appears frequently in flagged Medicare claims due to abnormal billing or procedure trends.	**Role:** Submits claims to Medicare for reimbursement.
Pain Points: Fragmented tools, delayed data, limited automation.	**Pain Points**: Potential detection risk tries to exploit blind spots in claims validation.	**Pain Points:** Risk of being incorrectly flagged; long appeal/review cycle.
Needs: Fast access to clean, ready-to-analyze datasets; an automated fraud flagging system.	**Needs**: (From CMS perspective) Enhanced surveillance and claim pattern monitoring.	**Needs:** Transparent flagging process, fair review tools, feedback loop with CMS reviewers.

9. Journey Map – CMS Fraud Investigation Workflow



This customer journey map shows how CMS can improve the process of identifying and auditing potential Medicare fraud cases. It follows the experience of the CMS Program Integrity Team as they move through key stages—from receiving flagged claims to selecting providers for audit. The map highlights current pain points and introduces opportunities to make the process faster, clearer, and more efficient using data-driven tools like a risk scoring dashboard and audit timelines.

10.Process flow diagram



1. **Claims Submission :**
   Providers submit Medicare Part D prescription claims to CMS.
2. **Data Collection :**
   CMS stores all submitted claims in a centralized database.
3. **Initial Screening :**
   The system runs basic checks to ensure data completeness and correct formatting.
4. **Risk Scoring :**
   A machine learning model analyzes each claim and assigns a fraud risk score with reasoning.
5. **Is the Risk Score High?**
   The system checks if the fraud risk score crosses a defined threshold.

- **If No:**

   i. **Mark as Valid Claim:**
   Low-risk claims are automatically marked as valid.

   ii. **Claim Accepted:**
   These claims exit the fraud detection process.

- **If Yes:**

    i. **Dashboard Visualization:**

        1. High-risk claims are flagged and displayed in an interactive dashboard.

        2. **Review & Audit Actions:**
           Investigators review flagged claims, prioritize them for audit, and notify providers if selected.

        3. **Audit Outcomes & Feedback Loop:**
           The results of the audit (valid, error, or fraud) are recorded and used to retrain the fraud model for continuous improvement

11. Detailed Use Cases

Based on our analysis and solution design, we recommend the following three use cases to strengthen CMS's fraud detection efforts-

Use Case 1:High-Risk Provider Identification

Context: CMS receives thousands of claims daily, and some providers consistently submit higher-cost claims for routine procedures.
Suggestion: Apply anomaly detection and percentile-based thresholds to identify providers billing significantly above the norm for common services.
Outcome: CMS investigators can proactively audit outliers, reducing financial exposure and deterring high-volume fraud.
Stakeholders: CMS investigators, auditors, policy analysts.

Use Case 2: Regional Fraud Hotspot Mapping

Context: Fraudulent activities often cluster in specific ZIP codes or provider groups, especially where CMS oversight is limited.
Suggestion: Use Power BI to map flagged claims by region and visualize fraud density over time.
Outcome: Helps CMS allocate audit resources efficiently and detect emerging fraud trends geographically.
Stakeholders: CMS regional offices, fraud enforcement units.

Use Case 3: Patient Demographic Fraud Analysis

Context: Fraudulent claims might be skewed toward certain age or gender groups.

Suggestion: Use demographic segmentation to identify vulnerable populations and abnormal fraud rates.

Outcome: Supports CMS in targeting fraud awareness or stricter verification in high-risk age/gender clusters.

Stakeholders: Public health policymakers, fraud detection teams, CMS analysts.

11. Analytics Use Case

The core analytics use case in this project focuses on detecting potentially fraudulent Medicare claims using structured Medicare Part D datasets. The publicly available data from Kaggle includes simulated claim records across inpatient, outpatient, and beneficiary levels, allowing for pattern analysis across providers, services, and regions.

Our initial data preparation involved cleaning and merging the datasets using common identifiers such as provider_id and patient_id. Key features include claim amount, diagnosis categories, provider specialty, chronic conditions, and demographic details. We performed data profiling to assess outliers, missing values, and frequency distributions, and applied basic transformations such as scaling numerical fields and encoding categorical variables.

Descriptive analytics helped uncover initial trends, including high-risk regions, provider billing patterns, and anomalies in patient demographics. The cleaned dataset supports use cases like high-risk provider identification and regional fraud hotspot mapping by enabling sorting, filtering, and aggregation of suspicious activity.

While we did not train a predictive model as part of this project, the dataset has been structured to support future classification efforts. Potential extensions include applying supervised learning to flag suspicious claims and generate fraud risk scores. These scores could be accompanied by model interpretability techniques (e.g., SHAP values) to help investigators understand the reasoning behind each flag.

The prepared dataset now serves as the foundation for interactive Power BI dashboards that enable fraud analysts to explore trends, filter by risk indicators, and prioritize audits based on geographic or provider-level insights.

The Dataset Can be Accessed Using the Following Link-

https://www.kaggle.com/datasets/jaiswalmagic1/healthcare-fraud-detection-dataset

The cleaned and pre-processed datasets used in this project are included in the submission as separate .csv files:

1. claims_cleaned.csv
2. providers_cleaned.csv
3. patients_cleaned.csv
4. payments_cleaned.csv

Each file was created by applying the following cleaning and transformation steps:

- Handled missing values using removal or imputation (depending on field criticality)
- Standardized column names and data formats (e.g., dates, categorical labels)
- Joined datasets on shared identifiers such as claim_id, provider_id, and patient_id
- Created derived fields such as Age Group, Claim_Year, and risk flags for dashboard analysis

12. Dimensional Model

To support our analytics and dashboarding use cases, we structured our data using a star schema. The central fact table captures claims-level data, while supporting dimension tables describe providers, patients, and payments. Below is a breakdown of tables, measures, and hierarchies used.

Fact Table:

1. Claims_Fact

This table captures all Medicare Part D claims with fraud-related attributes.

Column	Role	Description
claim_id	Primary Key	Unique claim identifier
patient_id	Foreign Key	Links to patient demographic info
provider_id	Foreign Key	Links to provider data
claim_date	Date	Date of claim submission
claim_amount	Measure	Total amount billed
status	Dimension	Approved / Rejected
FraudRiskStatus	Label	Model output: Fraud / Not Fraud / Error
Claim_Year	Dimension	Extracted year (for time analysis)

Dimension Tables:

1. Provider_Dim: Can be used to group claims by specialty, location, etc.

Column	Description
provider_id	Unique provider identifier (PK)
name	Provider or organization name
specialty	Medical specialty (e.g., GP, Ortho)
city	Provider city
state	Provider state
zip_code	Zip code
Phone Numbers	Contact information

2. **Patients_Dim** - Used to analyze fraud patterns by demographics

Column	Description
patient_id	Unique patient ID (PK)
first_name	Patient first name
last_name	Patient last name
age	Numeric age
gender	Gender (Male/Female)
state	Patient's state of residence
Age Group	Derived column: 18–30, 31–50, 51+

3. **Payments_Dim** - Enables comparison of claimed vs paid amounts, payment status trends.

Column	Description
payment_id	Unique payment ID (PK)
claim_id	FK to Claims_Fact
payment_date	Date of payment
payment_amount	Actual paid amount

Measures (Built from Claims_Fact and Payments)

Measure	Based On
Total Claim Amount	SUM(claim_amount)
Average Claim Amount	AVG(claim_amount)
Total Payment Amount	SUM(payment_amount)
Fraud Rate	% of claims labeled Fraud
High-Risk Providers	COUNT of providers with high fraud claims
Claims by State or Specialty	COUNT, grouped by dimension

Dimensions

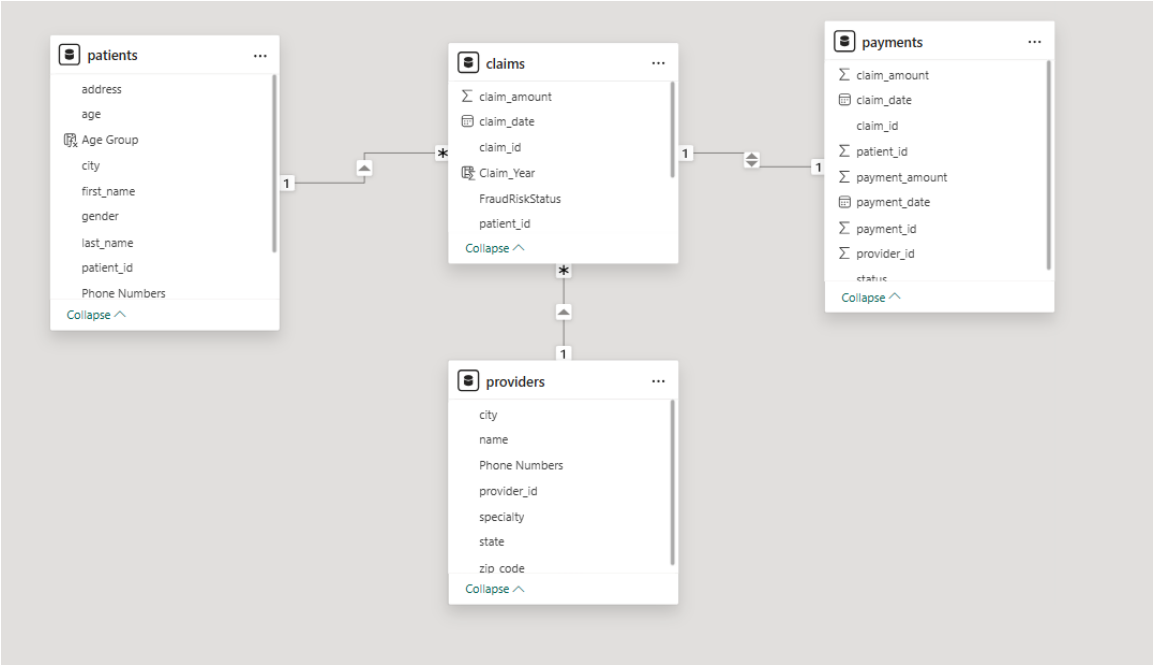
Dimension	Based On
Time	claim_date, Claim_Year
Location	state in Providers and Patients

Provider Specialty	specialty in Providers
Demographics	Age, Gender, Age Group in Patients
Payment Status	Derived from claim vs payment data

Hierarchies

Hierarchy Name	Path
Time Hierarchy	Claim_Year → Quarter (derived) → Month (if needed)
Geography Hierarchy	State → City

The model looks like -



The model supports slicing and dicing claims across time, geography, provider specialty, and patient conditions, enabling CMS to drill into high-risk areas with precision.

13. Governance Model & Glossary

- **Data Dictionary –**

Table	Column	Friendly Name	Description	Tags
Claims	claim_id	Claim ID	Unique identifier for each Medicare Part D claim	Primary Key, Identifier
	patient_id	Patient ID	Link to patient receiving the service	Foreign Key, Identifier
	provider_id	Provider ID	Link to provider who submitted the claim	Foreign Key, Identifier
	claim_date	Claim Date	Date when the claim was submitted	Date, Time Dimension
	claim_amount	Claim Amount	Total billed amount on the claim	Measure, Financial
	status	Claim Status	Approval status (Approved, Rejected, Pending)	Dimension, Categorical
	FraudRiskStatus	Fraud Risk Status	Model-determined risk label: Fraud / Not Fraud	Target, Prediction, Categorical
	Claim_Year	Claim Year	Extracted year from claim date	Time Dimension, Derived
Patients	patient_id	Patient ID	Unique ID for the patient	Primary Key, Identifier
	age	Patient Age	Age of the patient	Demographic, Numerical
	gender	Gender	Gender of the patient	Demographic, Categorical
	Age Group	Age Group	Derived age category (e.g., 18–30, 31–50)	Derived, Segment
	state	Patient State	State of residence	Geography, Location
Providers	provider_id	Provider ID	Unique ID for the provider	Primary Key, Identifier
	specialty	Medical Specialty	Area of medical practice (e.g., GP, Cardiology)	Dimension, Risk Attribute
	state	Provider State	Provider's practice state	Geography, Location
Payments	payment_id	Payment ID	Unique ID for the payment transaction	Primary Key, Identifier
	claim_id	Claim ID	Link to associated claim	Foreign Key, Identifier
	payment_amount	Payment Amount	Actual paid amount	Measure, Financial
	payment_date	Payment Date	Date payment was issued	Date, Financial

Governance ensures that all datasets, fields, and reporting outputs are well-documented, trusted, and traceable. Our data dictionary covers four primary tables: Claims, Providers, Patients, and Payments. Each column is defined with friendly names, source context, and descriptive metadata to support clear understanding.

Tags have been applied across major subject areas, including Provider_Info, Patient_Info, Claim_Stats, and Risk_Metrics. In addition to tagging, each data element includes source lineage and contextual purpose within the fraud detection process. This structured governance supports consistent data usage and enhances transparency in model outputs and dashboard reporting.

- **Business Glossary-**

Term	Definition	Tagged In
Claim ID	A unique identifier assigned to each Medicare claim	Claims, Payments
Fraud Risk Status	Predicted classification of claim risk using analytical techniques	Claims
Claim Amount	Total dollar value submitted by a provider for a patient's treatment	Claims
Payment Amount	Dollar amount paid by CMS to the provider for a claim	Payments
Provider ID	Unique identifier for healthcare providers submitting claims	Claims, Providers
Patient Demographics	Age, gender, and location of patients used for fraud segmentation	Patients
Specialty	Medical discipline or area of provider focus	Providers
High-Risk Provider	A provider identified through rules or outlier detection as likely fraudulent	Claims, Providers
Fraud Score	Numeric score indicating the probability that a claim is fraudulent	Claims (Derived field, optional)
Claim Status	Administrative state of a claim: approved, pending, or denied	Claims
Audit Trigger	Rule or model-based reason a claim was flagged for investigation	Claims (optional future attribute)

The business glossary defines key terms used across the Medicare fraud detection project to ensure consistency, clarity, and shared understanding among stakeholders. It includes core concepts such as Claim ID, Fraud Risk Status, High-Risk Provider, and Payment Amount, along with standardized definitions and applicable business rules. Each term is

linked to relevant datasets and reporting fields through tagging, enabling traceability and proper interpretation in analytics, dashboards, and model outputs. This glossary plays a vital role in maintaining data integrity and supporting communication between technical teams, investigators, and policy analysts.

14. AI Use Case – Fraud Prediction Model

As part of this project, we designed a predictive analytics use case focused on identifying potentially fraudulent Medicare Part D claims. The goal of the model is to assign a fraud risk score to each claim based on historical patterns in provider behavior, billing activity, and patient profiles. This score could then be used to prioritize investigations and flag suspicious claims before payment is disbursed.

The model would use publicly available, structured Medicare datasets (from Kaggle) as its input, including claims, patient demographics, and provider details. We prepared and engineered over 20 potential features such as claim amount, diagnosis frequency, chronic condition flags, prescription category, and provider specialty. The target column (label) would be a field like PotentialFraud, identifying whether a historical claim was found to be fraudulent.

While we did not build or deploy the model in this project, the dataset has been cleaned and formatted to support future use of supervised learning algorithms like logistic regression or XGBoost. Model outputs is displayed in a Power BI dashboard, enabling CMS investigators to filter claims by fraud risk score, provider type, or geographic region.

Feature Columns: claim_amount, diagnosis_count, chronic_conditions, drug_category, provider_type, geographic_region
Label Column: PotentialFraud (binary: Yes/No)

Business Benefit:
This model would help CMS reduce manual audit workloads, increase the ROI of fraud investigations, and intercept fraudulent claims earlier—ultimately protecting Medicare funds and public trust.

15. Ethical Challenges & Mitigation

Ethical considerations in fraud detection are essential due to the potential impact on provider reputations, patient trust, and systemic fairness. Below are five key ethical challenges we identified, along with corresponding mitigation strategies:

1. False Positives

Risk: Incorrectly flagging legitimate providers may cause reputational harm or trigger unnecessary audits.
Mitigation: Implement human-in-the-loop review processes and allow for provider appeal mechanisms.

2. Bias Against Regions or Provider Types

Risk: Models may unintentionally overflag providers in certain locations or specialties.
Mitigation: Conduct fairness audits regularly and normalize for region/specialty-based disparities

3. Opacity in Model Decisions

Risk: Black-box predictions may reduce transparency and accountability in CMS actions.
Mitigation: Integrate explainability tools (e.g., SHAP) to provide per-claim reasoning and generate reason codes.

4. Privacy Concerns

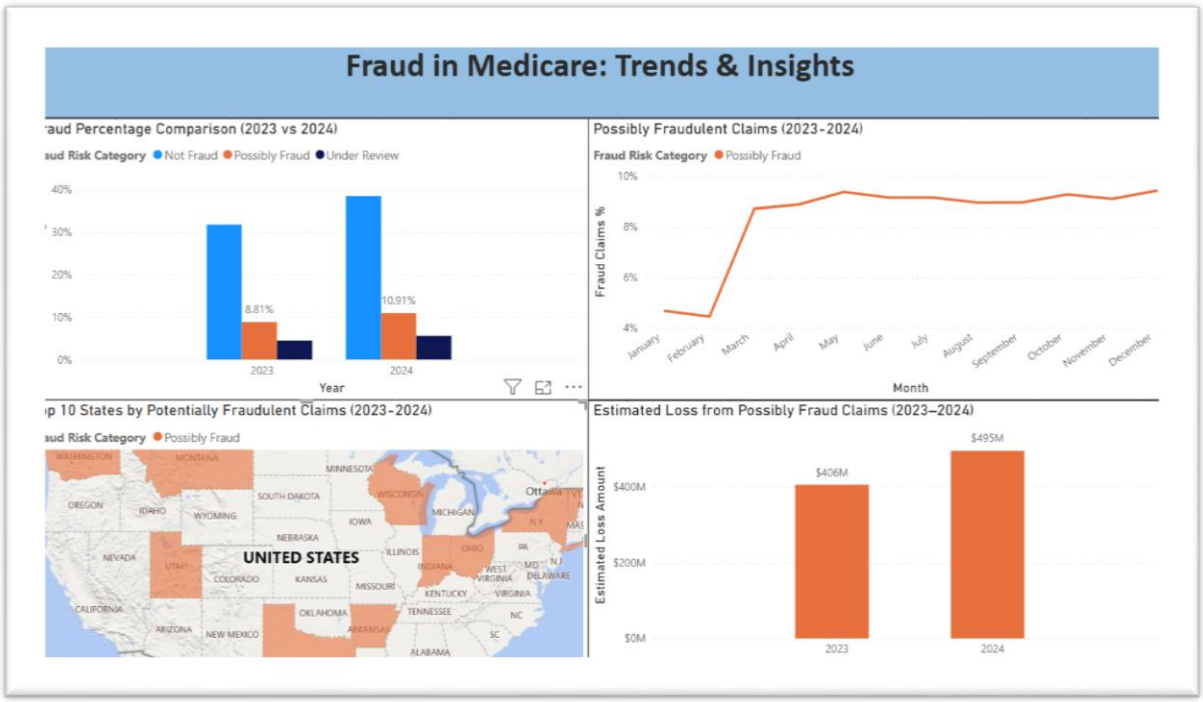
Risk: Although data is anonymized, feature combinations could inadvertently enable re-identification.
Mitigation: Use only de-identified public data and ensure model access is restricted to secure CMS environments.

5. Over-Reliance on Automation

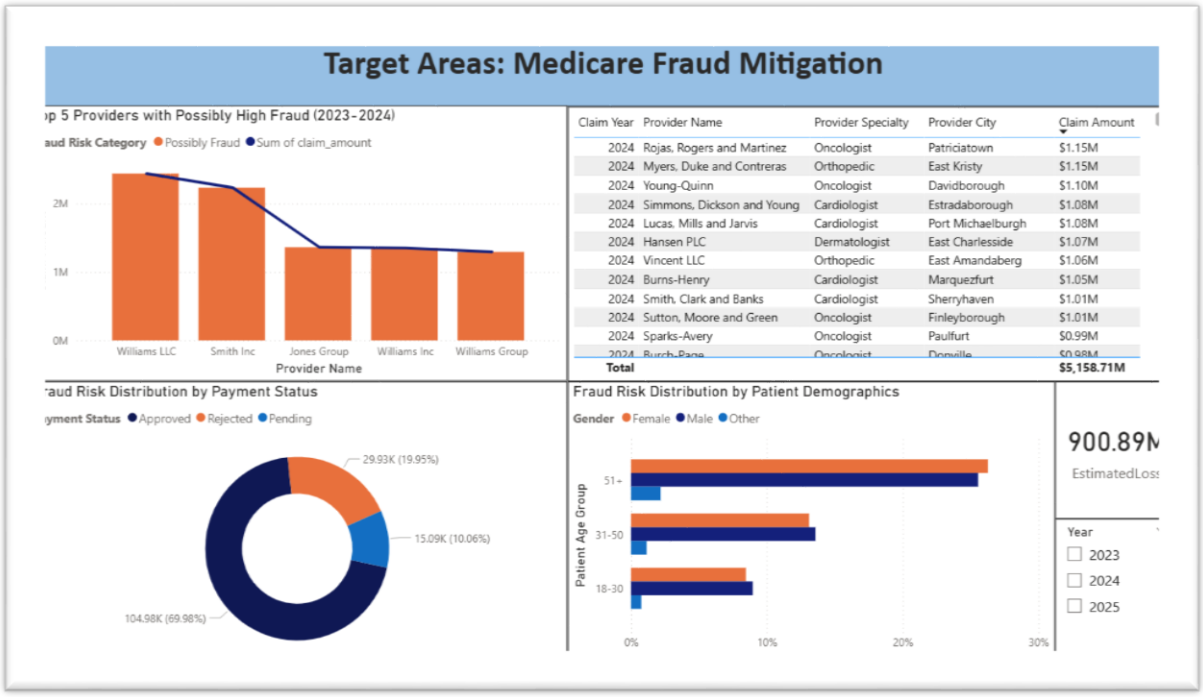
Risk: Auditors may defer entirely to model output, risking blind acceptance of flags.
Mitigation: Position models as decision-support tools and train investigators to interpret model scores critically.

16.Visualizations

1. Fraud Overview Dashboard



2. Action Insights Dashboard



17. Conclusion & Recommendations

Fraud in Medicare claims—particularly within Part D—continues to erode trust and divert critical resources away from legitimate care. This project demonstrates the value of combining analytics, machine learning, and business intelligence tools to create a more responsive, transparent, and efficient fraud prevention system.

By embedding risk scoring models within CMS's claim review pipelines and integrating dashboards that highlight fraud trends by provider and region, CMS can improve audit precision, enhance cost efficiency, and respond more rapidly to emerging fraud schemes.

We recommend pilot testing this solution in one or two high-risk states, expanding the model scope to include inpatient and outpatient data, and evolving the system through ongoing feedback from CMS investigators. Additionally, training teams on model logic and deploying SHAP-based dashboards will further strengthen CMS's leadership in data-driven healthcare oversight.

References

- **Centers for Medicare & Medicaid Services (CMS). (2024). Medicare Improper Payments Report.-** https://www.cms.gov/files/document/cms-financial-report-fiscal-year-2024.pdf

- Kaggle: Healthcare Fraud Detection Dataset – https://www.kaggle.com/datasets/jaiswalmagic1/healthcare-fraud-detection-dataset
- National Health Care Anti-Fraud Association. (2023). Health Care Fraud Overview.

- SHAP Python Library. https://shap.readthedocs.io/en/latest/index.html

- U.S. Department of Health and Human Services, Office of Inspector General (OIG**).** (2023). *Combating Fraud in Medicare*. Retrieved from: https://oig.hhs.gov

- McKinsey & Company**.** (2022). *Applying AI to Healthcare Claims Fraud Detection*. Retrieved from: https://www.mckinsey.com

- News Articles used for Research Purpose **-**https://www.kff.org/medicare/issue-brief/medicare-program-integrity-and-efforts-to-root-out-improper-payments-fraud-waste-and-abuse