# The Use of Foundation Models to Detect Medicare Fraud, Waste, and Abuse

## Experimental Results

Ben Goodwin

University of Denver

Objective

To determine if foundation models can outperform traditional statistical methods in detecting fraudulent healthcare providers using CMS Medicare Part D & B utilization data.

Data

Approximately 1.4 million providers aggregated by National Provider Index (NPI) with the List of Excluded Individuals/Entities (LEIE) as the ground truth for fraud. These providers listed on the LEIE have been banned from participating in Medicare.

Challenges

The dataset consists of aggregated provider data organized by each provider's HCPCS claims. Columns include:

| Medicare Physician & Other Practitioners - by Provider and Service Data Dictionary | | |
|---|---|---|
| Term Name | Variable Name | Definition |
| National Provider Identifier | Rndrng_NPI | National Provider Identifier (NPI) for the rendering provider on the claim. The provider NPI is the numeric identifier registered in NPPES. |
| Last Name/Organization Name of the Provider | Rndrng_Prvdr_Last_Org_Name | When the provider is registered in NPPES as an individual (entity type code='I'), this is the provider's last name. When the provider is registered as an organization (entity type code = 'O'), this is the organization name. |
| First Name of the Provider | Rndrng_Prvdr_First_Name | When the provider is registered in NPPES as an individual (entity type code='I'), this is the provider's first name. When the provider is registered as an organization (entity type code = 'O'), this will be blank. |
| Middle Initial of the Provider | Rndrng_Prvdr_MI | When the provider is registered in NPPES as an individual (entity type code='I'), this is the provider's middle initial. When the provider is registered as an organization (entity type code = 'O'), this will be blank. |
| Credentials of the Provider | Rndrng_Prvdr_Crdntls | When the provider is registered in NPPES as an individual (entity type code='I'), these are the provider's credentials. When the provider is registered as an organization (entity type code = 'O'), this will be blank. |
| Entity Type of the Provider | Rndrng_Prvdr_Ent_Cd | Type of entity reported in NPPES. An entity code of 'I' identifies providers registered as individuals and an entity type code of 'O' |

| | | |
|---|---|---|
| | | identifies providers registered as organizations. |
| Street Address 1 of the Provider | Rndrng_Prvdr_St1 | The first line of the provider's street address, as reported in NPPES. |
| Street Address 2 of the Provider | Rndrng_Prvdr_St2 | The second line of the provider's street address, as reported in NPPES. |
| City of the Provider | Rndrng_Prvdr_City | The city where the provider is located, as reported in NPPES. |
| State Abbreviation of the Provider | Rndrng_Prvdr_State_Abrvtn | The state where the provider is located, as reported in NPPES. The fifty U.S. states and the District of Columbia are reported by the state postal abbreviation. The following values are used for all other areas: |
| State FIPS Code of the Provider | Rndrng_Prvdr_State_FIPS | FIPS code for rendering provider's state. |
| Zip Code of the Provider | Rndrng_Prvdr_Zip5 | The provider's zip code, as reported in NPPES. |
| RUCA Code of the Provider | Rndrng_Prvdr_RUCA | Rural-Urban Commuting Area Codes (RUCAs), are a Census tract-based classification scheme that utilizes the standard Bureau of Census Urbanized Area and Urban Cluster definitions in combination with work commuting information to characterize all of the nation's Census tracts regarding their rural and urban status and relationships. The Referring Provider ZIP code was cross walked to the United States Department of Agriculture (USDA) 2010 Rural-Urban Commuting Area Codes. |
| RUCA Description | Rndrng_Prvdr_RUCA_Desc | Description of Rural-Urban Commuting Area (RUCA) Code |

| Medicare Physician & Other Practitioners - by Provider and Service Data Dictionary | | |
|---|---|---|
| Term Name | Variable Name | Definition |
| Country Code of the Provider | Rndrng_Prvdr_Cntry | The country where the provider is located, as reported in NPPES. The country code will be 'US' for any state or U.S. possession. For foreign countries (i.e., state values of 'ZZ'), the provider country values include the following: |

| | | |
|---|---|---|
| Provider Type of the Provider | Rndrng_Prvdr_Type | Derived from the provider specialty code reported on the claim. For providers that reported more than one specialty code on their claims, this is the specialty code associated with the largest number of services. |
| Medicare Participation Indicator | Rndrng_Prvdr_Mdcr_Prtcptg_Ind | Identifies whether the provider participates in Medicare and/or accepts assignment of Medicare allowed amounts. The value will be 'Y' for any provider that had at least one claim identifying the provider as participating in Medicare or accepting assignment of Medicare allowed amounts within HCPCS code and place of service. A non-participating provider may elect to accept Medicare allowed amounts for some services and not accept Medicare allowed amounts for other services. |
| HCPCS Code | HCPCS_Cd | HCPCS code used to identify the specific medical service furnished by the provider. HCPCS codes include two levels. Level I codes are the Current Procedural Terminology (CPT) codes that are maintained by the American Medical Association and Level II codes are created by CMS to identify products, supplies and services not covered by the CPT codes (such as ambulance services). CPT codes, descriptions and other data only are copyright 2016 American Medical Association. All rights reserved. CPT is a registered trademark of the American Medical Association (AMA). Please review the complete CMS AMA CPT License agreement which is presented to users when accessing the data. For additional information on HCPCS codes, visit the HCPCS general information page. |

| | | |
|---|---|---|
| HCPCS Description | HCPCS_Desc | Description of the HCPCS code for the specific medical service furnished by the provider. HCPCS descriptions associated with CPT codes are consumer friendly descriptions provided by the AMA. CPT Consumer Friendly Descriptors are lay synonyms for CPT descriptors that are intended to help healthcare consumers who are not medical professionals understand clinical procedures on bills and patient portals. CPT Consumer Friendly Descriptors should not be used for clinical coding or documentation. All other descriptions are CMS Level II descriptions provided in long form. Due to variable length restrictions, the CMS Level II descriptions have been truncated to 256 bytes. As a result, the same HCPCS description can be associated with more than one HCPCS code. For complete CMS Level II descriptions, please visit the HCPCS Release Code Sets page. |

| Medicare Physician & Other Practitioners - by Provider and Service Data Dictionary | | |
|---|---|---|
| Term Name | Variable Name | Definition |
| HCPCS Drug Indicator | HCPCS_Drug_Ind | Identifies whether the HCPCS code for the specific service furnished by the provider is a HCPCS listed on the Medicare Part B Drug Average Sales Price (ASP) File. Please visit the ASP drug pricing page for additional information. |
| Place of Service | Place_Of_Srvc | Identifies whether the place of service submitted on the claims is a facility (value of 'F') or non-facility (value of 'O'). Non-facility is generally an office setting; however other entities are included in non-facility. The following values are entities included in facility and non-facility: |
| Number of Medicare Beneficiaries | Tot_Benes | Number of distinct Medicare beneficiaries receiving the service for each Rndrng_NPI, HCPCS_Cd, and Place_Of_Srvc. |
| Number of Services | Tot_Srvcs | Number of services provided; note that the metrics used to count the number provided can vary from service to service. |

| | | |
|---|---|---|
| Number of Distinct Medicare Beneficiary/Per Day Services | Tot_Bene_Day_Srvcs | Number of distinct Medicare beneficiary/per day services. Since a given beneficiary may receive multiple services of the same type (e.g., single vs. multiple cardiac stents) on a single day, this metric removes double-counting from the line service count to identify whether a unique service occurred. |
| Average Submitted Charge Amount | Avg_Sbmtd_Chrg | Average of the charges that the provider submitted for the service. |
| Average Medicare Allowed Amount | Avg_Mdcr_Alowd_Amt | Average of the Medicare allowed amount for the service; this figure is the sum of the amount Medicare pays, the deductible and coinsurance amounts that the beneficiary is responsible for paying, and any amounts that a third party is responsible for paying. |
| Average Medicare Payment Amount | Avg_Mdcr_Pymt_Amt | Average amount that Medicare paid after deductible and coinsurance amounts have been deducted for the line item service. Note: In general, Medicare FFS claims with dates-ofservice or dates-of-discharge on or after April 1, 2013, incurred a 2 percent reduction in Medicare payment. This is in response to mandatory across-the-board reductions in Federal spending, also known as sequestration. |
| Average Medicare Standardized Payment Amount | Avg_Mdcr_Stdzd_Amt | Average amount that Medicare paid after beneficiary deductible and coinsurance amounts have been deducted for the line item service and after standardization of the Medicare payment has been applied. Standardization removes geographic differences in payment rates for individual services, such as those that account for local wages or input prices and makes Medicare payments across geographic areas comparable, so that differences reflect variation in factors such as physicians' practice patterns and beneficiaries' ability and willingness to obtain care. Please refer to the methodology document for more details on the standardization of Medicare payments. |

As in most instances of fraud, the minority class is extremely rare, documented fraud occurs >0.1% within the dataset, making standard supervised learning difficult. Traditional statistical methods also struggle with the detection of Medicare fraud due to the subtle context gaps that exist within healthcare data. Statistically a neurosurgeon looks like a fraudster compared to a family practitioner because their charges are typically astronomically higher. Without the

semantic knowledge of “what a doctor does,” numerical models struggle to differentiate “expensive” from “illegal.”

Data Engineering & Pipeline Architecture:

To transition from raw administrative claims to action forensic intelligence, a multi-stage pipeline was created. This process reduced dimensionality while preserving the statistical signals required for both ML and foundation reasoning.

A. Source Data & Aggregation
   The raw dataset consists of the CMS Medicare Provider Utilization and Payment Data, which details services provided to Medicare beneficiaries
   Raw Granularity: Service-Line Level (HCPCS Code per provider)
   Target Granularity: Provider Level (Unique NPI)

   A groupby(‘Rndrng_NPI’) aggregation to create a “Provider Profile”, collapsing thousands of service lines into a single feature using the following aggregation logic:

   1. Volume Metrics: Sum of Services and Beneficiaries (measuring total output).
   2. Financial Metrics: Mean and Max of Submitted Charges.

      Note: Retaining the Max charge is critical for fraud detection, as "unbundling" schemes often involve burying a few extremely high-cost claims amidst thousands of normal ones.
   3. Behavioral Metrics: Unique Count of States (detecting geographic anomalies).

B. Enrichment: Ground Truth & Taxonomy
   To label the dataset for training, two external government databases:
   1. The LEIE (Ground truth): We joined the dataset with the Office of the Inspector General’s List of Excluded Individuals/Entities.
      Target Variable: on_lexi (Boolean): If a provider appeared on the exclusion list, they were labeled as fraud
   2. NPPES (Taxonomy): Provider epeciality codes to human-readable descriptions.

C. Feature Engineering: “The Forensic Features”
   Standard raw metrics are often misleading. To normalize these inputs, two features were engineered:
      1. The Greed Ratio (charge_Ratio): This isolates the markup
      2. The Risk Score (Type_risk_score): The historical fraud rate for each specialty

D. The Model Context Layer (RAG Preperation)
   Foundation models don’t generally know that 5,000 services is “high” for a dentist, but “low” for a lab. To solve this, a peer group database was built to facilitate RAG.
   Benchmarking Process:
      1. Grouping: Segmenting the entire dataset of 1.4M providers by Rndrng_Prvdr_type
      2. Baseline calculation: For every specialty, we calculated the median volume, mean charge, and standard deviation of charges.

3. Z-Score Pre-computation: Before prompting the model, the providers statistical distance from the peer group was calculated.

The data engineering in this step offloaded the mathematical heavy lifting to python, allowing the LLM to focus purely on forensic reasoning and pattern recognition.

Experiment I: The Statistical Baseline (Logistic Regression)

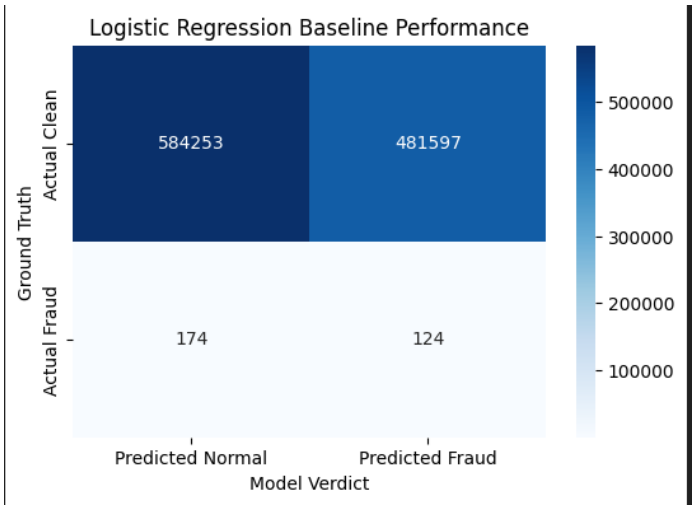
Setup:

Model: Logistic Regression with class_weight='balanced'

Features: Standardized numerical metrics (Service Volume, Beneficiary Count, Avg Charges)

Results: ROC-AUC Score: 0.4871

Confusion Matrix:




Analysis of Failure: The model learned that "High Charges = Fraud "

However, because it didn't know Peer Group benchmarks, it flagged every high-cost specialist (Surgeons, Oncologists) as fraudulent, leading to a precision of 0.00.

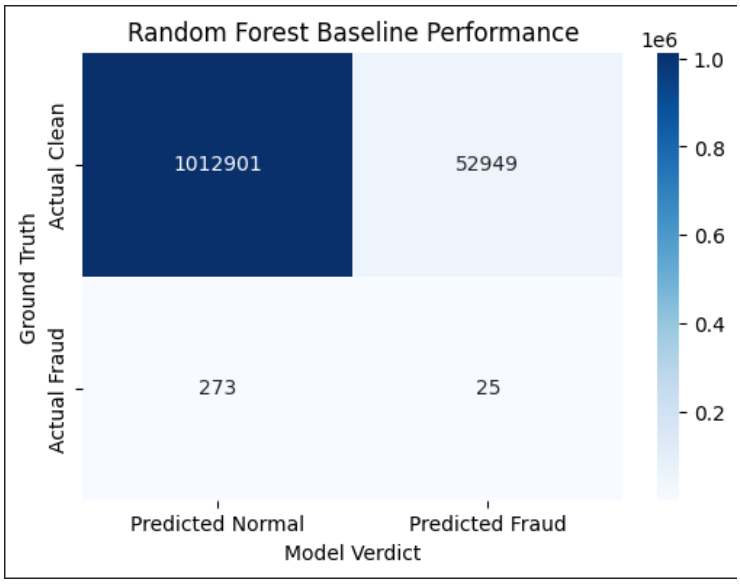
Experiment II: Random Forest

Setup:

Model: Random Forest

Results: ROC-AUC: 0.5959

Confusion Matrix:




Analysis of Failure: The model achieved 95% accuracy; it missed 92% of the actual fraud cases (recall 0.08). The model learned that since 99.9% of doctors are honest, the safest bet is to predict “clean” every time.

Feature Importance Analysis: The random forest revealed what matters, even if it couldn’t predict the outcome:

1. Type_Risk_Score (24.2%): The historical fraud rate of the providers specialty.
2. Charge_Ratio (12.2%): The ratio of submitted charges to Medicare allowed amounts (read, Greed)
3. Tot_Benes_sum (11.5%): Patient volume

Conclusion: Numerical ML models can identify risk factors, but they lack the sematic reasoning to convict specific providers.

Experiment III: The LLM Auditor (Zero-shot & RAG)

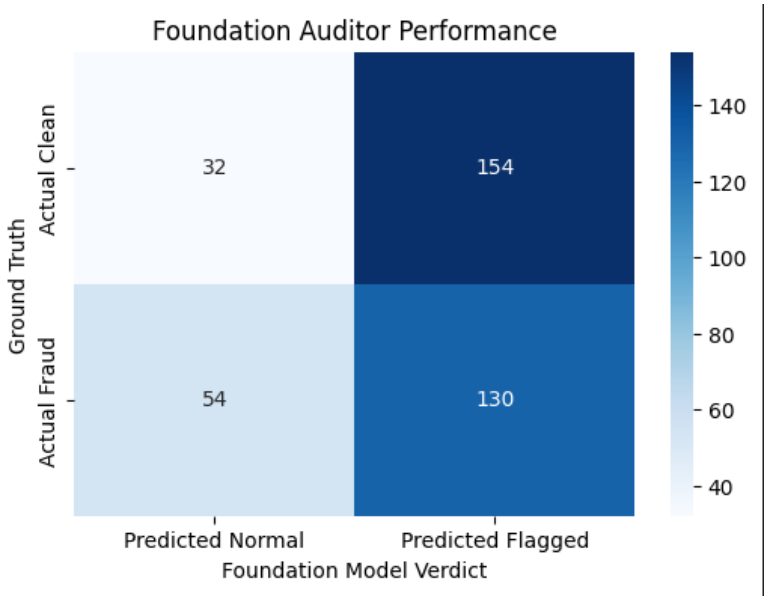
Hypothesis: An LLM can use “Semantic Reasoning” to interpret the relationship between volume, specialty, and charges, rather than just looking at raw numbers.

Model: Ollama, mistral

A. The “Blind” Test (Zero-shot)
   Approach: Presented model with summary of the provider’s stats versus their peer group benchmarks
   Result: The model achieved 75% accuracy on a balanced (small) test set.
   Behavior: It correctly identified 100% of the fraud cases (high recall), but suffered from hallucination regarding “under-utilization” (flagging part time doctors as suspicious).
B. The “Prosecutor” Agent (Logic-Gated)
   Approach: The system was refined by adding Python logic gates (to filter out low-volume) providers as a “prosecutorial” system prompt.

Results:

Confusion Matrix:




Performance Metrics:

The agent was tested on a balanced dataset of 200 fraudulent providers and 200 clean providers.

The most critical metric in fraud detection is recall and the foundation model caught 71% of fraud cases.

Implication: While the traditional ML models “played it safe” to maximize accuracy, the LLM successfully identified the semantic patterns of fraud (churning, upcoding) that evaded numerical detection.

However, the trade-off for high recall was low precision (0.46). The model flagged many clean providers as suspicious. This could mean a few different things, one could be that the model is simply acting as a dragnet, it captures anyone exhibiting statistical anomalies. It could also be that these providers are engaged in some sort of Fraud, Waste, or Abuse and have not been flagged.

Since accuracy was quite low, in an audit workflow, this is valuable. It means that for every two providers the model flags for review, 1 is an actual fraud case, this is a significant hit rate.

Conclusion

While the random forest offered the highest statistical stability, it was functionally useless for fraud detection because it ignored the minority class.

The LLM prosecutor, despite lower overall accuracy, proved to be the superior forensic tool. By achieving 71% recall, it demonstrated that generative AI can detect complex fraud patterns that evade traditional statistical modeling. The experiment proves that the model acts as a hypersensitive anomaly detector.

Some next steps:

1. Hard logic filter: automatically discard any provider with volume < 50% of peer median
2. Only review the remaining providers
3. Focus on verdicts labeled as abuse or fraud, while treating waste as low-priority inefficiencies.