INFER: A Multi-Agent Framework for Detecting Fraud, Waste, and Abuse in Medicare Claims

Anonymous Authors[1]

Abstract

Medicare Fraud, Waste, and Abuse (FWA) costs the U.S. healthcare system between $100B and $300B annually, yet current detection approaches recover only a fraction of these losses. We identify three structural gaps in existing methods that limit their effectiveness and present INFER, a multi-agent platform that addresses each gap through purpose-built architectural components. Applied to the CMS 5% Limited Data Set (LDS) spanning 3.6M DME claims, 254K hospice claims, and 42.8M Part B claim lines (2022–2024) as part of the CMS Chili Cook-Off Challenge, INFER achieved precision of 0.98 on clinician-reviewed high-confidence cases, an approximately 55% reduction in false positives attributable to the multi-agent design, and 9 fully-executed FWA detection pipelines. We detail the FWA taxonomy co-designed with clinical experts, present ablation analysis quantifying each agent's contribution, and provide worked examples demonstrating how the multi-agent design detects patterns invisible to prior approaches.

1. Introduction

In the United States, the National Health Care Anti-Fraud Association estimates that 3–10% of total healthcare spending, between $100B and $300B annually, is lost to fraud (National Health Care Anti-Fraud Association, 2024). The scale of the problem demands computational approaches, yet current detection systems face fundamental limitations that reduce their effectiveness. Recent federal oversight reports document expected recoveries of $7.13B (HHS Office of Inspector General, 2024), and targeted interventions have prevented over $4.2B in improper payments in specific categories alone (CMS Center for Program Integrity, 2024). The predominant paradigm in Medicare FWA detection relies on three categories of methods: (a) rule-based edit systems that enforce known billing constraints, (b) single-model statistical or ML classifiers that flag outliers on individual claim features, and (c) general-purpose large language models (LLMs) applied without domain-specific guardrails.

However, these approaches suffer from distinct shortcomings:

S1 High false-positive rates and investigator burden. Single-model classifiers optimized for recall produce unmanageable case volumes. Without adversarial validation, errors propagate unchallenged to human reviewers.

S2 Inability to detect coordinated, multi-actor schemes. Rule-based edits and claim-level classifiers operate on individual claims or provider profiles. They cannot capture network-level fraud, kickback arrangements, telemarketing rings, or coordinated billing across ostensibly independent entities, which represents a growing share of Medicare losses.

S3 Vulnerability to emerging AI-assisted fraud. General-purpose LLMs introduced without guardrails confabulate, generating plausible but unfounded assertions (Nigam, 2025b), and carry data leakage risks incompatible with federal data sharing agreements.

We present INFER (Agentic Inference Framework for Evidence and Multi-Tiered Reasoning), a purpose-built multi-agent platform that addresses each shortcoming through architectural design: (1) A Critique agent adversarially stress-tests every Finder detection, filtering false positives before they reach human reviewers. (2) Graph-based anomaly detection within the Finder agent models physician–supplier–beneficiary networks, surfacing coordinated schemes invisible to claim-level methods. (3) Domain-specific training on 10M+ patient records eliminates confabulation risk, while dedicated LLM-generated content detectors identify AI-fabricated documentation.

We evaluate INFER on the CMS Chili Cook-Off Challenge datasets, covering DME, Hospice, and Part B Limited Data

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Sets (2022–2024), identifying 34 clinically defined FWA scenarios co-developed with nurses, physicians, and Special Investigations Unit (SIU) experts.

2. The CMS Limited Data Set

Medicare is the U.S. federal health insurance programme covering approximately 67 million beneficiaries aged 65 and older. It operates primarily as a fee-for-service (FFS) system across three major benefit categories relevant to this work: Part B (Carrier) covers physician/outpatient services (E&M visits, procedures, diagnostics). DME (Durable Medical Equipment) covers wheelchairs, braces, oxygen equipment, and supplies. Hospice covers end-of-life care for beneficiaries with a terminal prognosis of six months or less.

The CMS 5% Limited Data Set (LDS) is a beneficiary-level sample: all claims for each sampled beneficiary are included, preserving within-patient longitudinal integrity. Table 1 in the appendix summarizes the dataset structure.

3. FWA Taxonomy: Clinical Co-Design of 34 Detection Scenarios

A central contribution of this work is the systematic co-design of FWA detection scenarios with domain experts. We convened a panel of registered nurses, physicians, and private-payer SIU investigators to enumerate fraud, waste, and abuse patterns across DME, Hospice, and Part B, grounded in OIG audit typologies, commercial payer experience, and clinical practice knowledge. This process yielded 34 distinct FWA scenarios organized across three analysis levels:

- **Beneficiary-level signals (micro):** post-mortem billing, utilization anomalies, demographic inconsistencies, coverage exploitation.
- **Provider/supplier-level signals (meso):** billing pattern outliers, patient population anomalies, service inconsistencies, temporal anomalies.
- **Network-level signals (macro):** referral concentration, geographic dispersion, multi-entity collusion, coordinated billing rings.

Table 2 presents the nine scenarios for which full detection pipelines were executed within the challenge window, alongside their detection approach and a concrete illustrative example of each fraud type.

4. System Architecture: INFER

INFER operates through three coordinated agents that mirror adversarial judicial processes: generation, challenge, and adjudication. This architecture is designed to address the three shortcomings identified in Section 1.

4.1. Agent 1: Finder (Signal Generation)

The Finder surfaces candidate FWA signals across micro (claim-level), meso (provider-level), and macro (network-level) patterns. It addresses **S2** by incorporating graph-based detection alongside statistical and ML methods, and **S3** through domain-specific training that eliminates confabulation. The Finder is not a single model but a method-selection engine that chooses from algorithm families based on scenario characteristics, including transformer-based architectures (Vaswani et al., 2017) for documentation analysis and gradient-boosted trees (Chen & Guestrin, 2016) for tabular claim features.

4.2. Agent 2: Critique (Adversarial Validation)

The Critique directly addresses **S1**. It stress-tests every Finder output against plausible benign explanations. For each flagged case, the Critique generates counter-hypotheses—for example, that a DME duplicate might reflect a legitimate supplier transition, or that a long hospice stay might be appropriate for a dementia patient. It operates on a configurable precision–recall threshold: for this analysis, we optimized for precision to minimize investigator burden.

4.3. Agent 3: Judge (Confidence-Scored Adjudication)

The Judge synthesizes Finder evidence and Critique analysis to render a final determination with explicit confidence calibration. High-confidence cases are separated from moderate-confidence cases and grey-area cases (with structured next steps and specification of what additional data would resolve uncertainty). Every determination includes a traceable reasoning chain supporting reproducibility and downstream legal processes.

5. Detection Methodology: Worked Examples

To illustrate how INFER's architecture provides advantages over prior approaches, we detail one representative scenario.

5.1. Kickback Detection via Graph Analytics

Detection (Finder): A bipartite network is constructed from DMERC base and line files, where physician nodes connect to supplier nodes via DME claim lines. An Isolation Forest identifies physicians whose network topology is structurally inconsistent with specialty peers.

Validation (Critique): Counter-hypotheses include legitimate multi-state telemedicine practice, complex patient populations, or regional supply constraints. The Critique checks

whether the physician's specialty and telehealth share are consistent with a legitimate remote practice model.

Why prior methods fail: Claim-level methods cannot see the network; they would flag the physician's high volume but miss the structural relationship with specific suppliers. Graph-based detection reveals that the physician routes 80% of volume to two suppliers across 15 states, a topology that claim-level analysis cannot represent.

6. Evaluation

6.1. Precision and Component Ablation

Manual review by clinical experts (nurses and physicians) of high-confidence flagged cases yielded a measured precision of 0.98. Formal recall measurement is infeasible due to the absence of labelled ground truth in LDS data.

To quantify each agent's contribution, we conducted an ablation analysis on the DME phantom billing scenario (Table 3 in the appendix). The Finder alone (without Critique or Judge) produces a larger candidate set with 0.72 precision. Adding the Critique dramatically improves precision to 0.94 by eliminating false positives. The Judge further refines by stratifying into actionable confidence tiers and increases the precision to 0.98.

7. Discussion

Conventional FWA detection in Medicare relies on CMS's Comprehensive Error Rate Testing (CERT) program for post-payment review (Centers for Medicare & Medicaid Services, 2024a;b) and National Correct Coding Initiative (NCCI) edits for pre-payment rule enforcement. These are effective for known, well-defined violations but cannot adapt to novel fraud patterns. Prior systematic reviews have catalogued machine learning approaches to healthcare fraud detection (du Preez et al., 2024), but these largely focus on single-model classifiers rather than multi-agent architectures. To our knowledge, INFER is the first deployed FWA system to incorporate LLM-generated documentation detection (Nigam, 2025a).

Broader implications: The multi-agent architecture—detection, adversarial validation, and confidence-scored adjudication—is applicable beyond Medicare. Any domain requiring high-stakes anomaly detection with low false-positive tolerance and auditability requirements (financial fraud, insurance claims, regulatory compliance) could benefit from this paradigm (McKinsey & Company, 2024). The key insight is that the placement of human judgment matters as much as its presence: human-in-the-loop review fails when positioned as final validation rather than architected into the inference process.

Limitations: The 5% beneficiary sample constrains network-level detection power: provider-level statistics are partial, and some apparent networks may dissolve at full population scale. Recall cannot be formally measured without ground truth. The system was evaluated on U.S. Medicare data; applicability to other healthcare systems (e.g., NHS) would require adaptation to different billing structures, but the multi-agent architecture is transferable.

8. Conclusion

We presented INFER, a multi-agent agentic AI platform for Medicare FWA detection that addresses three structural limitations of existing approaches: false-positive burden, inability to detect coordinated schemes, and vulnerability to AI-assisted fraud. Applied to CMS Chili Cook-Off Challenge datasets spanning DME, Hospice, and Part B (2022–2024), the system identified high-confidence FWA patterns across 34 clinically defined scenarios with 0.98 precision, projecting \$1.07–1.75B in recoverable Year 1 savings at population scale. The architecture is GovCloud-ready, commercially validated ($11\times$ ROI), and designed for continuous monitoring rather than batch analysis. Future work includes full-population validation, expanded recall measurement, and adaptation to international healthcare systems.

Impact Statement

This paper presents work whose goal is to advance the detection of fraud, waste, and abuse in healthcare systems. Improved FWA detection has the potential to recover billions of dollars in improper payments, ultimately benefiting taxpayers and the integrity of public healthcare programs. We note that any automated detection system must be deployed with appropriate human oversight and due process protections to avoid unfairly targeting legitimate healthcare providers.

References

Centers for Medicare & Medicaid Services. CERT 2024: Improper payment rates – DMEPOS. Technical report, Centers for Medicare & Medicaid Services, 2024a.

Centers for Medicare & Medicaid Services. 2024 supplemental improper payment data. Technical report, Centers for Medicare & Medicaid Services, 2024b.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, New York, NY, USA, 2016. ACM.

CMS Center for Program Integrity. Urinary catheter case

study – early actions prevent over $4.2 Billion in potential improper payments. Technical report, Centers for Medicare & Medicaid Services, 2024.

du Preez, A. et al. Fraud detection in healthcare claims using machine learning: A systematic review. *Artificial Intelligence in Medicine*, 148:102770, 2024.

HHS Office of Inspector General. Fall 2024 semiannual report to congress: $7.13B in expected recoveries. Technical report, U.S. Department of Health and Human Services, 2024.

McKinsey & Company. Payment integrity in the age of AI and value-based care. Technical report, McKinsey & Company, 2024.

National Health Care Anti-Fraud Association. The challenge of health care fraud. https://www.nhcaa.org/tools-insights/about-health-care-fraud/, 2024. Accessed: 2025.

Nigam, A. Beyond fraud detection: Agentic AI and changes in healthcare integrity. *Forbes*, November 2025a.

Nigam, A. Should your business use a generalist or specialized AI model? *Harvard Business Review*, July 2025b.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

A. Dataset Overview

Table 1. CMS LDS dataset overview (5% beneficiary sample, 2022–2024).

File	Claims	Line Items	Size
DME (DMERC)	3.6M	4.9M	∼1.5 GB
Carrier (Part B)	42.8M	84M	∼8 GB
Hospice	254K	5.5M	∼130 MB
MBSF	—	—	∼30 KB

B. Executed FWA Scenarios

Table 2. Executed FWA scenarios with concrete examples.

Domain	Scenario	Method	Specific Scenarios Captured
DME	Phantom Billing	Statistical + Z-score	Supplier bills 500+ orthotic braces with positive submitted charges but zero payments, zero allowed amounts, and denial codes across hundreds of beneficiaries.
DME	Overutilization	Statistical + IF	Beneficiary receives 10× the expected monthly supply of diabetic test strips with weak diagnosis alignment, driven by a single supplier.
DME	Duplicate Billing	Rule-based interval overlap	Same brace billed by two different suppliers for the same beneficiary within overlapping service dates.
DME	Kickbacks	Graph + IF	Physician orders braces for 200+ beneficiaries across 15 states via telehealth, with 80% of volume routed to two suppliers.
Hospice	Long Stays	Statistical + IF	Provider's mean LOS exceeds 400 days; >40% of patients enrolled >365 days, most discharged alive.
Hospice	GIP/CHC Overuse	Statistical peer comp.	Hospice bills 20% of days as General Inpatient vs. 2% national average, without corresponding symptom crisis documentation.
Hospice	Cap Evasion	Pattern rules	Hospice discharges ≥10 patients alive in Q4 and readmits them in Q1, cycling around the aggregate payment cap.
Part B	E&M Upcoding	Statistical benchmarking	Provider bills 95% of visits as 99215 (highest-complexity E&M) vs. specialty peer average of 25%.
Part B	Unbundling	Code-pair adjacency graphs	Provider consistently bills 82247 + 82310 + 82374 separately instead of the bundled Comprehensive Metabolic Panel (80053).

C. Ablation Analysis

Table 3. Ablation analysis: DME phantom billing detection.

Configuration	Candidates	Precision	Effect
Finder only	∼3× baseline	∼0.72	Broad detection; high false-positive rate
Finder + Critique	∼1.5× baseline	∼0.94	Adversarial filtering removes ∼55% of FPs
Finder + Critique + Judge	Baseline (high-conf.)	∼0.98	Confidence stratification; actionable triage

D. Consolidated Savings Projections

Table 4. Consolidated savings projections (full Medicare population).

Category	Annual Spend	IPR	Year 1 Est.	Year 5 Est.
DME	\$9.1B	21.4%	\$150–250M	\$500–800M
Hospice	\$25B	7.1%	\$120–200M	\$400–650M
Part B	\$553B	∼6.5%	\$800M–1.3B	\$2.5–4.0B
Total			**\$1.07–1.75B**	**\$3.4–5.45B**

E. Complete FWA Scenario Taxonomy (34 Scenarios)

The full taxonomy of 34 FWA scenarios is organized by domain and detection approach below. Nine were fully executed within the challenge window (marked with ★); the remainder represent the complete analytical roadmap.

Table 5. Complete FWA scenario taxonomy.

Domain	FWA Scenario	Detection Approach	Key LDS Variables
DME	★ Phantom Billing	Statistical + Z-score	LSBMTCHG, LALOWCHG, PMTDNLCD
DME	★ Overutilisation	Statistical + Iso. Forest	HCPCS, LINE_SRVC_CNT, ICD_DGNS
DME	★ Duplicate Billing	Rule-based overlap	HCPCS, CLM_THRU_DT, SUP_NPI
DME	★ Kickbacks	Graph + Iso. Forest	RFR_NPI, SUP_NPI, BENE_STATE_CD
DME	Post-Mortem Billing	Rule-based	MBSF death date, CLM_THRU_DT
DME	Upcoding	Unsupervised ML	HCPCS, LINE_ALOWD_CHRG_AMT
DME	Prolonged Rentals	Rule-based	HCPCS, service dates, rental flag
DME	New Supplier Spikes	Statistical	SUP_NPI, claim dates, volume
DME	Excessive Util./Bene	XGBoost/RF	Per-bene aggregates
Hospice	★ Long Stays	Statistical + Iso. Forest	CLM_THRU_DT, discharge codes
Hospice	★ GIP/CHC Overuse	Statistical peer comp.	Revenue centre codes (0656, 0652)
Hospice	★ Cap Evasion	Pattern rules	Discharge/readmit dates, provider ID
Hospice	High Live-Discharge Rate	Statistical	Discharge status codes
Hospice	Suspect Diagnosis Mix	Statistical	ICD principal diagnosis
Hospice	Short Stays/Revocations	Statistical	LOS, revocation indicators
Hospice	Insufficient Services	Rule-based proxies	Revenue centre units, visit counts
Hospice	Concentrated Referrals	Graph-based	Referring provider NPI
Part B	★ E&M Upcoding	Statistical benchmarking	HCPCS E&M codes, specialty
Part B	★ Unbundling	Code-pair adjacency	HCPCS pairs, same-day clusters
Part B	Duplicate Billing	Rule-based	HCPCS, date, beneficiary, provider
Part B	Modifier Abuse	Statistical	Modifiers 25, 59, 76/77, 91, XU
Part B	Services Not Rendered	Statistical	Cross-claim consistency
Part B	Telehealth Abuse	Supervised ML	POS codes, volume, specialty
Part B	Cloned Claims	Clustering/outlier	Claim similarity metrics
Part B	Medically Unnecessary	XGBoost	Diagnosis-procedure coherence
Part B	Provider ID Misuse	Rule-based	NPI, multi-state billing patterns
Cross	Bene ID Compromise	Clustering	Multi-domain utilisation vectors
Cross	Coordinated Fraud Ring	Graph-based	Multi-entity network topology
Cross	DME During Hospice	Cross-file rules	Hospice + DME date overlap
Cross	Part B During Hospice	Cross-file rules	Hospice + Carrier date overlap
Cross	Phantom Doctor	Cross-file check	RFR_NPI presence in Carrier
Cross	Physician–DME Collusion	Graph-based	NPI linkage across files
Cross	Hospice Referral Kickbacks	Graph-based	Referring NPI concentration

F. Data Validation Framework

All processing occurred within a dedicated CMS-only AWS VPC (us-east-1) with no external egress, copy/paste/print disabled, and KMS-encrypted S3 storage. The validation framework comprised four layers: (1) structural schema validation against LDS data dictionaries with key uniqueness enforcement; (2) financial reconciliation of claim-level vs. line-level

payments; (3) statistical benchmarking of feature distributions against full LDS to ensure flagged outliers were genuine rather than sampling artefacts; and (4) expert clinical review by registered nurses and SIU investigators of anonymised high-risk cases, whose feedback refined detection thresholds. All outputs enforced cell suppression (no count $\leq$10 exposed), and CloudTrail logging enabled evidence production within 24 hours upon CMS/OIG request.

G. Algorithm Selection Rationale

INFER's method-selection engine maps scenario characteristics to algorithm families. Scenarios with well-defined rules (duplicate billing, post-mortem billing) use statistical/rule-based methods for maximum interpretability. Scenarios requiring distributional reasoning (overutilisation, upcoding) use gradient-boosted models with feature importance. Scenarios requiring relational structure (kickbacks, collusion) use graph-based methods. Documentation analysis uses domain-specific LLMs. Autoencoder-based anomaly detection was evaluated but incorporated only as one input among many, given recent concerns about reliability in high-stakes applications. All methods produce explainable outputs with traceable reasoning chains.