ACAVIS PUBLISHERS

Trends in Public Health and Community Medicine

Research Article

Volume 1, Issue 1

Integrating Natural Language Processing and Machine Learning for Documentation-Driven Billing Anomaly Detection in Medicare and Medicaid: A Quantitative Framework for Long-Term Care Facilities

Agyapong MK*

School of Business, Tiffin University, United States

ABSTRACT

Improper payments remain one of the most consequential operational and fiscal risks in U.S. public health insurance. For fiscal year 2025, the Centers for Medicare & Medicaid Services (CMS) reported improper payments of $28.83 billion in Medicare fee-for-service, $23.67 billion in Medicare Part C, $4.23 billion in Medicare Part D, $37.39 billion in Medicaid, and $1.37 billion in the Children's Health Insurance Program. A large share of these losses is documentation-related rather than purely utilization-related, yet most operational detection systems still analyze structured claims without reading the clinical narratives that should substantiate those claims. This paper rebuilds the documentation-driven anomaly detection problem as a joint information problem: what was clinically documented, what was coded, and what was ultimately billed. Using current CMS, HHS, AHRQ, and OIG data, combined with the peer-reviewed literature on clinical natural language processing (NLP), automated coding, and healthcare fraud analytics, the paper develops a hybrid framework for long-term care (LTC) settings that links (1) NLP extraction of diagnoses, services, and severity indicators from unstructured notes, (2) machine-learning detection of anomalous claims behavior, and (3) a cross-referencing engine that scores documentation-billing consistency. Because linked, public LTC note-claim corpora are not available for true patient-level validation, the quantitative contribution of the paper is national and programmatic rather than encounter-level: it documents the financial exposure in nursing care, identifies documentation-dominant error structures, and models conservative savings scenarios from better documentation integrity and automated pre-bill review. The results show that directly quantified documentation-related exposure in FY2025 Medicare fee-for-service and Medicaid alone was approximately $47.65 billion. If integrated documentation-aware anomaly detection reduced that exposure by 10% to 30%, annual savings could plausibly range from $4.77 billion to $14.30 billion before considering secondary spillovers into Medicare Advantage, Part D, or downstream audit efficiencies. The article concludes that documentation-grounded anomaly detection is not merely a technical enhancement; it is a payment-integrity strategy with measurable fiscal and compliance relevance for LTC operators and public payers.

***Correspondence:**
Michael Kwakye Agyapong
School of Business, Tiffin University, Tiffin, Ohio, United States.
Email: mkwakyeagyapong@gmail.com

Article Info:
Received Date: April 28, 2026
Published Date: May 25, 2026

Citation:
Agyapong MK. Integrating Natural Language Processing and Machine Learning for Documentation-Driven Billing Anomaly Detection in Medicare and Medicaid: A Quantitative Framework for Long-Term Care Facilities. Trends Publ Health Commun Med. 2026;1(1):1-14.

KEYWORDS: Natural language processing, Machine learning, Medicare, Medicaid, Improper payments, Clinical documentation integrity, Long-term care, Skilled nursing facilities, Payment integrity, Anomaly detection.

INTRODUCTION

Medicare and Medicaid sit at the center of the U.S. health financing system. CMS reported that national health expenditures reached $5.3 trillion in 2024, equal to 18.0% of gross domestic product, with Medicare spending of $1.118 trillion and Medicaid spending of $931.7 billion. In the CMS brief summaries published in late 2025, Medicare, Medicaid, and CHIP together accounted for just over $1.8 trillion in health care goods and services in 2023, or roughly 40% of total national health spending. These magnitudes mean that even modest inefficiencies in reimbursement, coding, documentation, and oversight translate into very large public losses.

Improper payments are a central part of that problem. Under federal payment-integrity rules, an improper payment is not synonymous with fraud. It may involve overpayment, underpayment, or the inability of a reviewer to determine whether payment was proper because required information was absent, incomplete, or unsupported. This distinction matters. It means that payment integrity is partly a fraud problem, partly a documentation problem, partly a coding problem, and partly a process-control problem. In FY2025, CMS reported $95.49 billion in improper payments across Medicare fee-for-service, Medicare Part C, Medicare Part D, Medicaid, and CHIP. Medicaid alone accounted for $37.39 billion, and CMS stated that 77.17% of those improper payments stemmed from insufficient documentation. In Medicare Part C, CMS identified unsupported diagnosis documentation as the dominant reason for error. In Medicare fee-for-service, the CERT program reported that insufficient documentation was the single largest error category and that skilled nursing facilities were one of the leading drivers of improper payments.

Long-term care facilities deserve special attention in this context. They operate at the intersection of medically complex populations, evolving reimbursement rules, longitudinal documentation, and multiple payment streams. Residents often have multiple chronic conditions, functional limitations, cognitive impairment, therapy needs, medication complexity, and repeated transitions between acute and post-acute settings. These realities create dense and fragmented documentation. They also create fertile conditions for mismatch between the clinical narrative and the billing abstraction. A diagnosis may appear in the Minimum Data Set (MDS) but not in the supporting physician or therapy notes. A therapy intensity or severity assignment may exceed the strength of the charted evidence. A service may have been delivered appropriately but documented incompletely, leading to an improper payment that is operationally costly even when no fraud occurred.

The core premise of this paper is that many billing anomalies cannot be understood correctly from claims data alone. A claims-only system can identify statistical outliers—unusual code frequencies, atypical charge profiles, suspicious timing patterns, and abnormal utilization clusters—but it cannot determine whether those outliers are clinically justified, poorly documented, miscoded, or intentionally inflated. Clinical notes, orders, assessments, certifications, and therapy records contain the missing evidence. Yet a large share of that evidence remains in unstructured text. AHRQ has noted that approximately 80% of EHR data is unstructured, including clinical notes, reports, and other narrative forms. That fact creates both the problem and the opportunity. The problem is that manual review is slow and expensive. The opportunity is that modern NLP systems can extract clinical meaning from notes at scale and allow documentation to be analyzed jointly with billing.

This article reconstructs the original framework paper into a stronger quantitative and policy-oriented manuscript. Rather than treating the model as a purely conceptual architecture, it anchors the argument in current official data and develops a formal documentation-billing consistency framework for LTC payment integrity. The contribution is fourfold. First, the paper synthesizes the latest national evidence on improper payments, LTC spending, nursing-facility exposure, and documentation-driven error categories. Second, it formalizes a hybrid documentation-aware anomaly detection model that combines NLP, machine learning, and explainable review outputs. Third, it quantifies directly observable documentation-related fiscal exposure and develops scenario-based savings estimates under plausible reduction assumptions. Fourth, it translates the framework into implementation steps relevant to skilled nursing facilities, compliance teams, and Medicaid agencies.

The paper proceeds as follows. Section 2 reviews the literature on healthcare fraud analytics, clinical documentation integrity, and automated coding. Section 3 presents the policy and fiscal landscape with emphasis on LTC and nursing-facility risk. Section 4 introduces the data sources, modeling logic, and mathematical framework. Section 5 develops the integrated NLP-ML architecture. Section 6 presents quantitative findings and scenario analysis. Section 7 discusses implementation and governance in LTC settings. Section 8 concludes with implications for research, payment integrity, and federal oversight.

Table 1. FY2025 improper payments and documentation-related findings across major CMS programs

Program	FY2025 improper payment rate	FY2025 improper payments ($ billions)	Documentation-related finding
Medicare fee-for-service	6.55%	28.83	Insufficient documentation was 51.5% of overall improper payments; no documentation added 11.7%.

Program	FY2025 improper payment rate	FY2025 improper payments ($ billions)	Documentation-related finding
Medicare Part C	6.09%	23.67	CMS identified unsupported diagnosis documentation submitted by MA organizations as the dominant source of error.
Medicare Part D	4.00%	4.23	Improper payments continue to reflect sponsor documentation and payment support issues.
Medicaid	6.12%	37.39	77.17% of improper payments were due to insufficient documentation.
CHIP	7.05%	1.37	56.07% of improper payments were due to insufficient documentation.

Note. Improper payment rates and dollar values are from CMS FY2025 improper payment reporting. Documentation findings summarize dominant support issues described in the official program materials.

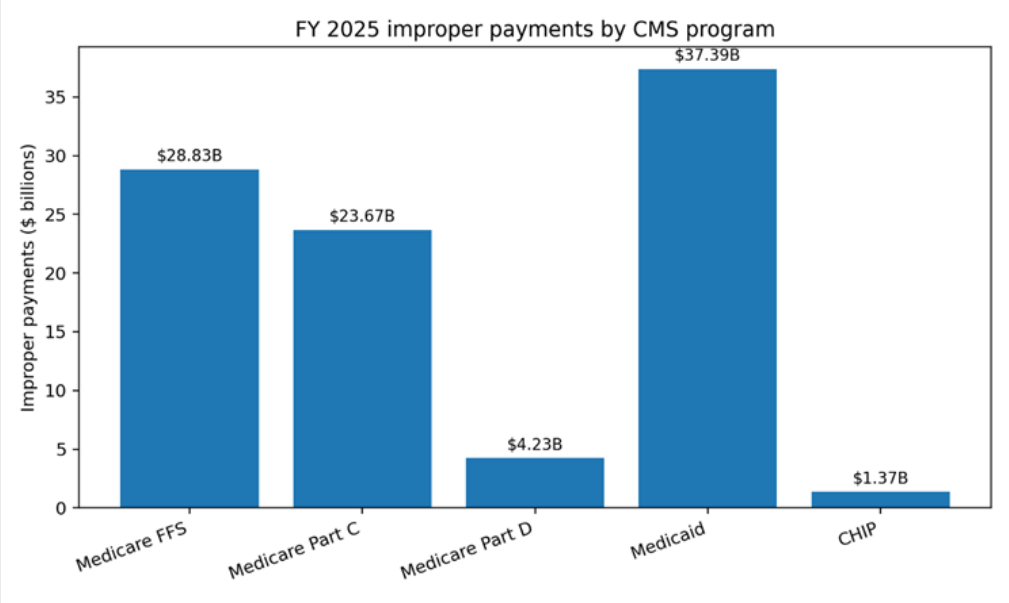

Figure 1. FY2025 Improper Payments by CMS Program

Note. Official improper payment dollar amounts reported by CMS for FY2025.

LITERATURE REVIEW

Claims-Based Healthcare Fraud and Abuse Detection

The early healthcare fraud literature was dominated by rule-based systems and statistical outlier detection. Li et al. (2008) surveyed traditional approaches such as threshold rules, peer-group comparisons, and descriptive statistical screening, showing that these methods were helpful for identifying egregious anomalies but often produced large volumes of false positives. Rule-based approaches remain operationally useful because auditors and payers can interpret them easily, yet their static nature makes them brittle in the face of changing billing patterns and reimbursement rules.

Later work moved toward machine learning. Ahmed et al.[1] documented the rise of anomaly detection methods in finance and related domains, including clustering, isolation forests, one-class methods, and ensemble systems. In healthcare, Bauder et al.[2] surveyed upcoding fraud detection and argued that the core analytical challenge is not simply classifying fraud, but separating legitimate case-mix variation from suspicious coding inflation. Johnson and Khoshgoftaar[3] showed that neural-network approaches could improve Medicare fraud screening relative to simpler classifiers, especially when high-dimensional billing features were available. More recent reviews, such as du Preez et al.[4] confirm that ensemble learning, tree-based methods, and hybrid fraud pipelines now dominate the health-claims literature, largely because they accommodate nonlinear interactions, missingness, and high-cardinality coding variables.

Yet the literature remains structurally limited. Most studies treat the claim as the unit of truth. Inputs usually include diagnosis

codes, procedure codes, provider characteristics, service counts, charges, geography, or timing. The systems learn from patterns in claims submissions, but they generally do not inspect the underlying documentation that should justify the claim. That omission is especially problematic for Medicare Advantage risk adjustment, skilled nursing claims, therapy services, and other contexts where payment depends heavily on diagnosis specificity, functional status, intensity classification, or medical necessity language.

Clinical Documentation Integrity and Coding Accuracy

A parallel literature has developed around clinical documentation integrity (CDI). Davis and Shepheard[5] argue that documentation integrity is foundational to health data quality, patient safety, coding fidelity, and downstream performance measurement. Recent CDI scholarship emphasizes that coding accuracy is inseparable from documentation quality. The record must show not only what condition or service existed, but why it mattered clinically, when it occurred, and how it supported the code billed. Sanderson et al.[6] further show that CDI programs can alter measured case mix and improve the representation of clinical severity when documentation becomes more complete and specific.

This literature is important because payment integrity in public programs often fails at the documentation layer before it fails at the billing layer. A coder can only abstract what is present. An auditor can only validate what is documented. A compliant service can still become an improper payment if supporting records are missing, incomplete, internally inconsistent, or temporally incoherent. That insight aligns strongly with current CMS findings: the largest error category in Medicare fee-for-service is insufficient documentation, and the majority of FY2025 Medicaid improper payments were attributed to insufficient documentation as well.

Clinical NLP, Automated Coding, and EHR Analytics

NLP has become central to extracting information from clinical narratives. Hossain et al.[7] provide a systematic review demonstrating broad use of NLP in EHRs for classification, information extraction, decision support, and risk stratification. Clinical language models such as BioBERT and ClinicalBERT significantly improved entity recognition, assertion detection, and context extraction relative to generic text models. Alsentzer et al.[8] showed that publicly available contextual embeddings trained on clinical notes can enhance downstream medical NLP tasks. Dong et al.[9] and more recent systematic reviews of automated clinical coding conclude that automated code assignment is progressing rapidly, though challenges remain around label complexity, class imbalance, clinical nuance, and human oversight.

What matters for this paper is not fully automated coding in isolation, but document-grounded validation. In LTC, the question is not only "what code should be predicted from the note?" but "does the note support the code that was billed?" That distinction reframes the task as one of evidence alignment rather than purely generative coding. It encourages systems that can extract diagnoses, service descriptions, functional scores, certification indicators, and severity markers from narrative text, and compare them against billed codes, dates, and payment classes. Human-in-the-loop automated coding research supports this orientation, because operational use generally requires transparent evidence rather than black-box predictions alone.

Explainability, Compliance Review, and the Long-Term Care Gap

Explainability is especially important in regulated payment contexts. Healthcare compliance teams need to know why a model flagged a claim, what documentation was missing or contradictory, and what action is recommended. Interpretable ensemble models, SHAP-style feature attribution, and evidence-linked NLP outputs are increasingly favored because they support audit defense and organizational learning. However, the LTC literature remains underdeveloped relative to hospital, emergency, and ambulatory settings. Publicly accessible, linked LTC clinical-note and claim datasets are scarce. As a result, the sector has abundant need but limited methodological infrastructure.

The gap is therefore clear. Claims-based anomaly detection is mature but documentation-blind. CDI scholarship is documentation-rich but often operationally separate from anomaly analytics. Clinical NLP is technically capable but not consistently embedded into payment integrity pipelines. The LTC sector, where documentation complexity and reimbursement sensitivity are high, is precisely where these literatures should meet. This paper addresses that intersection.

POLICY AND FISCAL LANDSCAPE OF LONG-TERM CARE PAYMENT INTEGRITY

The scale of LTC financial exposure can be seen from both national spending data and improper payment reports. CMS reported that spending for nursing care facilities and continuing care retirement communities rose 7.3% in 2024 to $219.9 billion. In the 2025 CMS brief summaries, Medicaid was reported to have paid $69.7 billion for nursing facility services in calendar year 2024, or 30.5% of national nursing facility spending. Medicaid also paid $56.5 billion for home health and other home- and community-based long-term services. These figures underscore that LTC is not a niche payment environment; it is a large and growing component of the public financing system.

The Medicare fee-for-service CERT supplemental report provides additional detail on the compliance intensity of the sector. For the 2025 reporting period, skilled nursing facilities generated a projected $4.3 billion in improper payments and an improper payment rate of 11.8%. That rate exceeded the overall Medicare fee-for-service improper payment rate of 6.55%. More importantly, the top cited root causes in SNF claims were overwhelmingly documentation-related: changes in HIPPS level based on submitted documentation, inadequate case-mix-group documentation, missing orders, missing documentation for primary or active diagnoses reported on the MDS, and absent or inadequate functional-score support.

The operational implication is straightforward. In SNF settings, improper payment risk is often embedded in the chain linking assessment, documentation, coding, and claim submission. A claim can fail because the recorded diagnosis lacks support

in the narrative chart. A case-mix grouping can fail because the documentation needed to justify functional or clinical scoring is incomplete. A therapy-related claim can fail because the certification or recertification trail is deficient. These are precisely the kinds of failures that a documentation-aware detection system is suited to identify before or during pre-bill review.

The policy environment is also moving in this direction. OIG's Nursing Facility Industry Segment-Specific Compliance Program Guidance urges facilities to build risk-based compliance structures tailored to their own operational exposures. CMS quality and reporting programs continue to reinforce evidence capture, coding integrity, and documentation-dependent accountability. HHS has also elevated AI and data modernization as operational priorities, while TEFCA reached nearly 500 million health records exchanged in early 2026. As interoperability expands, the quality and consistency of source documentation becomes even more consequential because inaccurate records do not remain local; they move across systems, affecting payment, quality metrics, and downstream care decisions.

Taken together, these developments make LTC documentation-billing alignment a strategic issue rather than a back-office concern. The question is no longer whether documentation matters for payment integrity. CMS and OIG data already answer that. The real question is whether facilities and payers can operationalize documentation review at scale in time to reduce losses, triage audits, and improve coding quality before claims crystallize into improper payments.[10-19]

Table 2. Improper payment rates by program, FY2021-FY2025

Program	FY2021	FY2022	FY2023	FY2024	FY2025
Medicare fee-for-service	6.26%	7.46%	7.38%	7.66%	6.55%
Medicare Part C	10.28%	5.42%	6.01%	5.61%	6.09%
Medicare Part D	1.33%	1.54%	3.72%	3.70%	4.00%
Medicaid (rolling rate)	21.69%	15.62%	8.58%	5.09%	6.12%

Note. Medicare fee-for-service, Part C, and Part D rates are from CMS improper payment measurement pages; Medicaid rates are rolling PERM rates.

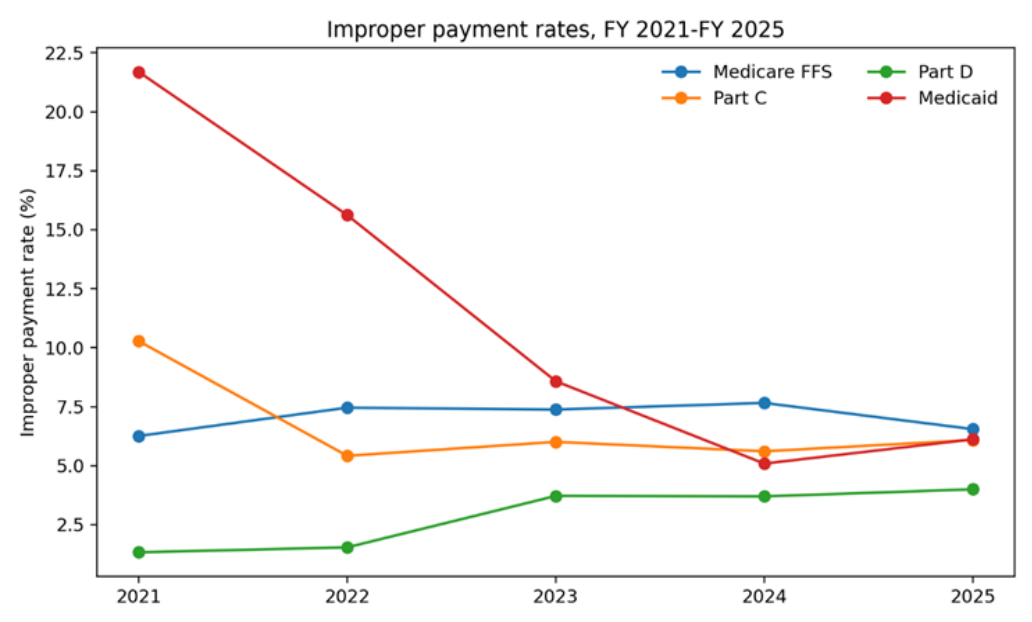

Figure 2. Improper payment rate trends, FY2021-FY2025

Note. The Medicaid series uses rolling PERM rates and therefore is not perfectly comparable to single-year program measurements.

Table 3. Long-term care financial exposure in current federal data

Indicator	Value	Interpretation
National health expenditures, 2024	$5.3 trillion	Medicare and Medicaid payment integrity affects the core of U.S. health financing.

Indicator	Value	Interpretation
Spending for nursing care facilities and continuing care retirement communities, 2024	$219.9 billion	LTC is a large national spending category, not a marginal one.
Medicaid nursing facility spending, 2024	$69.7 billion	Medicaid financed 30.5% of national nursing facility spending.
Skilled nursing facility improper payments in Medicare FFS, FY2025	$4.3 billion	SNFs were one of the principal improper-payment drivers in CERT 2025.
SNF improper payment rate, FY2025	11.80%	This exceeded the overall Medicare FFS rate of 6.55%.

Note. The LTC values combine CMS national expenditure data and Medicare fee-for-service improper payment reporting for skilled nursing facilities.

Data Sources

This paper does not claim patient-level predictive validation using a linked public LTC note-claims corpus, because such a dataset is not currently available in the public domain. Instead, the empirical base of the paper combines two evidence streams.

First, the national quantitative analysis uses official federal data, including: CMS FY2025 improper payment estimates for Medicare fee-for-service, Medicare Part C, Medicare Part D, Medicaid, and CHIP; the 2025 Medicare fee-for-service CERT supplemental report; the 2025 Medicaid and CHIP supplemental improper payment report; the CMS National Health Expenditure Accounts for 2024; CMS brief summaries of Medicare and Medicaid published in November 2025; AHRQ materials on unstructured EHR data; OIG compliance guidance for nursing facilities; and HHS information on TEFCA and AI policy.

Second, the model specification is benchmarked against the peer-reviewed literature on healthcare fraud analytics, CDI, clinical NLP, and automated coding. These sources establish the feasibility of the components used in the proposed framework, even though the fully integrated LTC implementation remains a design blueprint rather than a completed trial.

Quantitative Design

The quantitative design has three layers.

1. Fiscal exposure analysis. Current official data are used to measure the scale, trend, and composition of improper payments, with special attention to documentation-related error categories and SNF-specific drivers.
2. Documentation-related loss modeling. Where the official reports quantify documentation-related components directly, those values are aggregated to estimate a conservative lower bound for directly measurable documentation-driven exposure. For FY2025, this paper uses Medicare fee-for-service no-documentation and insufficient-documentation categories together with the documentation-derived share of Medicaid improper payments.
3. Scenario-based savings analysis. The paper models the annual savings associated with 10%, 20%, 30%, and 40% reductions in directly quantified documentation-related exposure. These are scenario estimates rather than experimental results. They are intended to show the order of magnitude of the opportunity if documentation-aware review tools are implemented effectively.

Documentation-Billing Consistency Index

Let each claim or billing episode be indexed by i. Let the clinical record supply a set of documentation-derived features D_i, including diagnoses, services, severity indicators, temporal markers, and certification variables. Let the submitted claim supply a corresponding set of billed features B_i. The documentation-billing alignment for episode i can be summarized using a weighted Documentation-Billing Consistency Index (DBCI):

DBCI_i = [sum over k of w_k * m(d_ik, b_ik)] / [sum over k of w_k]

In this expression, m(d_ik, b_ik) is a matching function taking values between 0 and 1, and w_k assigns greater importance to clinically and financially material features such as primary diagnosis support, therapy intensity support, certification evidence, and timing coherence. A value near 1 indicates strong alignment; a value near 0 indicates substantial mismatch.

Hybrid Anomaly Score

The model does not rely on the DBCI alone. Claims can be fully documented yet still be statistically anomalous, and they can be statistically ordinary yet poorly documented. Therefore, the system combines documentation alignment with claims-based anomaly detection. Let X_i denote structured claims and provider features and T_i denote temporal or sequence features. A generalized anomaly risk score can be written as:

P(A_i = 1) = logistic[alpha + beta1*f(X_i) + beta2*(1 - DBCI_i) + beta3*T_i + beta4*Z_i]

Here, A_i indicates a materially suspicious or noncompliant claim, f(X_i) is a machine-learning score derived from structured claims features, T_i captures temporal irregularities, Z_i includes

provider- or facility-level context, and logistic(.) denotes the logistic link. In operational deployment, f(X_i) could come from an ensemble of XGBoost, random forest, and isolation-forest outputs.

Fiscal Savings Function

Let E_d denote directly quantified documentation-related improper payments. Let r denote the fractional reduction achieved by pre-bill and post-bill documentation-aware anomaly detection. The first-order annual gross savings estimate is:

$$S = r * E\_d$$

This is deliberately conservative because it excludes several second-order benefits: reduced audit labor, lower recoupment risk, faster claim correction, improved coder productivity, better training feedback, and documentation spillovers into programs for which documentation-related shares are not numerically reported.

Evaluation Logic

Because this is a framework paper with real national program data but without a public LTC patient-level validation set, model success is defined conceptually in two ways. At the national level, success is the amount of directly preventable documentation-related exposure identified by the analysis. At the system level, success would later be measured through prospective validation: precision of anomaly flags, reduction in false positives versus claims-only screening, time-to-review, auditor agreement, pre-bill correction rates, and net recovered or avoided improper payments. This distinction between macro-level fiscal analysis and micro-level predictive validation is essential for interpreting the contribution honestly.[20-28]

INTEGRATED NLP-MACHINE LEARNING ARCHITECTURE FOR LTC BILLING INTEGRITY

Design Principles

The architecture is designed around five principles: accessibility, modularity, interpretability, auditability, and privacy. Accessibility means using tools and workflows that smaller LTC operators and state agencies can realistically deploy. Modularity means each component can operate independently if an organization lacks the capacity for a full-stack implementation. Interpretability means every alert should be traceable to evidence. Auditability means the system should leave a review trail showing what the model saw and why it flagged a claim. Privacy means the workflow should operate under HIPAA-compliant de-identification, minimum-necessary access, and secure governance rules.

Pipeline 1: Documentation Intelligence

The first pipeline processes unstructured notes. Inputs include progress notes, nursing assessments, therapy notes, physician or nonphysician practitioner orders, MDS-linked descriptions, discharge summaries, and certification documents. The workflow proceeds in four stages.

Stage 1: Preprocessing and normalization: Text is segmented into episodes, standardized, and stripped of obvious formatting noise. Abbreviation expansion is important in LTC because facility-specific shorthand is common. Dates, author roles, note types, and document provenance should be preserved because they are part of the evidence trail.

Stage 2: Clinical extraction: A domain-adapted NLP layer identifies diagnosis mentions, treatments, services, medications, functional statements, severity cues, certification language, and temporally anchored events. ClinicalBERT-like encoders are useful for contextual representation, while rule-based layers remain valuable for highly structured payment concepts such as certification phrases, MDS-linked documentation cues, and therapy support language.

Stage 3: Support assessment: Extracted concepts are tested against billing-relevant support rules. Examples include whether the note supports the coded diagnosis at the specificity billed, whether the chart documents the skilled need for therapy, whether the recertification trail is present, whether functional limitations supporting a payment classification are explicitly charted, and whether required order language exists.

Stage 4: Documentation-derived billing profile: The result is not a full replacement claim. It is a structured profile of what the documentation appears to justify: likely diagnoses, services, severity indicators, and evidence quality. This profile is the documentation-side comparator for the cross-referencing engine.

Pipeline 2: Structured Claims Analytics

The second pipeline ingests structured claim data and provider context. Variables include ICD-10-CM diagnoses, CPT/HCPCS codes, HIPPS-related billing variables where relevant, service dates, provider types, place of service, units, charges, prior denial or edit history, claim submission timing, resubmission activity, and peer-group utilization patterns.

A hybrid detection strategy is recommended. Supervised models such as XGBoost and random forests are effective when historical outcomes or adjudication results are available. Unsupervised methods such as isolation forests are important for surfacing novel or low-frequency anomalies that do not resemble prior labeled cases. Temporal features should be included because documentation-billing mismatch often has a sequence signature: clustered end-of-period billing, service dates without contemporaneous notes, or recertification timing that consistently lags expected windows.

Cross-referencing engine

The cross-referencing engine is the centerpiece of the framework. It compares the documentation-derived billing profile with the actual claim along four axes.

1. **Diagnosis support:** Do the documented assessments, physician narratives, and MDS-linked evidence substantiate the billed diagnoses?
2. **Service support:** Is the billed service explicitly or implicitly documented in the chart?
3. **Severity and intensity alignment**: Do functional limitations, therapy minutes, acuity indicators, or case-mix elements support the billed payment level?
4. **Temporal coherence:** Does the charted timeline justify

the service date and billing episode, or does the record suggest retrospective or incomplete support?

The engine produces a DBCI and a set of evidence-linked mismatches. For example, a claim may receive a low DBCI because the primary diagnosis appears on the claim but not in contemporaneous notes, the order is missing, and the chart lacks support for the functional score that drives classification. This is much more useful operationally than a generic anomaly score because it tells reviewers exactly where to look.

Explainability Layer and Human Review

Compliance workflows require more than classification. They require usable work products. Therefore, the output should include: a risk score; the mismatch category; the relevant note excerpts; the billed elements in question; a short explanation in plain compliance language; and a recommended next action such as "request supporting documentation," "downgrade code," "hold claim," "escalate to compliance," or "educate department."

This human-review layer is central to adoption. In practice, many facilities will accept a model that catches obvious documentation gaps and helps prioritize review, even if it is not perfect. Fewer will trust a black-box model that cannot explain itself. Explainable outputs also help transform anomaly detection into process improvement. If a facility repeatedly sees alerts around missing recertifications, absent diagnosis support in MDS-linked records, or weak functional-score documentation, the model becomes a training and governance tool rather than a narrow policing tool.

Why LTC is an Especially Suitable Use Case

LTC is particularly suitable for this model because the payment logic and documentation burden are tightly linked. Skilled nursing claims, therapy records, assessments, and recertification requirements all leave a narrative footprint. Many common error causes in the CERT report are exactly the types of deficiencies that NLP and rules-based evidence checks can find. In other words, the sector's pain points are not opaque to computation; they are buried in text and process fragmentation.

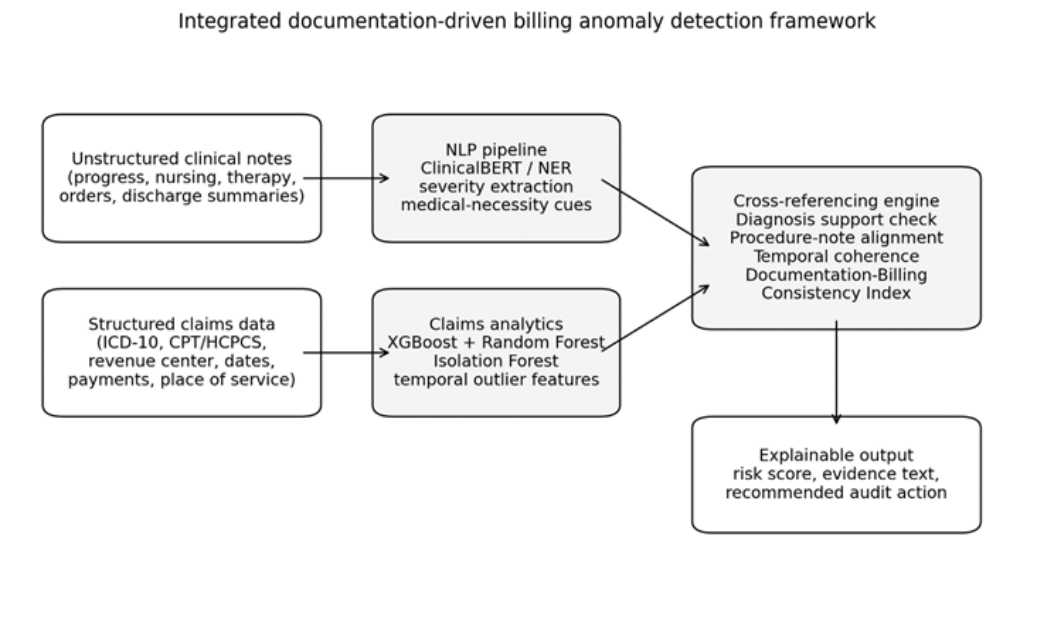

Figure 3. Integrated documentation-driven billing anomaly detection framework

Note. The proposed system links documentation intelligence, claims analytics, and evidence-based review outputs.

QUANTITATIVE FINDINGS AND SCENARIO ANALYSIS

National Improper Payment Exposure in FY2025

The first empirical result is the sheer scale of exposure. CMS reported FY2025 improper payments of $28.83 billion for Medicare fee-for-service, $23.67 billion for Medicare Part C, $4.23 billion for Medicare Part D, $37.39 billion for Medicaid, and $1.37 billion for CHIP. Summed across these five programs, the measured exposure reached $95.49 billion. Medicaid represented the largest single dollar amount, followed by Medicare fee-for-service and Medicare Part C.

The second result is that documentation problems sit at the center of this exposure. In Medicaid, 77.17% of FY2025 improper payments were due to insufficient documentation. In Medicare fee-for-service, insufficient documentation accounted for 51.5% of overall improper payments and no documentation for another 11.7%. Thus, in Medicare fee-for-service alone, documentation-related categories represented approximately $18.8 billion out of roughly $29.7 billion in unadjusted improper payments. At the national program level, the evidence strongly suggests that the payment-integrity problem is not purely about unusual utilization or fraud in the narrow sense; it is heavily shaped by information quality and evidentiary support.

Trends, FY2021-FY2025

Trend analysis shows that improper payment rates evolved differently across programs. Medicare fee-for-service fluctuated within a relatively narrow range, from 6.26% in FY2021 to 7.66% in FY2024 before declining to 6.55% in FY2025. Medicare Part C fell sharply from 10.28% in FY2021 to the 5%–6% range and then increased modestly to 6.09% in FY2025. Medicare Part D rose steadily from 1.33% in FY2021 to 4.00% in FY2025. Medicaid experienced the largest swing, declining from 21.69% in FY2021 to 5.09% in FY2024 before edging up to 6.12% in FY2025, partly reflecting post-pandemic eligibility and verification dynamics.

These patterns matter for model design. They imply that static rules are not enough. The fraud and payment-integrity environment changes with policy, eligibility redeterminations, payment-system changes, and program administration. A machine-learning layer is useful because it can adapt to new distributions and peer groups. But a documentation layer remains necessary because the reason for many program-level shifts is not simply behavioral change by providers; it is also change in how documentation and verification are evaluated.

Long-term Care Exposure and Skilled Nursing Findings

The LTC-specific evidence is particularly striking. Nursing care facilities and continuing care retirement communities absorbed $219.9 billion in national spending in 2024. Medicaid paid $69.7 billion for nursing facility services alone, equivalent to 30.5% of national nursing facility spending. On the Medicare side, skilled nursing facilities generated a projected $4.3 billion in FY2025 improper payments with an 11.8% improper payment rate.

The detailed SNF root causes reinforce the central thesis of this paper. The highest-frequency SNF root cause in the CERT sample was a HIPPS level changed based on submitted documentation (327 sample claims), followed by inadequate case-mix-group documentation (176), missing orders (160), missing support for primary or active diagnoses reported on the MDS (145), and missing case-mix-group documentation (125). Additional root causes involved missing or inadequate support for PT/OT or nursing functional scores and inadequate physician or nonphysician practitioner certification/recertification.

These are exactly the types of failures that a documentation-aware model is meant to intercept. None of them can be fully resolved by looking at a claim line or a code frequency table in isolation. They require reading the chart.

Conservative Lower-Bound Estimate of Directly Quantified Documentation-Driven Exposure

To avoid overstating the case, the paper computes a conservative lower bound using only documentation-related amounts that are directly quantified in official sources. For Medicare fee-for-service, the no-documentation and insufficient-documentation categories total approximately $18.8 billion. For Medicaid, 77.17% of $37.39 billion yields approximately $28.85 billion attributable to insufficient documentation. Together, these directly measurable documentation-related components amount to about $47.65 billion.

This lower-bound estimate does not include documentation-related exposure in Medicare Part C or Part D because the current official sources describe documentation as a dominant driver but do not provide a simple percentage on the summary pages used here. It also excludes any cost associated with provider labor, appeal cycles, post-payment audits, external consultants, delayed cash flow, or recoupment management. In that sense, it is more useful as a floor than as a full estimate.

Scenario Savings Analysis

Using the fiscal savings function \(S = r \times E_d\), the paper models annual gross savings under different reduction assumptions applied to the directly quantified documentation-related exposure of $47.65 billion.

- At a **10%** reduction, annual savings would be about **$4.77 billion**.
- At a **20%** reduction, annual savings would be about **$9.53 billion**.
- At a **30%** reduction, annual savings would be about **$14.30 billion**.
- At a **40%** reduction, annual savings would be about **$19.06 billion**.

The 10% case is the most defensible short-run benchmark for practical pilots because it assumes modest success and no dramatic behavioral change. Even that conservative case implies a multibillion-dollar national opportunity. The 20% and 30% scenarios are plausible over longer horizons if documentation-aware anomaly detection is paired with coder education, template redesign, physician feedback, therapy-documentation standardization, and recurring compliance audits.

Interpretation

The economic interpretation is important. A documentation-aware anomaly detection system does not need to eliminate fraud to justify itself. It only needs to reduce a fraction of documentation-related improper payments, shorten review time, and improve claim support enough to lower losses and administrative burden. Given the size of the documented exposure, the return on such systems could be substantial even when deployed initially as a targeted pre-bill review tool in high-risk categories such as SNF claims, therapy-related services, and diagnosis support for payment classification.

The results therefore support a stronger claim than the original conceptual paper: documentation-driven anomaly detection is not merely technically feasible; it targets one of the largest identifiable error mechanisms in current public program payment integrity.

Table 4. Top skilled nursing facility root causes in CERT 2025

Top SNF root cause (CERT 2025)	Sample claim count	Primary implication
HIPPS level changed based on documentation submitted	327	Payment classification changed because submitted documentation did not support the billed HIPPS level.
CMG component documentation inadequate	176	Case-mix-group evidence existed but was inadequate for payment support.
Order missing	160	Core ordering documentation was missing from the record.
Diagnosis support on MDS missing	145	Diagnosis reported on the MDS lacked sufficient support in the chart.
CMG component documentation missing	125	Required case-mix-group documentation was absent.
PT/OT or nursing functional scores missing	98	Functional scoring used for therapy or nursing support lacked evidence.
PT/OT or nursing functional scores inadequate	94	Functional-score evidence existed but was not adequate.
Physician/NPP certification inadequate	66	Certification or recertification language was insufficient.

Note. Sample claim counts come from the Medicare fee-for-service CERT supplemental report for the skilled nursing facility driver.

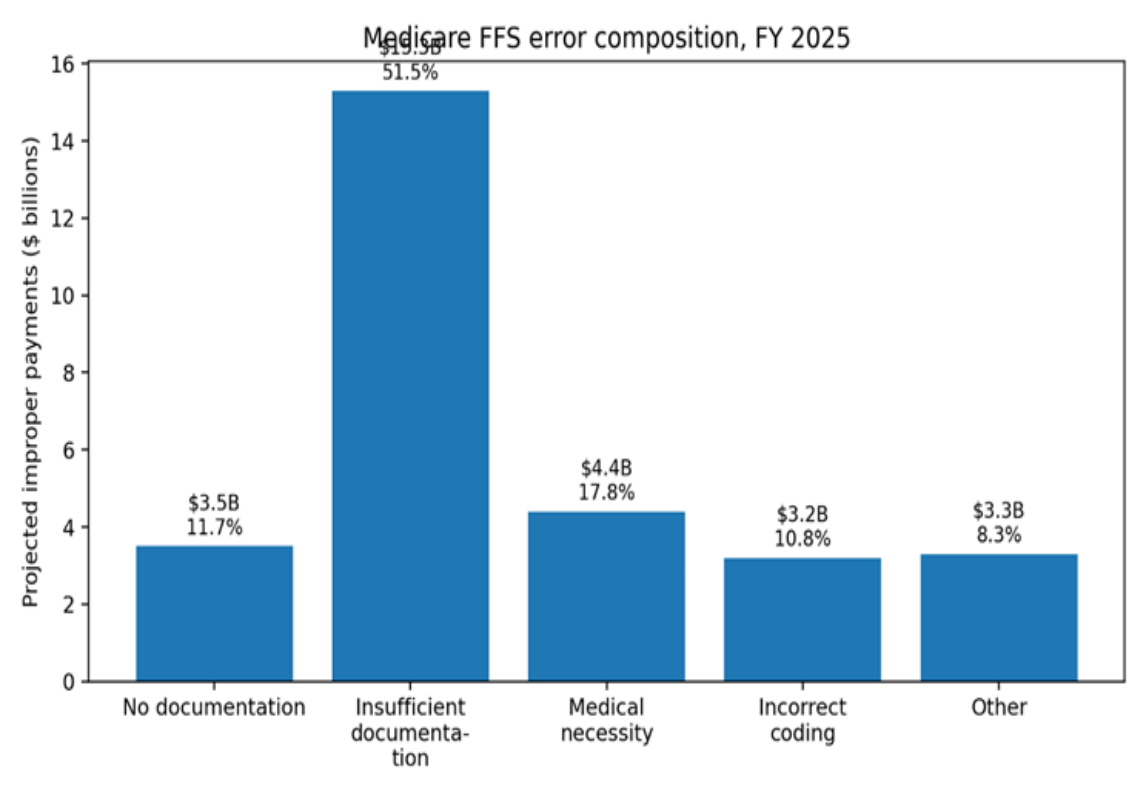

Figure 4. Medicare fee-for-service error composition, FY2025

Note. Dollar amounts use error-category totals from the CERT supplemental report.

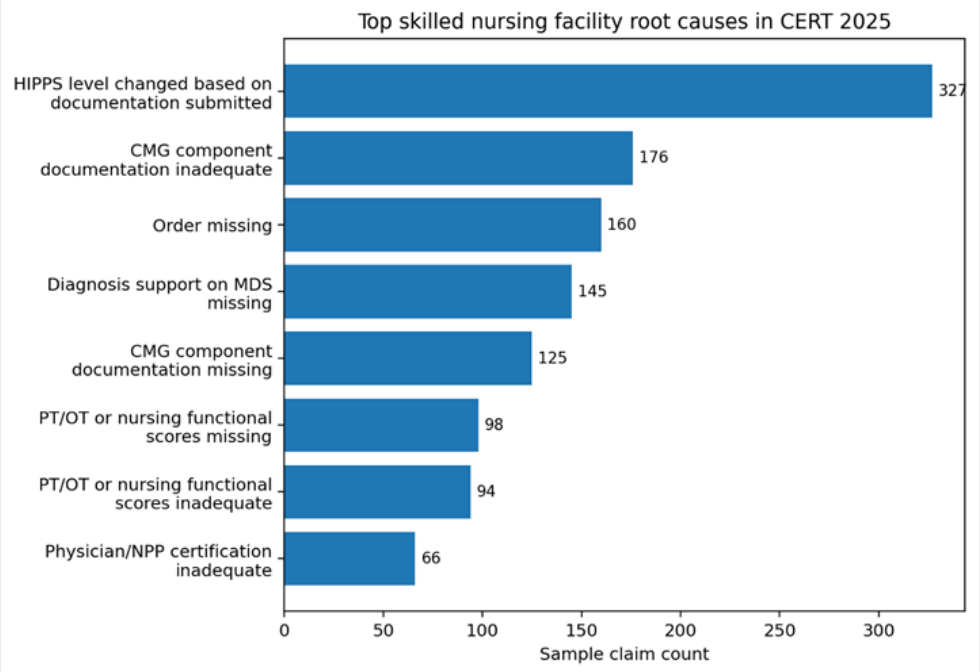

Figure 5. Top skilled nursing facility root causes in CERT 2025

Note. Counts represent sample claim frequencies rather than projected national dollars.

Table 5. Scenario-based annual savings from reducing directly quantified documentation-related improper payments

Scenario	Reduction applied to directly quantified documentation-related exposure	Estimated gross annual savings ($ billions)	Interpretation
Conservative	10%	4.77	Illustrative savings from reducing documentation-driven improper payments in Medicare FFS and Medicaid only.
Moderate	20%	9.53	Illustrative savings from reducing documentation-driven improper payments in Medicare FFS and Medicaid only.
Strong	30%	14.3	Illustrative savings from reducing documentation-driven improper payments in Medicare FFS and Medicaid only.
Transformational	40%	19.06	Illustrative savings from reducing documentation-driven improper payments in Medicare FFS and Medicaid only.

Note. Savings are calculated from a direct documentation-related exposure base of $47.65 billion.

OPERATIONAL IMPLEMENTATION IN LONG-TERM CARE FACILITIES

Workflow integration

The most practical deployment model for LTC is phased integration. A facility should not begin with autonomous claim holds across all services. It should begin with a narrow, high-yield pilot. Skilled nursing claims, therapy-intensive episodes, recertification-heavy cases, and diagnoses known to be sensitive in payment classification are natural candidates. The system can run as a pre-bill review layer that generates ranked worklists for coders and compliance staff.

An effective workflow would typically have four stages: (1) ingest documentation and claims from the EHR, billing, and MDS workflows; (2) score episodes using the hybrid anomaly framework; (3) present evidence-linked alerts to a human reviewer; and (4) capture reviewer outcomes to retrain thresholds and reduce noise.

Over time, the same system can support retrospective audits, appeal preparation, and education dashboards.

Governance and Privacy

Any real implementation must be governed carefully. Documentation analysis necessarily touches protected health information. Facilities therefore need role-based access, secure data pipelines, logging, retention rules, and model-monitoring procedures. Minimum-necessary principles should be enforced so that reviewers see only what is needed for the specific payment-integrity task. Where feasible, model development should occur on de-identified or limited datasets, with re-identification restricted to authorized compliance operations.

Governance must also address fairness and over-alerting. Facilities serving residents with severe complexity, behavioral comorbidity, or frequent transitions may naturally look different from peer averages. That is why the proposed system combines statistical anomaly detection with documentary evidence rather than relying only on outlier scores. The model should inform investigation, not replace judgment.

Human Capital and Organizational Learning

One of the underappreciated benefits of a documentation-aware system is that it produces structured feedback. If alerts consistently cluster around certain note types, disciplines, or units, the organization can respond with focused interventions. For example, if diagnosis support gaps are concentrated in MDS-linked records, training can focus there. If certification failures are concentrated in therapy episodes, the workflow can be redesigned. If missing orders dominate a subset of claims, order-entry controls can be strengthened. This turns anomaly detection into a learning system.

Vendor and Payer Applications

The same architecture can be adapted for payers, managed-care organizations, Medicaid agencies, and external audit vendors. Payers may emphasize provider-level risk stratification and prepayment review. Facilities may emphasize pre-bill correction and coding support. State Medicaid agencies may use the system to prioritize technical assistance and targeted audit resources. Because the underlying logic is evidence alignment, the system is flexible across governance contexts.

DISCUSSION

This paper advances a simple but consequential argument: payment integrity in LTC should be treated as an information-alignment problem, not merely a statistical claims problem. Official CMS data already show that documentation failures account for a large share of measured improper payments. The Medicare fee-for-service CERT report places insufficient documentation at the center of national error categories and identifies skilled nursing facilities as major improper-payment drivers. Medicaid reporting goes further by quantifying the documentation burden directly. In that environment, it is analytically inefficient to ignore the clinical text that gives claims their meaning.

The proposed framework matters because it closes the logic gap between coding and evidence. Claims-based models can tell us that a billing pattern is unusual. A documentation-aware system can tell us whether it is unusual because it lacks support, because the code appears too aggressive for the chart, because key supporting records are absent, or because the timeline is inconsistent. That additional information is exactly what compliance teams need.

The paper also contributes economically by translating documentation quality into fiscal magnitude. The conservative lower-bound estimate of $47.65 billion in directly quantified documentation-related exposure is not an abstract system failure. It is a large, measurable policy opportunity. Even a 10% reduction would be material at the federal level. Moreover, the true economic value is likely higher because the estimate excludes Medicare Advantage documentation failures, Part D sponsor documentation issues, internal labor costs, and the deterrence effect of better controls.

At the same time, the paper is intentionally cautious. It does not claim that a proposed model has been prospectively validated on a public LTC note-claims dataset. It does not claim that all documentation-related improper payments are preventable. It does not collapse documentation error into fraud. These distinctions are essential for credibility. A facility may have a proper service but poor records. Another may have inflated coding. Another may simply have weak internal controls. A high-quality detection system should help separate these realities rather than conflate them.

Limitations

Three limitations deserve emphasis. First, the absence of a public linked LTC note-claim dataset means the integrated model cannot be fully validated in this paper. Second, some official program summaries identify documentation as a major root cause without publishing a clean headline percentage for every program, which constrains national aggregation. Third, documentation-aware systems are only as good as the data they receive. If source records are fragmented, scanned poorly, inconsistently dated, or inaccessible in structured workflows, performance will suffer.

Future Research

Future work should move from design and national fiscal analysis to prospective field evaluation. The strongest next step would be a multi-facility pilot in which de-identified LTC notes and claims are linked, manually adjudicated, and used to evaluate pre-bill and post-bill detection performance. That study should compare claims-only models against documentation-aware models on precision, recall, false-positive burden, auditor agreement, avoided improper payments, and time-to-resolution. Additional work should also test retrieval-augmented and large-language-model architectures for note evidence extraction, provided their outputs are constrained and auditable.

CONCLUSION

Public payment integrity in Medicare and Medicaid cannot be strengthened adequately if documentation and billing remain analytically separate. In long-term care, where reimbursement logic depends heavily on diagnoses, severity, therapy support, certifications,

and temporally coherent records, the separation is especially costly. Current official data show that documentation-related failures already account for a very large share of measured improper payments. Skilled nursing facility findings from the CERT program make the issue concrete: the highest-frequency root causes are largely documentation and support failures, not exotic fraud schemes.

This paper has shown, using current federal data, that documentation-aware anomaly detection has both technical and economic justification. A hybrid system that combines clinical NLP, claims-based machine learning, and an evidence-alignment engine can target the exact class of errors that official oversight reports repeatedly identify. The conservative lower-bound estimate of directly quantified documentation-related exposure in FY2025 Medicare fee-for-service and Medicaid is approximately $47.65 billion. That figure alone is sufficient to justify serious investment in documentation-aware controls, pilot programs, and governance structures.

The larger implication is strategic. The future of payment integrity will not be won only by better claims outlier models. It will be won by systems that can read, compare, explain, and learn from the clinical evidence that underlies payment. Long-term care facilities, Medicaid agencies, and Medicare oversight bodies should treat documentation-grounded anomaly detection as a core component of modern compliance infrastructure.

ETHICAL APPROVAL

This article does not contain any studies with human participants or animals performed by the author.

ACKNOWLEDGEMENT

None

FUNDING

The author received no financial support for the research, authorship, and/or publication of this article.

CONFLICT OF INTEREST

The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

1. Ahmed M, Mahmood AN, Islam MR. A survey of anomaly detection techniques in financial domain. Future Generation Computer Systems. 2016;55:278-288.
2. Bauder RA, Khoshgoftaar TM, Seliya N. A survey on the state of healthcare upcoding fraud analysis and detection. *Health Services and Outcomes Research Methodology.* 2017;17(1):31-55.
3. Johnson JM, Khoshgoftaar TM. Medicare fraud detection using neural networks. *Journal of Big Data.* 2019;6:63.
4. Anli du Preez A, Bhattacharya S, Beling P, Bowen E. Fraud detection in healthcare claims using machine learning: A systematic review. *Artif Intell Med.* 2025;160:103061.
5. Davis J, Shepheard J. Clinical documentation integrity: Its role in health data integrity, patient safety and quality outcomes and its impact on clinical coding and health information management. *Health Inf Manag.* 2024;53(2):53-60.
6. Sanderson AL, Lo L, Arafiles W, et al. The Impact of Clinical Documentation Integrity Programs on Diagnosis Documentation. *Adv Health Inf Sci Pract.* 2025;1(2):CGJA8827.
7. Hossain E, Rana R, Higgins N, et al. Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review. *Comput Biol Med.* 2023;155:106649.
8. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019; pp.72-78.
9. Dong H, Suárez-Paniagua V, Tikk D, et al. Automated clinical coding: What, why, and where we are? *NPJ Digital Medicine.* 2022;5:159.
10. Agency for Healthcare Research and Quality. Challenges and opportunities for improvement. 2024.
11. Centers for Medicare & Medicaid Services. Fiscal year 2025 improper payments fact sheet. 2026a.
12. Centers for Medicare & Medicaid Services. Medicare fee-for-service supplemental improper payment data, 2025. 2026b.
13. Centers for Medicare & Medicaid Services. Medicare Part C improper payment measurement. 2026c.
14. Centers for Medicare & Medicaid Services. Medicare Part D improper payment measurement. 2026d.
15. Centers for Medicare & Medicaid Services. 2025 Medicaid and CHIP supplemental improper payment data. 2026e.
16. Centers for Medicare & Medicaid Services. National health expenditure accounts: 2024 highlights. 2026f.
17. Centers for Medicare & Medicaid Services. NHE fact sheet. 2026g.
18. Centers for Medicare & Medicaid Services. Brief summaries of Medicare & Medicaid. 2026h.
19. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. *Journal of Big Data.* 2014;7:113.
20. Li J, Huang KY, Jin J, Shi J. A survey on statistical methods for health care fraud detection. Health Care Management Science. 2008;11(3):275-287.
21. Office of Inspector General. Nursing facility industry segment-specific compliance program guidance. 2024.
22. Office of Inspector General. Nursing homes. 2026.
23. U.S. Department of Health and Human Services. HHS unveils AI strategy to transform agency operations. 2025a.
24. U.S. Department of Health and Human Services. Request for information: Accelerating the adoption and use of artificial intelligence as part of clinical care. 2025b.

25. U.S. Department of Health and Human Services. TEFCA, America's national interoperability network, reaches nearly 500 million health records exchanged. 2026.

26. Woo BFY, Cato K, Cho H, You SB, Song J. The use of large language models in clinical documentation: A scoping review. *International Journal of Nursing Studies*. 2025;176:105322.

27. Yu H, Fan L, Li L, et al. Large Language Models in Biomedical and Health Informatics: A Review with Bibliometric Analysis. *J Healthc Inform Res.* 2024;8(4):658-711.

28. Zhang Y, Lyu C, Chang L, et al. A systematic review of automated International Classification of Diseases coding models using the Medical Information Mart for Intensive Care dataset. *Digit Health.* 2025;11:20552076251404518.