Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Automating Clinical Validation in Claims Adjudication: An NLP/LLM Systems Approach

Triveni Kolla

Marist College, USA

ARTICLE INFO

Received: 08 Nov 2025

Revised: 12 Dec 2025

Accepted: 20 Dec 2025

ABSTRACT

Healthcare claims adjudication suffers a lot of challenges in its operations because of manual clinical validation processes that use huge resources and yield inconsistent results when applied to different populations of reviewers. The use of Natural Language Processing and Large Language Models to automate clinical validation offers a transformational potential for payment integrity activities in payer organizations. Domain-adapted models of language models, such as BioBERT and ClinicalBERT, exhibit advanced performance in the task of clinical entity detection, medical terminology comprehension, and documentation pattern reasoning that is typical of real-world healthcare delivery. The introduction of the implementation strategies that focus on the gradual implementation, human-AI interaction, and ongoing quality control allows integrating the novel approach into the existing adjudication processes without deteriorating the quality of the results that are needed to comply with the regulations. The deployed systems are shown to have significant gains in processing efficiency, consistency of decisions, and the productivity of reviewers in comparison to traditional manual validation techniques. The features of enhanced explainability create transparent justifications to support appeals management and audit requests and establish trust among the clinical staff. The technology can resolve the underlying capacity issues of payer organizations by facilitating the processing of greater claim volumes without similarly large projections of staffing. Financial gains are attained in the form of lower processing fees, lower rate of appeals, and reduced adjudication time, which enhances the satisfaction of stakeholders. The development of such systems proceeds via the increased training data, improved architectures, and experience of operation that brings the ability to solve more and more complicated validation problems. The smart fusion of the sophisticated computational systems with human clinical knowledge generates validation workflows that will reach performance levels that are not feasible by either of the preceding methods.

Keywords: Natural Language Processing, Large Language Models, Claims Adjudication, Clinical Validation, Healthcare Automation

1. Introduction

Healthcare claims adjudication is among the most labor-intensive operations in payer operations, and clinical validation is one of the key bottlenecks of such operations and influences millions of claims every year. To ensure that clinical validation is a major factor in payment integrity issues, the Centers for Medicare and Medicaid Services has developed holistic programs of improper payment measurement that identify and mitigate payment errors in a wide range of healthcare programs [1].

These surveillance programs are based on the application of statistical sampling techniques and stringent review procedures to determine the accuracy of payments made under Medicare, Medicaid, and other federally funded healthcare programs and generate the necessary data that measures the extent of validation issues in the industry.

Conventional methods of medical record review in paper-based format require a large number of human resources, add variability to the decision-making process, and cause great delays in payment processing, which spreads throughout the healthcare ecosystem. Due to the sophistication of the current healthcare documentation, the growing complexity of the coding systems, and the changing coverage policies, the difficulties in maintaining the accuracy of claims adjudication have also increased. The processing cost benchmarks indicate a significant difference in the amount of resources used to handle various types of claims and different organizational environments, with complex claims that involve clinical validation taking significantly more resources than the typical administrative processing [3]. The demand to have scalable, consistent, and efficient scalable validation systems has grown in both public and private payer organizations as healthcare costs become more and more expensive and regulatory requirements continue to become stricter.

The paper discusses the use of Natural Language Processing and Large Language Models to automate clinical validation in claims adjudication, regarding how state-of-the-art computational linguistics and machine learning structures can revolutionize the way things are performed. The systems can provide a way to overcome the long-standing operational challenges and keep the accuracy and compliance standards required in healthcare financial operations by making use of advanced algorithms that can understand unstructured medical documentation, extract all the relevant clinical entities, and reason about the medical necessity criteria.

2. Current Challenges in Manual Claims Validation

Multiple inherent limitations in the manual clinical validation process affect operational efficiency and quality of adjudication in healthcare payer organizations. To identify whether claimed payments were in accordance with the medical necessity requirements and codes of conduct, clinical reviewers have to review a lot of medical documentation, such as physician notes, diagnostic reports, treatment plans, and procedural records. The programs of improper payment measurement implemented by CMS show that the percentage of error rates among various classes of services and types of providers is different, and some expensive procedures and specialized services record higher levels of payment errors due to insufficient documentation or clinical validation [1]. These measurement programs use well-planned sampling procedures and elaborate scrutiny standards that reveal systemic flaws in existing validation methods, especially on claims where critical clinical circumstances have to be looked at or where subtle coverage standards must be sought.

The cognitive load required of reviewers results in inconsistency in decision making, whereby one reviewer might arrive at one decision following a clinical scenario compared to another reviewer, depending on his or her experience, specialty background, training, or personal judgment. Studies on the cost of processing claims have noted that the manual review process would comprise a disproportionate portion of the overall administrative costs, and that there were large cost differences depending on the complexity of the claim, the level of expertise of the reviewer, and the amount of supporting documentation needed to be reviewed [3]. This variability presents downstream complexity (such as an appeal process, provider dispute, and possibly a compliance risk), as well as poor relationships between payers and healthcare delivery organizations. The financial consequences do not just limit themselves to direct processing costs but also to opportunity costs due to delays in adjudication and the resources used in managing appeals, along with possible regulatory fines that may be imposed due to systematic adjudication errors.

Moreover, the workload of claims to be checked is significantly higher than the number of clinical personnel in many payer organizations, which results in permanent bottlenecks and worsens the efficiency of the operation and the satisfaction of the stakeholders. A limited number of qualified clinical reviewers possess both the right medical qualifications and knowledge of claims adjudication, which compounds capacity, forcing them to make tough prioritization decisions that could lead to the poor review of high-risk claims or the unnecessary delays in the time-sensitive process of authorizing them. This capacity crisis is symptomatic of workforce issues in healthcare administration more generally, as specialized knowledge requirements of effective clinical validation provide obstructions to quick staffing solutions. The collision of the rising volumes of claims, the rising volume of documentation, and the scarcity of reviewers has resulted in an operational model that is unsustainable and requires new technological solutions to reach certain acceptable functioning levels.

Challenge Category	Manifestation	Operational Impact	Stakeholder Effect
Processing Time	Extended review duration for complex documentation	Delayed adjudication cycles	Provider cash flow constraints
Decision Consistency	Variable interpretations across reviewers	Elevated appeals rates	Provider administrative burden
Reviewer Capacity	Workforce shortages and high turnover	Processing bottlenecks	Patient authorization delays
Cost Structure	High administrative expense ratios	Reduced operational margins	Increased premium pressures
Documentation Complexity	Unstructured clinical narratives	Cognitive workload burden	Reviewer fatigue and errors
Policy Application	Nuanced coverage criteria interpretation	Inconsistent determinations	Regulatory compliance risks

Table 1: Manual Claims Validation Challenges and Impacts [3, 4]

3. NLP and LLM Technology architecture.

The current NLP and LLM systems are based on the use of transformer neural network structures that have proven to possess outstanding abilities in comprehending and writing human language, with applications to many fields, including biomedicine. BioBERT is a breakthrough in the field of biomedical text mining, where a pre-trained language representation model is further trained on domain-related corpora, such as PubMed abstracts and full-text articles of PubMed Central [2]. This domain adaptation procedure allows the model to learn subtle knowledge about medical vocabulary, disease-gene interactions, pharmacology, and clinical terms that would not otherwise be well captured in general-purpose language models that are solely trained on non-medical text. The architecture uses contextualized word representations that encode semantic meaning using contextual information as opposed to prior definitions, so that ambiguous medical words can be better interpreted depending on context than on predefined definitions, which could have varying meanings in a clinical specialty or documentation setting.

ClinicalBERT also further optimized language model architectures in clinical tasks, training on real clinical notes on patient intensive care unit admissions, and producing representations specifically optimized to the vocabulary, abbreviations, and patterns of documentation found in real-world clinical practice [4]. This model shows improved performance in tasks that are directly related to claims validation, such as clinical named entity recognition, medical concept extraction, identification of temporal relationships, and prediction of clinical outcome based on information recorded for a patient. The training method used in ClinicalBERT deals with the large domain gap between published biomedical literature and clinical documentation, noting that physician notes, discharge summaries, and procedural reports have different linguistic patterns, shorthand notation, and contextual conventions, all of which need a specialized model adaptation. ClinicalBERT discerns the lingo of the real-world healthcare delivery, which is reflected in the text behind insurance claims, by training on clinical narratives instead of research publications.

The proposed pipeline, with the help of LLM, combines several expert elements that collaboratively work together to process claim data in stages of processing. The document ingestion layer receives clinical documentation in many different forms, including structured electronic health records in standard formats and unstructured narratives of physicians that demand advanced parsing and scanned paper that is read using optical character recognition algorithms. These heterogeneous documents are broken down into semantically meaningful segments by the use of advanced parsing algorithms that maintain important contextual relationships to guide clinical reasoning and determination of medical necessity. Core NLP engine will recognize entities to identify clinical concepts, which may be diagnoses, procedures, medications, laboratory values, anatomy, and time to describe the clinical presentations and treatment course of a patient comprehensively.

Connection extraction algorithms transform relationships between recognized entities, building a formal model of the clinical history of the patient that can be used to make automated inferences about the medical necessity, suitability of care, and coverage policy congruence. These algorithms are required to be able to represent some complicated dependencies, such as causal relationships among symptoms and diagnoses, temporal orders of treatment interventions, and hierarchical relationships between primary conditions and complications. The LLM component subsequently gathers this mined and formatted data and compares clinical data with policy requirements, coding specifications, and clinical necessity provisions with the system's stored knowledge base and develops early validation tests with express references to pertinent documentation passages and policy provisions.

Component Layer	Primary Function	Technology Foundation	Output Products
Document Ingestion	Multi-format data acceptance	OCR and parsing algorithms	Normalized text streams
Entity Recognition	Clinical concept identification	Domain-adapted BERT models	Structured entity sets
Relationship Extraction	Semantic connection mapping	Neural relation classifiers	Clinical narrative graphs
Temporal Processing	Event sequencing and timeline construction	Temporal expression parsers	Chronological representations
Policy Reasoning	Coverage criteria assessment	Large language models	Validation recommendations
Explainability Generation	Justification articulation	Attention-based attribution	Human-readable rationales

Table 2: NLP and LLM Architecture Components [5, 6]

4. Implementation and Workflow Integration.

The effective implementation of the LLM-based validation systems must be thoroughly integrated with the current claims adjudication workflows, technical infrastructure, and organizational processes that have developed over decades of manual operation. Basic capabilities involving biomedical word embeddings and improved lexical representations make possible the capabilities of clinical terminology processing, and special-purpose vectorization methodologies more effectively preserve semantic relationships among medical concepts compared with general-purpose word embedding methods [5]. Such improved representations allow the system to identify synonymous terms, comprehend hierarchical relationships among general and specific medical terms, and interpret abbreviations and shorthand notation that is widespread in clinical documentation but not in standardized medical vocabularies. Combining the subword information with ordered medical knowledge provided by resources such as Medical Subject Headings forms powerful representations that are highly generalizing in different documentation styles and clinical settings.

Implementation strategy is usually a step-by-step strategy that starts with well-chosen use cases that have definite value propositions and have reduced risks of deployment and organizational disturbance. Clinical information extraction applications have spread to many fields of healthcare, and systematic reviews have found various applications of the technology, such as automated coding support, pharmacovigilance, clinical decision support, and quality measurement [6]. These applications indicate that NLP technologies are feasible to be implemented in actual healthcare settings where implementation strategies deal with critical issues such as compatibility with legacy systems, assurance of extraction quality, system error and maintenance, and system performance as documentation patterns change. The large number of successful extraction applications is a source of technical methods, best practices, and lessons learned that inform the implementation of claims validation.

When first implemented, initial deployments are usually focused on a particular type of claim, like prior authorization requests, high-cost procedural areas, or types of services with the highest rates of denials, where the impact of a bottleneck in the manual review process is highest and maximum improvement in efficiency is potentially the greatest. The system is designed not as an autonomous system but as an intelligent assistant that supplements the work of a human reviewer and does not fully replace the functions of a human reviewer, but performs the routine validation processes. In the simplest situations, where clinical records are clear in support or opposition of policy demands, the LLM produces high-confidence reports that may be immediately accepted by reviewers with minimal further research needed; a large and significant amount of time is saved as opposed to manual analysis of documents. In cases that are more complex or ambiguous, the system will offer structured feeds of pertinent clinical data, areas where human judgment is necessary, and possible policy interpretation issues, which, in effect, triage work on review, according to the level of case complexity and confidence levels.

Connection to claims management systems can be achieved via standardized APIs, which allow two-way data flow, with the system receiving claim submissions along with related clinical documentation, running them through the NLP and LLM pipeline, and returning structured validation assessments, which are processed into review queues and decision support interfaces. The quality assurance mechanisms are put in the workflow, and random sampling of system recommendations is subject to an expert review to ensure the accuracy, any possible model drift, and systematic error that might manifest as documentation patterns or as policy criteria change. Feedback loops record any adjustments made by reviewers on system recommendations, which form useful training data, maintaining the model refinement process and advancement with organizational policies and priorities.

Implementation Phase	Activities	Integration Points	Quality Mechanisms
Use Case Selection	Target high-impact claim types	Claims management systems	Performance baseline establishment
Pilot Deployment	Limited production testing	Review queue interfaces	Expert validation sampling
Workflow Integration	API development and data flow	Legacy system connections	Inter-rater reliability monitoring
Reviewer Training	System interaction protocols	User interface deployment	Feedback collection processes
Scaling Operations	Expanded claim type coverage	Enterprise platform integration	Continuous accuracy assessment
Continuous Improvement	Model refinement cycles	Machine learning pipelines	Drift detection and correction

Table 3: Implementation and Integration Strategies [7, 8]

5. Results and Performance Impact.

Initial implementations of early LLM-based clinical validation systems have led to significant gains in various areas of operation that influence payer performance, financial performance, and stakeholder satisfaction. Neural network-based methods utilized in the healthcare domain have demonstrated a strong potential in a wide range of applications. Systematic reviews have found this pattern of results to be consistent in terms of performance improvement once neural network models are appropriately adapted to healthcare-specific problems, such as quality issues with data, regulatory issues, and the necessity to ensure the interpretability of results [8]. These reviews focus on the need to pay close attention to the validation methodologies, bias detection and mitigation, clinical workflow integration, and continuous monitoring in order to achieve sustained performance when implementation of AI in healthcare takes place as input data distributions change over time. Deep learning solutions to claims validation have the advantages of experience with other fields of AI in healthcare, as well as coping with special issues with financial decisions and regulatory compliance needs.

Deep learning implementations at scale with electronic health records have reported impressive accuracy levels on predictive datasets such as diagnosis, length of stay prediction, and mortality prediction, and are showing that neural network models can successfully scale up to high-dimensional healthcare data. These applications used specially developed data preprocessing pipelines, model architectures that are specialized to temporal clinical data, and validation strategies that evaluate the performance on a variety of patient groups and clinical contexts. The methods designed to analyze large-scale EHR can be applied to scalable claims validation applications, where data heterogeneity, time constraints, and issues with processing millions of records create requirements similar to those of scalable architecture. The proven high predictive capability and minimal computational cost to ensure real-time or near-real-time capability to run a validation system give reason to believe that LLM-based validation systems can support the throughput of a production claims adjudication environment.

In deployed systems, accuracy measures show that validation based on LLM is highly in agreement with expert human assessments on both standard types of cases, and the accuracy of validation changes depending on the complexity of claims, the completeness of documentation, and the specific coverage policies applicable. Studies of clinical documentation and prediction tasks have determined that machine learning models can perform at levels as good as human experts on well-defined clinical tasks when trained on large and representative datasets [10]. Nevertheless, these articles also underline the necessity of strict validation on held-out test sets, exploring the model calibration and uncertainty quantification, and measuring performance on subpopulations of demographics to identify possible biases. The implementation of claims validation should apply the same strict evaluation system, and continuous controls should be used to make sure that the performance of the system is within reasonable limits, as the type of claims, documentation activities, and policy standards change over time.

The explainability capabilities added to the architecture of modern LLM models have especially been useful when being used in appeals or in regulatory audits, where the documentation of the reasoning underpinning adjudication decisions, or in reviewing challenged determinations. These systems leave audit trails more transparent and accountable than the unaccounted algorithmic decisions through the production of explicit justifications that reference particular passages in clinical documentation and refer to relevant policy provisions. Analysis of the workload of the reviewer shows that AI-assisted validation allows clinical personnel to devote their skills to truly complicated situations that need delicate clinical judgment, and the routine cases can be identified significantly faster by using automated initial assessments. Depending on their organization, organizations that have installed such systems experience remarkable increases in viable validation capacity without corresponding increases in staffing, as well as an increase in job satisfaction among the reviewers because repetitive documentation review work is fully automated, and time spent by the staff is redirected to other professionally engaging work that requires expert clinical judgment.

Performance Dimension	Traditional Manual Process	AI-Assisted Process	Improvement Domain
Processing Duration	Extended multi-day cycles	Accelerated same-day completion	Operational efficiency
Decision Consistency	Moderate inter-rater agreement	High algorithmic uniformity	Quality standardization
Reviewer Throughput	Limited daily case volume	Substantially increased capacity	Productivity enhancement
Appeals Frequency	Elevated contestation rates	Reduced dispute incidence	Decision quality
Accuracy Performance	Variable by reviewer expertise	Consistently high agreement levels	Clinical validation reliability
Financial Impact	High administrative cost burden	Significant expense reduction	Economic sustainability

Table 4: Performance Outcomes and Impact Metrics [9, 10]

Conclusion

The adoption of Natural Language Processing and Large Language Model as a component of the clinical validation of claims adjudication is a paradigm shift in the healthcare payment integrity operations, fixing the problematic aspects of the processes, and improving the quality and consistency of decisions. Language models that have been domain-adapted based on biomedical literature and clinical documentation show advanced knowledge of medical terminology, clinical reasoning, and documentation conventions required to provide the right interpretation of supporting evidence provided when insurance claims are submitted. The gradual introduction of implementation plans that focus on human-AI cooperation and do not rely on full automation can help organizations achieve significant performance improvements without losing proper control over the complex cases that need professional clinical decision-making. Implemented systems deliver better processing speeds, consistency, and accuracy rates that match or outperform historical manual validation models with and are able to generate a measurable value through a lower cost of administration, shorter adjudication timeframes, and enhanced satisfaction among stakeholders in both providers and patients. The explainability features incorporated into the current architectures bring about clear justifications that enable compliance with regulation, ease the control of appeals, and establish trust among clinical reviewers who are dependent on the system recommendations in their daily operations. Through the effective introduction of these technologies, organizations are placed in a better position to handle the increased number of claims arising due to the increasing number of people covered and the rising number of people utilizing healthcare services without a commensurate increase in the number of reviewers, solving the inherent capacity limitations that have long afflicted the operations of payers. The financial gains include direct processing cost savings, reduction in the appeals handling costs, better payment accuracy leading to less compliance risk, and operational efficiency that liberates clinical capabilities to do more high-value activities that require human judgment. As these systems keep maturing with more training data, with different documentation styles and clinical cases, fine-tuned model structures with the new developments on transformer-based language understanding, and a wealth of operational experience that guides their daily improvement process, they will be able to do more and more sophisticated validation cases, and other related revenue cycle management services. Intelligent cooperation between sophisticated computational systems that can handle information scaleably and consistently, efficiently, and human clinical specialists who can offer subtle judgment on dubious cases, contextual analysis of atypical circumstances, and ethical regulation of consequential financial choices to patient care is the future of claims adjudication.

References

[1] Centers for Medicare & Medicaid Services, "Improper Payment Measurement Programs," 2024. [Online]. Available: https://www.cms.gov/data-research/monitoring-programs/improper-payment-measurement-programs

[2] Jinhyuk Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, 2020. [Online]. Available: https://academic.oup.com/bioinformatics/article/36/4/1234/5566506

[3] APQC, "Total cost to perform healthcare claims processing (excluding benefits and claims expense) per FTE that performs the process group "adjudicate claims and process reimbursement"," [Online]. Available: https://www.apqc.org/what-we-do/benchmarking/open-standards-benchmarking/measures/total-cost-perform-healthcare-claims

[4] Kexin Huang et al., "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," arXiv:1904.05342, 2019. [Online]. Available: https://arxiv.org/abs/1904.05342

[5] Yijia Zhang et al., "BioWordVec, improving biomedical word embeddings with subword information and MeSH," Scientific Data, 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31076572/

[6] Yanshan Wang et al., "Clinical information extraction applications: A literature review," Journal of Biomedical Informatics, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046417302563

[7] Junaid Bajwa et al., "Artificial intelligence in healthcare: transforming the practice of medicine," Future Healthc J. 2021. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8285156/

[8] Riccardo Miotto et al., "Deep learning for healthcare: review, opportunities and challenges," Briefings in Bioinformatics, 2017. [Online]. Available: https://academic.oup.com/bib/article/19/6/1236/3800524

[9] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," npj Digital Medicine, 2018. [Online]. Available: https://www.semanticscholar.org/paper/Scalable-and-accurate-deep-learning-with-electronic-Rajkomar-Oren/1f7eec4c76963a4ba7516ca00e6a2f855667b3f2

[10] Thomas H McCoy et al., "A clinical perspective on the relevance of research domain criteria in electronic health records," American Journal of Psychiatry, 2015. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25827030/