AI-Driven DRG Validation in Healthcare RCM: Challenges, Solutions, and the Path Forward

Applied Doctoral Project Manuscript

Submitted to National University

School of Health Professions

In Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF HEALTH ADMINISTRATION

by

TIFFANY K. REEVES

San Diego, California
April 2026

Abstract

Artificial intelligence-enabled natural language processing tools are increasingly used in diagnosis-related group validation workflows to improve coding accuracy and reduce revenue leakage. However, there is no clear evidence of the types of errors these systems generate and the organizational factors that influence their performance. The purpose of this applied doctoral project was to understand the nature, frequency, and operational impact of errors generated by artificial intelligence-driven diagnosis-related group validation tools and how human and organizational factors influence their identification and management. The project was guided by the Human Artificial Intelligence Integration Framework.

A qualitative descriptive design was employed, combining structured artifact analysis with thematic analysis of stakeholder interviews. A structured audit of 40 synthetic inpatient cases was conducted to assess differences between artificial intelligence-generated recommendations and manual coding validation. Root cause analysis was performed for different types of errors, and the operational and financial impacts were assessed. Ten semi-structured interviews were conducted with stakeholders in coding, clinical documentation improvement, auditing, and revenue cycle leadership.

The audit results showed that discrepancies occurred most often in clinically complex cases, particularly respiratory and sepsis-related cases. False positives were detected in 50% of cases and were the most common discrepancy types. Thematic analysis highlighted common elements: the need for human validation, limitations in contextual interpretation, documentation inconsistencies, workflow burden, and the importance of feedback loops. Artificial intelligence tools can help with case identification, but human review is needed to ensure correctness and compliance. We suggest enhancing human validation processes, standardizing the

documentation, making the AI logic transparent, and developing feedback loops. As such, we provide practical guidance for integrating artificial intelligence into revenue cycle workflows whilst maintaining coding accuracy, operational efficiency, and compliance.

Acknowledgments

I would first like to express my sincere gratitude to my committee chair, Dr. Rodney McCurdy, for his guidance, encouragement, and steady support throughout this doctoral journey. His mentorship, thoughtful feedback, and willingness to provide both structure and flexibility were instrumental in bringing this project to completion. I am also grateful to my subject matter expert, Dr. Mountasser Kadrie, for his careful review and insightful recommendations that strengthened this work and enhanced its clarity. In addition, I would like to thank Dr. Brian Allen and Dr. Lorie Shoemaker, my Academic Readers, for their time, thoughtful review, and guidance in strengthening this project. I would like to acknowledge Zac for generously spending hours of his personal time assisting with the technical aspects of my audit. This project would not have been possible without his expertise and willingness to help. I am also deeply appreciative of Candice for her constant support and for the time and flexibility she provided me to complete this project. Additionally, I extend my sincere thanks to the organization that allowed me to utilize the software required for this research. I am truly grateful for their support, which has enabled my professional growth and academic journey.

On a more personal note, I would like to recognize the individuals closest to me who provided unwavering personal support throughout this journey. To my husband, Bobby, my biggest supporter and my biggest fan: Thank you for always believing in me, especially when I did not believe in myself. Thank you for never telling me the odds. I love you more than words can express. We did it! Here is to many more years of adventures together. To my mom: Thank you for always wanting more for me, for making sure I understood my potential, and for pushing me to be the best version of myself. I hope I have made you proud. To my Dad, Daddy, and my brother Jeff: I wish you could have been here to see this. I hope you are proud, smiling, and

toasting together. To Maria: Thank you for being the support, encouragement, cheerleader, and friend I needed throughout this journey. Your presence meant more than I can tell you. Love you, mean it! To my sisters, my chosen family: Thank you for decades of sisterhood, encouragement, and for rooting me on throughout this long journey. Your support has meant everything. To my sweet Bellatrix: Thank you for being my constant companion, writing buddy, hiking partner, comforter, reminder to take a break, and the best Boxer I could ask for. I promise we will have more fun time now. Finally, to NWW: Thank you for the Demigod, the salsa boards, the encouragement and support, and for giving me a comfortable, quiet corner to research and write. You will always be my spot.

Table of Contents

List of Tables

List of Figures

Section 1: Foundation

Healthcare revenue cycle management (RCM) has increasingly become a fertile ground for artificial intelligence (AI) as a transformational technology. Natural Language Processing (NLP), a subset of AI focusing on the interaction between human language and computers, has gained increasing attention for processing unstructured data (such as clinical notes), improving coding accuracy, and mitigating documentation deficits. These features are critical tools to address long-standing challenges such as claim denials and coding mistakes. NLP allows for extracting insights from datasets far bigger than would otherwise be possible through traditional processes. NLP, for instance, has shown a 30% increase in documentation accuracy, reduced coding errors, and improved billing accuracy (Futterman et al., 2023).

Although AI and NLP can be very promising, their adoption in healthcare RCM is challenging. This includes ethical challenges, data privacy concerns, and the necessity of strong governance strategies. Workers' reluctance to work on this shift and the complexity of incorporating AI into existing systems make the landscape even muddier (Nazi & Peng, 2024; Gabel et al., 2024; Aldoseri et al., 2023; Soroush et al., 2024). To this end, the dissertation will investigate whether an artificial intelligence (AI)-enabled Diagnosis-Related Group (DRG) validation tool that utilizes NLP processing has a positive operational and financial impact, and what human organizational factors affect its adoption. This study will analyze quantitative outcomes, such as improvements in quality metrics and financial performance, and qualitative feedback, including the perceptions of RCM professionals. By examining these dimensions, the research aims to identify pathways to maximize the effectiveness of AI integration within healthcare RCM. This dual emphasis on technical performance and human factors provides a holistic understanding of the elements necessary for successful AI adoption and sustained

utilization. The study's findings will also offer practical recommendations for overcoming common barriers to AI integration, such as resistance to change and gaps in technological infrastructure.

The results of this research will significantly contribute to the existing body of literature on AI implementations in healthcare. By systematically evaluating the accuracy of AI solutions and identifying key barriers to their adoption, this study will provide actionable insights for healthcare administrators, policymakers, and technology developers. These findings are intended to improve the financial viability and operational responsiveness of healthcare organizations, paving the way for more sustainable and efficient revenue cycle management practices (Kilanko, 2023; Sahni & Carrus, 2023). The study's insights will facilitate better decision-making in the adoption of AI technologies and enhance the broader understanding of their transformative potential in healthcare.

Statement of Problem

Although the promise of AI-powered solutions to transform healthcare revenue cycle management (RCM) is immense, multiple pitfalls continue to hinder both their efficacy and adoption. Among the most critical challenges are poorly trained, inadequately governed, and insufficiently tailored AI-based medical coding processes (Peng et al., 2021; Liu et al., 2022; Nazi & Peng, 2024; Futterman et al., 2023). These inadequacies often result in unpaid claims, delayed payments, and an excessive administrative burden, collectively undermining healthcare organizations' financial performance and operational capabilities (Seery & Wieczorek, 2022; Eramo, 2024b). The failure to customize AI systems to address healthcare-specific business processes further exacerbates these issues, creating inefficiencies that ripple across supply chains.

Regulatory challenges also play a significant role in impeding AI adoption in healthcare RCM. Concerns about algorithmic bias and a lack of transparency in decision-making processes erode stakeholder trust, making it difficult for organizations to embrace these technologies (Nazi & Peng, 2024; Peng, et al., 2021). The absence of standardized or authoritative implementation frameworks further complicates the landscape, leaving organizations to navigate an uncoordinated and often fragmented regulatory environment (Nazi & Peng, 2024; Weber, 2024b). These challenges are particularly acute for smaller healthcare organizations, which often lack the financial and technological resources to adapt. As a result, many of these institutions persist with outdated and inefficient processes, which may contribute to disparities in healthcare RCM and limit their ability to compete with larger, more resource-rich organizations (Aldoseri et al., 2023; Soroush et al., 2024; Gabel et al., 2024b).

Overcoming these challenges is critical to unlocking the full potential of AI in healthcare RCM. Addressing these barriers requires a multi-faceted approach that includes developing robust training programs for AI systems, implementing comprehensive governance protocols, and tailoring AI solutions to meet the unique demands of healthcare workflows (Peng et al., 2021; Nazi & Peng, 2024; Soroush et al., 2024; Gabel et al., 2024b). Additionally, fostering stakeholder trust through increased transparency and addressing ethical concerns around algorithmic bias are essential steps toward broader acceptance and integration of AI technologies (Nazi & Peng, 2024; Peng et al., 2021; Liu et al., 2022). For smaller organizations, targeted support in the form of scalable solutions and accessible training resources can help bridge the gap and enable equitable access to advanced technologies (Aldoseri et al., 2023; Soroush et al., 2024; Gabel et al., 2024b).

This study seeks to contribute to this effort by generating data-driven recommendations to enhance AI integration, ensure compliance with regulatory standards, and expand access to emerging technologies. By addressing technical and organizational barriers, the research aims to provide actionable insights to facilitate the sustainable adoption of AI in healthcare RCM. These findings have the potential to not only improve financial performance and operational efficiency but also promote equity and innovation across the healthcare industry (Kilanko, 2023; Pounds, 2021).

Purpose Statement

This qualitative descriptive study examines the scale coverage and perceived impact of a natural language processing (NLP)-based Diagnosis-Related Group (DRG) validation audit solution in healthcare revenue cycle management (RCM). It examines how this NLP-enhanced tool can affect DRG assignment accuracy, decrease the denial of claims, and tackle financial waste that can degrade organizational performance. Through this emphasis, our goal is to build a deeper, context-dependent understanding of how the tool impacts financial and operational practices in the RCM.

The study captures the RCM professionals' perceptions and experiences and sheds light on how this AI-driven solution affects their daily operations, decision-making, and performance (Weber, 2024a; Montroy et al., 2023). Utilizing a structured qualitative approach with semi-structured interviews and thematic analysis, it investigates barriers to adoption (ethical issues, resistance from clinical staff, and issues of data integration) and key facilitators (greater efficiency and improved reliability of documentation).

This targeted qualitative study seeks to guide best practices for AI adoption in healthcare RCM, highlighting the importance of ethical governance, robust data quality standards, and

change strategies developed to align human and technology needs of staff. The results are intended to provide concrete recommendations for healthcare leaders, technology experts, and policy makers looking to promote responsible and effective AI health technology integration in an array of healthcare settings (Sahni & Carrus, 2023; Wagner, 2023; McCormack, 2024a).

Nature of the Project

This applied doctoral project is a qualitative descriptive study that seeks to understand the impact of artificial intelligence (AI) enabled natural language processing (NLP) tools on hospital-based revenue cycle DRG validation workflows. Under the framework of the Human-AI Integration Framework (HAF), this project seeks to describe the nature of the errors made by AI-based DRG validation systems, how such errors occur in practice, and the effect of organizational or workflow variables on these errors. This research project aims to explore practical, experience-based evidence for guiding human-centered recommendations to help optimize DRG validation and AI integration workflows within health information management (HIM) and revenue cycle.

The project design contains two consecutive phases. An initial focused audit was performed on four hundred random, de-identified inpatient cases to explore the nature of DRG errors created by the AI. Using proprietary data from the middle revenue cycle vendor significantly enhances the validity and reliability of the study's findings. We do not claim this audit to be a stand-alone quantitative dataset, but rather a data artifact that enriches real-world relationships and helps to craft interview questions. Semi-structured interviews will be conducted with ten to fifteen healthcare professionals (including coding specialists, CDI team, auditors, and RCM leadership) in the second phase. These interviews delve into reasons for AI errors, the collaboration of the human coders and the

AI tools, and what is needed for system optimization. The project's problem statements and interview guides overlap significantly with the HAF trust calibration, task sharing, feedback loops, and explainability.

This qualitative descriptive methodology is suitable for describing the actualities in operation, user experience, and workflow perspective, which are necessary to comprehend the AI integration in the DRG validation processes. The findings will be used to develop specific recommendations for optimizing the design, deployment, and oversight of AI-driven tools in hospitals' RCM. This highest-quality asset also keeps research rooted in real, actionable knowledge, connected to the nature of health or healthcare RCM environments.

Needs Assessment

The rising complexity of healthcare revenue cycle management (RCM) has necessitated better coding accuracy, faster claims processing, and financial optimization. Central to this process is the Diagnosis-Related Group (DRG) coding system, which serves to classify inpatient hospital services for reimbursement purposes. However, erroneous DRG assignments create substantial revenue losses, claims denials, and regulators' scrutiny. Research suggests that errors are made in seven to thirty percent of inpatient claims, leading to billions of dollars in variations in the data every year (Kilanko, 2023; Montroy et al., 2023; Weber, 2024).

AI solutions using Natural Language Processing (NLP) have been developed to resolve these inefficiencies and aid with DRG validation. AI-powered DRG validation automates coding audits, pinpoints documentation gaps, and improves coding accuracy, all of which could minimize claim denials and delays in reimbursement (McCormack, 2024a). Although greater automation and accuracy are appealing promises, major adoption barriers remain, including

stakeholder hesitation, ethical and compliance risks, and technical and financial constraints (Gabel et al., 2024b; Nazi & Peng, 2024). All this has led to doubts among RCM professionals and health administrators on whether the new technology will attempt to replace human expertise and whether it will become just another level of complexity in workflow integration (Russell et al., 2024). There are still some important barriers to be solved, including the potential for algorithmic bias, lack of transparency in AI decision-making, and difficulty in regulatory compliance (Nazi & Peng, 2024). Small hospitals and under-resourced healthcare facilities lack the financial and technical readiness to adopt an AI-based DRG validation solution at scale (Root & Samtani, 2025).

Multiple touchpoints in DRG coding and revenue cycle workflows contribute to errors that cascade impact on financial and operational performance. Seven to thirty percent of the inpatient claims submitted contain DRG-related errors each year (Kilanko, 2023; Montroy et al., 2023; Weber, 2024), and incorrect DRG assignments are among the most common causes of claim denials (Lo et al., 2025). These errors have a long-term effect due to lengthy claim adjudication cycles, rework efforts, and payer appeals that delay reimbursement for weeks or months (Pounds, 2021). The challenge spans hospitals, health systems, and third-party revenue cycle vendors covering all payers (Soroush et al., 2024). This poses a disproportionate challenge for smaller hospitals and rural facilities that either lack the specialized RCM teams or the AI-capable infrastructure to support their operations (McCormack, 2024a). DRG misclassification directly impacts hospital revenue, compliance, and audit exposure. Getting the coding wrong can lead to upcoding penalties, underpayment losses, or even fraud investigations, with the potential for significant financial ramifications (Chaturvedi et al., 2024).

Stakeholder perceptions about AI adoption in revenue cycle workflows are mixed. Although many RCM professionals foresee automation continuing the trend of human error reduction, skepticism remains about AI reliability, data integrity, and usability (Gabel et al., 2024). Furthermore, clinical documentation quality is an important dependency. No matter how sophisticated the AI tools are, they are only as good as the data employed (Kristiansen et al., 2022). With increasing regulatory pressures, including Centers for Medicare & Medicaid Services (CMS) audit scrutiny, Office of the Inspector General (OIG) audits, and commercial payer scrutiny, healthcare organizations need to find a balance between automation and compliance (McCormack, 2024b). This paper seeks to measure the effectiveness of these AI-based DRG validation tools in minimizing coding errors while also addressing organizational, financial, and regulatory factors that challenge the implementation of AI in revenue cycle management.

This study will analyze data from real-world healthcare settings to assess the impact of NLP-driven DRG validation tools. The research will rely on secondary data sources from a mid-revenue-cycle vendor specializing in AI-driven DRG validation solutions. This will incorporate a structured audit of four hundred inpatient encounters, which will help inform semi-structured interviews with healthcare professionals in the health information and RCM space. This project aims to develop actionable recommendations to improve workflows in AI/NLP-driven DRG validation and AI integration by identifying error patterns in stakeholder responses. The study is guided by the Human-AI Framework, which leans into the importance of trust, task-sharing, feedback, and transparency in AI integration. Findings from this project will be disseminated through professional healthcare forums, industry conferences, and organizational leadership channels to support evidence-based decision-making in revenue cycle operations. Dissemination

is important to ensure that insights related to AI-enabled DRG validation are translated into practice, helping healthcare organizations improve coding accuracy, strengthen governance, and optimize workflow integration.

Project Questions

The questions for this study are designed to investigate the impact of NLP/AI-driven Diagnosis-Related Group (DRG) validation tools on revenue cycle management (RCM) workflows, focusing on error patterns, performance impacts, and recommendations for optimization.

PQ1: What are the most common types of errors identified in NLP/AI-driven DRG validation for high-volume and high-error episodes of care?

PQ2: How do NLP/AI-driven DRG validation errors impact key performance indicators in revenue cycle management, such as claim denial rates and reimbursement timelines?

PQ3: What are the underlying factors contributing to the inaccuracies of NLP/AI-based DRG validation in revenue cycle workflows?

Introduction to the Conceptual Framework

The Human-AI Integration Framework (HAF) will be the conceptual lens that guides the development of methods, data collection strategy, and interpretation of findings. A brief overview of HAF is provided in this section. A more detailed discussion of HAF, including background, use in the scientific literature, and the application of HAF to this project, is presented in the literature review section below. HAF provides a structure for understanding the interactions, dependencies, and performance outcomes of collaborative processes between professionals, in this case, RCM professionals and AI systems.

The HAF framework emphasizes four integration components that can impact how professionals interact with technology: role allocation, trust calibration, error recovery mechanisms, and communication interfaces (Wang et al., 2019). The framework argues that successful integration depends not only on the technical competence of AI but also on the transparency, explainability, and adaptability of its outputs to human reasoning (Zhang, et al., 2020; Vasconcelos et al., 2023). Most notably, HAF distinguishes between automation and augmentation regarding the use of AI as one of many decision support tools to facilitate and enhance decision-making and not as a replacement for professional judgment (Sahni & Carrus, 2023). AI's ability to enhance decision-making relies on the importance of team cognition and co-adaptive learning loops between human users and AI algorithms (Ma et al., 2023). Assumptions in the framework include: 1) both humans and AI contribute unique strengths to the workflow; humans provide context sensitivity and ethical reasoning, while AI offers scalability and pattern recognition; and 2) alignment between AI system design and human workflow processes leads to greater accuracy, trust, and efficiency (Sankaran et al., 2022). HAF emerged from research in human factors and cognitive systems engineering and later evolved through studies in human-computer interaction and collaborative automation (Shneiderman, 2020). In the field of healthcare management, particularly healthcare finance, HAF has been used to evaluate the transparency of algorithmic decisions, the mitigation of automation bias, and the conditions under which users override or defer to AI outputs (Vössing et al., 2022; Kim et al., 2023).

The HAF framework is well aligned with this dissertation's problem, purpose, and questions and has been effectively used in several studies similar to this applied project. Zhang et al. (2020) examined using confidence scores and explanation types in AI-assisted decision-making. They found that transparent, adaptive feedback helped improve human trust in the

technology and improved decision quality. Vasconcelos et al. (2023) demonstrated that AI explanations enhanced the collaborative performance between technology and human professionals. The studies support the assumptions that explainability and adaptive transparency contribute to building trustworthy AI systems and facilitate greater trust among human professionals who use them (Thiebes et al., 2020)

Significance of the Project

Adopting AI-powered Diagnosis-Related Group (DRG) validation tools in the context of healthcare revenue cycle management (RCM) is a gamechanger for financial and operational efficiency. Amidst complex changes in medical coding, payer regulations, and reimbursement models, healthcare organizations need innovative technology-driven solutions to maximize accuracy, compliance, and revenue cycle performance. This study contributes to the field of healthcare administration by systematically assessing the effect of AI on DRG validation, quantifying financial and operational benefits, and identifying barriers to successful implementation (Kilanko, 2023; McCormack, 2024a).

With healthcare costs continuing to soar, organizations are pressured to minimize coding errors, avoid claim denials, and improve financial viability. Coding errors found in seven to thirty percent of inpatient claims result in substantial revenue and compliance risks (Montroy et al., 2023; Weber, 2024). AI-driven solutions can help address the above-mentioned challenges by improving reimbursement accuracy (Russell et al., 2024), reducing administrative burdens, and increasing regulatory compliance. These research results will inform evidence-based recommendations for maximizing the effectiveness of AI in revenue cycle workflows, contributing practical insights for healthcare leaders, policymakers, and technology developers.

This study is significant to leaders and practitioners in healthcare administration, health information management (HIM), and revenue cycle operations. Healthcare leaders and financial officers are pursuing data-led solutions to optimize reimbursement structures, aid financial forecasting, and limit compliance exposure (Nazi & Peng, 2024). However, these DRG validation tools based on AI not only show the potential to overcome such inefficiencies but also show how they can truly add value in this area depending broadly on their integration in the workflow, the readiness of the workforce, and the regulatory perception (Root & Samtani 2025).

The research will provide empirical evidence that builds on prior literature focusing on AI applications in RCM by systematically assessing the utility of AI in DRG validation. Most prior studies have concentrated on predictive analytics, AI-powered clinical decision-making, and revenue optimization (McCormack, 2024b). Little research has evaluated NLP-based DRG validation tools' actual financial and operational effects in real-world settings. This research bridges these gaps by evaluating AI's efficacy in decreasing coding errors, enhancing financial outcomes, and identifying critical barriers to adoption (Chaturvedi et al., 2024). The results will serve as a reference for future research on using AI in the revenue cycle systems of health care.

While the problem statement outlines the negative impact of inaccuracies and inefficiencies in revenue cycle management, this section focuses on the positive aspect of successful AI-based DRG validation solutions usage. The benefits of this study's findings include better coding accuracy and financial integrity, operational feasibility and workforce optimization, regulatory compliance and risk mitigation, and scalability and long-term sustainability. AI-driven DRG validation can reduce DRG misclassification, allowing claims to be coded accurately and per the guidelines set forth by payers for reimbursement (Russell et al., 2024). This can improve coding accuracy, which helps hospitals reduce claims denials, avoid

compliance risks, and optimize overall revenue integrity (Weber, 2024). AI can also take over repetitive manual tasks, freeing HIM professionals and RCM specialists to work on high-value tasks (Montroy et al., 2023). AI tools can be integrated into workflow processes to decrease administrative burden and enhance financial performance (Nazi & Peng, 2024). By integrating AI-driven validation tools, organizations can increase compliance with CMS, HIPAA, and payer-specific regulations, improving audit readiness while minimizing exposure to financial penalties (Root & Samtani, 2025).

Importantly, RCM solutions powered by AI can be deployed across multiple hospital systems, potentially yielding sustained savings and operational efficiencies over time (Chaturvedi et al., 2024). Advances in machine learning (ML) and natural language processing (NLP) driven DRG validation will continue to enhance the accuracy and automation capabilities in coding (McCormack, 2024b). This research will address these critical aspects and enhance insights for healthcare executives, policymakers, and RCM leaders tasked with optimizing financial sustainability and operational efficiency in AI-augmented revenue cycle management.

Definitions of Key Terms

Artificial Intelligence (AI):

A branch of computer science that enables machines to perform tasks typically requiring human intelligence, such as decision-making, pattern recognition, and language processing (Ferrara, 2023).

Bias in AI Algorithms:

Systematic errors in AI decision-making due to training data limitations, which can lead to disparities in claim approvals and reimbursement (Nazi & Peng, 2024).

Blockchain in Healthcare RCM:

Decentralized ledger technology enhances data security, interoperability, and fraud prevention in financial transactions (Chaturvedi et al., 2024).

Clinical Documentation Improvement (CDI):

A process aimed at improving clinical documentation's accuracy, completeness, and specificity to ensure appropriate coding and billing (Gabel et al., 2024).

Deep Learning (DL):

A type of ML that uses neural networks with multiple layers to analyze complex patterns, often applied in AI-driven medical coding and fraud detection (McCormack, 2024b).

Diagnosis-Related Group (DRG) Coding:

A system that categorizes hospital stays into groups based on diagnoses and procedures to determine reimbursement rates (Root & Samtani, 2025).

Diagnosis-Related Group (DRG) Validation (Auditing or Review):

The process of reviewing DRG assignments to ensure that coding, clinical documentation, and reimbursement align with payer policies and regulatory requirements. DRG validation audits help healthcare organizations prevent revenue loss, mitigate compliance risks, and identify potential overcoding or undercoding issues (Centers for Medicare and Medicaid Services, 2014).

Explainable AI (XAI):

AI models designed to provide human-interpretable justifications for their outputs, ensuring transparency and trust in automated decisions (Ferrara,

Large Language Models (LLMs):

Advanced NLP models, such as GPT-4, that can generate and process human language at scale, are used for tasks such as summarizing clinical documentation and automating billing queries (Chaturvedi et al., 2024).

Machine Learning (ML):

A subset of AI that enables computers to learn from data and make predictions or automate tasks without explicit programming (Nazi & Peng, 2024).

Natural Language Processing (NLP):

A subset of AI that enables computers to understand, interpret, and generate human language is often used in healthcare for extracting medical insights from unstructured data like clinical notes (Soroush et al., 2024).

Predictive Analytics in RCM:

AI and statistical techniques are used to forecast claim denials, revenue trends, and reimbursement issues (Russell et al., 2024).

Literature Review

This section presents a detailed discussion of the literature on the use of AI in healthcare revenue cycle management (RCM). The literature review is structured in four broad topic areas. The review begins with an overview of AI, Natural Language Processing (NLP), Machine Learning (ML), and other advanced technologies. Next, a brief history of the rise of these technologies in healthcare revenue cycle management is presented. Third, a focused analysis of the evidence on the types of errors observed with NLP/AI technology is provided, and the implications of these errors on RCM performance and healthcare organizations are discussed. The literature review section concludes with an examination of the gaps in the evidence on AI

and RCM performance, followed by a discussion of the Human AI Integration Framework and the application of the framework to this applied project.

To compile this literature review, a search was conducted electronically to locate all relevant literature written in English, primarily peer-reviewed scholarly journals, through various databases, including CINAHL, Embase, Emerald, Google Scholar, Medline, ProQuest, PsycNet, PubMed, and Science Direct. Other applicable information (i.e., professional organizations and administrative reports) was gathered using Google and Google Scholar search engines. Resources include journal articles, books, government websites, professional association websites, and dissertations. The key terms utilized in the literature search were "artificial intelligence in healthcare", "AI in healthcare revenue cycle", "NLP in DRG validation", "artificial intelligence in clinical documentation improvement", "AI, machine learning, and NLP in health information management", "natural language processing in medical coding", "machine learning in revenue cycle management", "deep learning for fraud detection", "large language models in clinical documentation", "AI trust and explainability in healthcare", and "human-AI interaction framework". Variations of the key terms were utilized to ensure complete search results. The search scope was from 2009 to 2024, integrating pertinent seminal works, and most of the literature used in the review was published between 2018 and 2024.

Introduction to Artificial Intelligence and Its Technologies

Artificial Intelligence (AI) refers broadly to the development of computer systems capable of performing tasks that typically require human intelligence, such as decision-making, language interpretation, learning from data, and pattern recognition (Soroush et al., 2024). Under the AI umbrella, there are several distinct but related technologies, each contributing in unique ways to the automation and enhancement of complex processes across industries. Four of

the most widely adopted subsets of AI include Natural Language Processing (NLP), Machine Learning (ML), Deep Learning (DL), and Large Language Models (LLMs). Appendix A presents an overview of the various technologies that, while sharing a common goal of improving task performance, accuracy, and efficiency, differ in design, data requirements, computational complexity, and application domains (Montroy et al., 2023; Soroush et al., 2024).

Natural Language Processing (NLP).

NLP is a branch of AI that enables machines to understand, interpret, and respond to human language (Chaturvedi et al., 2024). NLP tools analyze unstructured text, such as clinical notes, insurance forms, and customer messages, and convert it into structured data. In healthcare, NLP is widely used for clinical documentation improvement and automated medical coding by extracting relevant content and mapping it to standardized billing codes like ICD-10 or CPT (Chaturvedi et al., 2024; Gabel et al., 2024). Outside healthcare, NLP powers applications like spam detection in email systems, sentiment analysis in social media monitoring, and virtual customer support in retail and banking (Soroush et al., 2024). The ability of NLP to bridge human language and machine-readable formats makes it a vital technology for industries that rely heavily on documentation and communication.

In RCM, NLP allows AI systems to read unstructured clinical documentation, transforming physician notes into billable medical codes. In October 2023, the technology was used to automate coding workflows, improve accuracy, and maintain compliance with evolving payer guidelines (Gabel et al., 2024a). NLP can be utilized in automated clinical coding by allowing coding platforms to extract relevant lexicon elements and match them with ICD-10, CPT, or HCPCS codes, mitigating human coding errors (Chaturvedi et al., 2024). It can assist in Clinical Documentation Improvement (CDI) by helping providers capture complete, accurate,

and payer-compliant documentation to mitigate claim denials (McCormack, 2024a). NLP can also analyze patterns from denied claims, which can help identify gaps in documentation, enabling NLP-driven systems to recommend enhanced documentation prior to claim submission, assisting in denials identification and prevention (Root & Samtani, 2025).

Machine Learning (ML) and Deep Learning (DL).

Both ML and DL are subfields of AI that involve training algorithms to learn from data and make predictions or decisions without explicit programming (Russell, 2024). ML refers to a broad range of techniques that can identify patterns in historical data and improve performance over time. DL, a more complex form of ML, uses neural networks with multiple layers to model intricate relationships, often achieving higher accuracy in tasks like image and speech recognition (Soroush, et al., 2024). In finance, these technologies support credit scoring and risk assessment, while in logistics, they forecast demand and optimize supply chain operations. DL's power to model complexity makes it particularly useful for high-volume, nonlinear problems, whereas traditional ML approaches offer greater transparency and interpretability (Soroush et al., 2024).

In the RCM process, ML and DL are deployed for predictive analytics, such as forecasting claim denials, detecting anomalies indicative of fraud, and optimizing prior authorization workflows (Kristiansen et al., 2022; Russell, 2024). ML and DL models help to make predictions in a way that improves with each cycle through large historical billing datasets. These models assist in sanitizing revenue cycle workflows by refining coding suggestions, reinforcing fraud detection, and forecasting claims denials (Russell, 2024). This technology can be utilized in predictive claim denial analytics when historical claim data is leveraged by ML models to predict potential future denials, enabling revenue cycle teams to

address such issues proactively (Kristiansen et al., 2022). In addition, ML technology can also assist with fraud prevention and anomaly detection. AI-based fraud detection systems indicate suspicious billing behaviors, which enable medical institutions to evade non-compliant claims (Ferrara, 2023). AI is helpful in the prior authorization workflow, where ML facilitates the preauthorization process by confirming medical necessity against payer policies, thus minimizing administrative bottlenecks (Nazi & Peng, 2024).

Large Language Models (LLMs).

LLMs represent a recent evolution of NLP and DL technologies, enabling machines to perform advanced language tasks by training on massive datasets (Montroy, et al., 2023). Generative models, such as OpenAI's GPT-4, can compose text, summarizing documents, generating code, and even answering domain-specific queries with human-like fluency. LLMs are revolutionizing fields such as education (e.g., tutoring platforms), legal services (e.g., contract summarization), and marketing (e.g., content generation). Unlike traditional NLP systems, LLMs do not rely solely on rule-based extraction or classification but use contextual understanding and generative reasoning, often outperforming previous models on a variety of language tasks (Montroy et al., 2023).

In healthcare RCM, LLMs enhance productivity through auto-generation of physician queries, automated appeal letters for denied claims, and real-time support via AI-enabled chatbots (McCormack, 2024b; Russell et al., 2024). The advent of LLMs has further advanced the capabilities of AI in RCM by enhancing natural language understanding, enabling documentation summaries, and aiding in complex billing inquiries (Montroy et al., 2023). LLMs have become helpful in CDI by providing LLM-Assisted physician queries. These not only help with real-time documentation queries but also lead to better communication between clinical

teams and coders (McCormack, 2024b). They are helpful in appeals, as well, by providing appeal letter generation, wherein AI-based systems analyze previous successful appeals and payer policies to create tailored appeal letters for denied claims (Russell et al., 2024). AI-enabled chatbots can also help patients and RCM personnel answer billing questions, claim status, and provide financial counseling (Root & Samtani, 2025).

While each AI technology is distinct, many industries are combining technologies to support end-to-end solutions (Soroush, et al., 2024). For example, an RCM platform may use NLP to extract data from clinical notes, ML to predict claim denials, and LLMs to draft appeal letters (Russell, et al., 2024). Combining AI technologies can create a more seamless, intelligent workflow. Healthcare administrators and industry leaders need to understand the unique roles and integration potential in seeking to implement reliable, scalable, and ethical AI systems (Montroy, Patel, & Brooks, 2023). With the rapid adoption of these technologies in healthcare, administrators must address transparency, governance, and explainability issues to ensure they can be effectively implemented to fulfill their transformative promise (Shneiderman, 2020).

The Rise of AI in Healthcare Revenue Cycle Management (RCM)

AI in healthcare revenue cycle management (RCM) is revolutionizing RCM at healthcare organizations. With guidance from NLP, ML, and LLMs, AI-enabled solutions offer an unprecedented opportunity to improve accuracy, eliminate tedious administrative tasks, and streamline billing workflows. These technologies are being applied to various aspects of RCM, including medical coding automation, real-time claims adjudication, and other predictive analytics used for revenue estimation (Kristiansen et al., 2022).

Environmental Factors Contributing to AI Adoption in RCM.

The widespread application of AI in RCM has been driven by many factors such as regulatory changes, increased administrative costs, and growing demand for enhanced accuracy in medical coding and billing (Russell et al., 2024). Organizations often face challenges related to healthcare regulations (e.g., HIPAA, CMS) that require strict billing compliance, necessitating AI-driven solutions that ensure regulatory compliance (McCormack, 2024a). AI can assist in ensuring that organizations remain in compliance. By automating tedious tasks such as medical coding, claims submission, and denial management, AI minimizes costs related to manual processes and reduces administrative burdens (Montroy et al., 2023). The use of AI improves payer-provider engagements by mitigating claim errors, improving documentation accuracy, and ensuring reimbursement of healthcare services provided (Nazi & Peng, 2024). Developments in deep learning and predictive analytics enable AI for claim adjudication, revenue projection, and fraud detection in healthcare organizations (Ferrara, 2023).

Improvements in RCM Efficiency.

Efficiency has been a long-standing issue in the healthcare industry, particularly in medical billing and claims processing. Manual coding and billing processes can be susceptible to human error, leading to claim denials, revenue losses, and higher administrative overhead costs (McCormack, 2024a). Due to these inefficiencies in RCM, AI applications wish to improve the processes through process automation, such as automating repetitive tasks, which are used in documentation, coding, claim processing, and accounts receivable management, and maintaining compliance with payer regulations that rapidly evolve depending on the policy formation (Russell et al., 2024). However, as expanded AI adoption continues, data integrity, algorithmic bias, transparency, and interoperability with existing health information systems remain growing challenges (Root & Samtani, 2025).

One of the most significant aspects of AI's application in the RCM workflow is its capacity to review massive volumes of both structured and unstructured data. NLP-driven AI tools analyze clinical documentation and notes to assign the most relevant billing codes that comply with payer regulations and reimbursement policies (Montroy et al., 2023). On the other hand, ML models analyze historical claims data to identify trends correlated with claim denials and recommend preemptive documentation enhancements to improve reimbursement rates (Soroush et al., 2024). Despite these improvements, AI-based solutions are not infallible and require human supervision to confirm coding correctness and adherence to healthcare financial policies (Chaturvedi et al., 2024).

The most notable improvement has been in the revenue cycle's efficiency, with AI-driven solutions automating repetitive manual work and augmenting financial fidelity. Research suggests that coding enhanced with AI can reduce claim denials by 25%, increase billing accuracy by 30%, and reduce processing time by 40% (Chaturvedi et al., 2024). In many areas, AI is assisting the revenue cycle with more efficient workflows. AI aids RCM professionals in making data-guided financial choices and improving reimbursement strategies (McCormack, 2024a). Given the dynamic nature of payer policies and coding guidelines, AI tools need to grow continuously (Nazi & Peng, 2024). AI should act as a tool that augments but does not replace human expertise, ensuring that human coders and financial professionals retain oversight (Russell et al., 2024).

Fraud Detection and Compliance.

Another important advancement of AI-powered RCM is its contribution to fraud detection and compliance. AI algorithms analyze big data to identify billing anomalies, flag suspicious claims, and improve audit practices (Nazi & Peng, 2024). These capabilities further

enable healthcare organizations to manage financial risk while ensuring compliance with federal and payer-specific regulations (Ferrara, 2023). Nevertheless, there are ongoing concerns about the ethics of AI-based fraud detection, particularly about false positives leading to unwarranted claim denials and resulting administrative issues for healthcare providers (Gabel et al., 2024b).

Financial Transactions.

As AI matures, its role in RCM will expand beyond automation concepts to support real-time financial decisions, become more integrated into patient financial interactions, and improve synergies between AI elements and human teams (McCormack, 2024b). Advancements in explainable AI (XAI), interoperability with electronic health records (EHRs), and ethical AI governance frameworks, including responsible AI, will refine the evolution of AI in RCM (Russell et al., 2024). AI has significant potential to revolutionize revenue cycle operations, but the overall success of AI in this domain will ultimately depend on overcoming implementation challenges, fine-tuning operational models between AI and human agents, and ensuring responsible governance of AI within healthcare financial endeavors (Root & Samtani, 2025).

AI extends beyond provider-facing financial operations, helping to revolutionize patient financial engagement by personalizing billing experiences and improving transparency (Montroy et al., 2023). AI-enhanced tools can provide patients with real-time breakdowns of medical expenses, allowing them to understand out-of-pocket costs and prepare for payments (Kristiansen et al., 2022). These tools can also assist with payment plans and patient assistance programs by analyzing patient financial history and recommending payment lines tailored to each patient, reducing medical debt and improving collection rates (Russell et al., 2024). AI chatbots can then respond to billing inquiries, insurance verification, and payment help, boosting patient satisfaction (McCormack, 2024a).

The increased usage of AI in healthcare revenue cycle management (RCM) is an emerging transformation in the conduct of financial transactions in medical organizations. Natural language processing (NLP), machine learning (ML), and deep learning (DL) are among the commonly used artificial intelligence functionalities to improve billing accuracy, optimize claim processes, and minimize administrative burdens (Kristiansen et al., 2022).

Interoperability with Important RCM Systems.

AI improves RCM efficiency by interoperating with electronic health records (EHR), billing and financial systems, and claim submission systems for healthcare payers and insurance. AI facilitates data exchange, minimizes administrative burden, and improves documentation accuracy (Chaturvedi et al., 2024). AI allows medical records to be automatically extracted and interpreted, allowing for accurate alignment between documentation and billing and minimizing the chances of manual entry errors (Root & Samtani, 2025). It also provides better data interfacing with payors. Through AI, insurance verification, eligibility checking, and prior authorization workflows become automated, speeding up the claim's approval process (Nazi & Peng, 2024). AI-supported blockchain technologies improve user data security on financial transactions, ensuring HIPAA compliance (Ferrara, 2023).

Barriers to AI Integration in RCM.

Though AI offers many benefits for RCM, its implementation does come with challenges. Evidence supports the fact that changing the healthcare billing and insurance policy requires high diligence and robust execution, implementation, verification, or surveillance to avoid any adverse effects (Kristiansen et al., 2022). The implementation of AI-powered RCM tools requires quality and structured data. The lack of standardized documentation practices and

multiple data sources can influence AI performance and yield claim inaccuracies (Gabel et al., 2024b).

There are also ethical concerns and biases in implementing AI in healthcare. Exposure to biases from training datasets may affect AI algorithms and contribute to disparities in claim approvals and reimbursement decisions (Ferrara, 2023). There can be issues integrating AI systems with the organization's legacy systems. Many healthcare providers operate on archaic financial management systems with little interoperability for AI-driven automation; therefore, integrating AI with existing systems remains challenging (Russell et al., 2024).

There are regulatory and compliance risks that must be considered. AI models require continuous updates to evolve payer policies and healthcare regulations. Not meeting these requirements may lead to refusals of claims and penalties in finances (Nazi & Peng, 2024). Some revenue cycle professionals remain skeptical of AI's decision-making processes, creating a need for robust explainable AI (XAI) frameworks to increase trust and adoption (Montroy et al., 2023).

Types of Errors in AI-Driven Medical Coding and Billing

Although AI has improved medical coding and billing efficiency, different types of errors remain. Such mistakes may arise from algorithmic misreading of data, inconsistencies, or the current limits of AI in understanding complex medical records. Knowledge of these errors is important for optimizing AI-based RCM systems and mitigating potential risks (Soroush et al., 2024).

Misclassified Medical Codes.

AI coding tools use NLP and ML models to analyze clinical notes to identify medical terms and assign the most relevant codes. However, misclassification has continued to be a

problem. AI systems may assign an incorrect code due to ambiguities in the physician documentation or incomplete contextual analysis (Gabel et al., 2024b). Improper secondary diagnoses coding can be an issue in AI coding. AI sometimes fails to encode associated comorbidities or complications that affect clinical decision-making and, therefore, reimbursement, leading to undercoding of claims (Russell et al., 2024). There may be discrepancies in code selection. AI trained on general datasets may have difficulty appropriately classifying procedures in specialty niches, including oncology, cardiology, and neurology (Montroy et al., 2023).

Two main challenges are overcoding and undercoding in AI-driven medical coding, which could lead to revenue cycle inefficiencies and compliance risks. AI could lead to the assignment of more advanced procedure codes than are warranted, thereby raising the risk for payer audits and potential repercussions from legal action (Nazi & Peng, 2024). On the other hand, AI may underestimate the severity of a patient's condition, resulting in lower reimbursement and financial penalties for providers (McCormack, 2024b). Medical necessity misalignment is another issue that has been found to result from errors in AI-generated coding. AI-generated coding recommendations could contradict payer medical necessity criteria, increasing denial rates and leading to extensive appeals (Chaturvedi et al., 2024)

Errors In Data Extraction and Saving of Incorrect Documentation.

NLP-powered AI tools access medical data from both structured and unstructured documentation, but discrepancies in physician notes or disjointed EHRs can introduce errors. There may be missing or omitted clinical information. Physicians sometimes use non-standard abbreviations, writing styles, or vague terminology, so AI models may lack the needed information (McCormack, 2024a). Misaligned data entries can cause complications, as well. AI

can misinterpret physicians' intention, resulting in incorrect coding recommendations that do not reflect actual diagnoses or procedures performed (Kristiansen et al., 2022). AI models also need to be regularly updated to reflect changes in medical coding standards, payer requirements, and new reimbursement policies (Nazi & Peng, 2024).

Bias in AI Algorithm Training.

AI coding tools are trained on historical billing data, and if the training datasets are biased, AI models may learn to perpetuate that bias in their processing of claims and billing decisions. The training models must contain representative samples, or AI may be biased towards over-representing common codes or under-representing conditions that are under-represented in the training data (Ferrara, 2023). AI may often miss or inaccurately flag claims from certain demographic groups, resulting in inequalities in reimbursement (Kristiansen et al., 2022). AI trained on high-volume procedures often has difficulty accurately coding low-volume, rare, or complex cases, leading to billing errors (Root & Samtani, 2025).

Mistakes In AI-Driven Claims Adjudication.

AI-powered claims adjudication tools analyze and process claims, determining eligibility, coverage, and reimbursement. However, mistakes made by the AI can cause improper judgments and the rejection of claims. One issue found is with AI fraud detection models. These models can mistakenly classify valid claims as fraudulent, leading to unnecessary audits and delayed reimbursements (Ferrara, 2023). Contextual comprehension is a limitation found in AI-driven tools. Advanced AI may struggle with complex clinical cases involving multiple coding inferences, leading to inappropriate claim categorizations (Soroush et al., 2024). The non-adaptive mechanisms in AI-driven claims adjudication tools are also a concern. AI-driven

adjudication models need to be continuously updated to keep up with changing regulations and payer reimbursement criteria (Russell et al., 2024).

Limits of AI in Fraud Detection and Compliance.

AI-driven fraud detection tools scan billing patterns for suspicious behavior, but these technologies are not perfect and can result in false claim denials or compliance issues. High rates of false positives have been identified as a top concern in these models. Models for AI fraud detection are prone to high false-positive rates, leading to increased administrative burden for healthcare organizations (Kristiansen et al., 2022). The fraud detection mechanisms must also be kept up to date. AI models should be updated over time to identify new dos and schemes that replace outdated tactics, but hesitance to implement updates can lead to periods of program failure (Root & Samtani, 2025). Organizations must ensure that their models remain compliant with regulations. AI fraud detection algorithms should take care to create a compliant model that meets CMS, HIPAA, and other healthcare guidelines (McCormack, 2024b).

Summary of the Evidence on AI in Revenue Cycle Management (RCM)

AI applications are being rapidly deployed in RCM due to the technology's demonstrated ability to enhance process efficiency and performance through automating medical coding, claim adjudication, and fraud detection. Standardization, ethical considerations, model transparency, and interoperability challenges remain. Addressing these challenges is important for improving AI's reliability, fairness, and effectiveness in RCM (Root & Samtani, 2025; McCormack, 2024a), which should be addressed in future research.

Strengths of the Evidence.

The integration of AI technologies into RCM processes has yielded numerous improvements in coding accuracy, claims processing efficiency, and regulatory compliance. Combining different technologies (NLP, ML, DL, and LLM) has been shown to be an effective strategy for automating data extraction from unstructured documentation and predicting claim outcomes. NLP systems have demonstrated value in mapping clinical narratives to accurate ICD-10 and CPT codes, decreasing claim denials due to incomplete or incorrect documentation (Chaturvedi et al., 2024; Gabel et al., 2024a). Furthermore, NLP-supported clinical documentation improvement tools guide providers in capturing more comprehensive, compliant medical records, thereby reducing downstream billing issues (McCormack, 2024a).

ML and DL models contribute further by analyzing historical billing and reimbursement data to identify patterns that signal claim denials, documentation gaps, or reimbursement risks. These models can dynamically learn from evolving datasets and provide predictive insights that support proactive decision-making and denial prevention (Russell et al., 2024; Kristiansen et al., 2022). AI-enabled financial forecasting tools also empower healthcare administrators to anticipate cash flow changes and manage the revenue cycle more strategically (Montroy et al., 2023).

Another strength lies in AI's application to fraud detection and compliance monitoring. ML algorithms can identify unusual billing patterns indicative of fraud, such as upcoding or unbundling, while compliance tools automatically audit claims against payer policies (Ferrara, 2023). These tools are designed to alert organizations of compliance risks in real-time, enabling faster response and mitigating legal or financial penalties (Nazi & Peng, 2024). Collectively, these AI-driven capabilities reduce manual workloads, streamline administrative tasks, and position healthcare systems for improved financial performance and accountability.

These strengths suggest that AI technologies, when properly trained and aligned with healthcare regulations, can significantly enhance revenue integrity and operational efficiency. When deployed in combination (e.g., using NLP for data extraction, ML for predictive analytics, and LLMs for automated communication), AI platforms can support robust end-to-end RCM strategies that are both scalable and adaptive (Soroush et al., 2024; Root & Samtani, 2025).

Areas of Divergence and Uncertainty in the Evidence.

Despite promising outcomes, implementing AI in RCM has revealed inconsistencies and uncertainties, particularly in the accuracy and reliability of AI-generated outputs. A primary concern is the variability in clinical documentation styles and terminology, which can compromise NLP accuracy. For example, non-standard abbreviations or incomplete physician notes may cause AI systems to misinterpret clinical intent, leading to incorrect coding (Gabel et al., 2024a). This misinterpretation can result in overcoding or undercoding, affecting reimbursement accuracy and triggering compliance audits (Kristiansen et al., 2022).

Another area of uncertainty relates to the adaptability and maintainability of AI systems in the face of regulatory and policy changes. AI tools require frequent retraining and updates to remain aligned with evolving CMS guidelines, payer requirements, and coding standards. Delays in implementing such updates can render AI systems outdated and inaccurate, increasing the risk of claim rejections and financial loss (Root & Samtani, 2025). The lack of automated integration with real-time regulatory databases further complicates efforts to maintain compliance across different healthcare systems.

Bias in AI algorithms also presents a significant challenge. AI models trained on historical billing data may replicate existing inequities or inaccuracies, potentially skewing coding recommendations or claim predictions. This bias could disadvantage certain patient

populations or result in inconsistent coding outputs across similar clinical scenarios (Ferrara, 2023). In addition, transparency in AI decision-making remains limited. Many models function as "black boxes," offering limited visibility into how coding or billing decisions are derived, which undermines trust and makes auditing difficult (McCormack, 2024b).

Operational and integration challenges, such as user acceptance, system interoperability, and AI literacy among coders and RCM professionals, continue to hinder optimal integration. While some organizations report high productivity gains, others struggle with training burdens and poor system fit within legacy EHRs and RCM tools (Montroy et al., 2023). These divergences suggest that AI implementation outcomes vary widely depending on institutional readiness, technical infrastructure, and workforce capability.

Gaps in the Current Evidence.

While the growing literature on AI in RCM highlights notable innovations, several gaps persist in empirical understanding and system-level validation. One significant gap is the limited number of peer-reviewed longitudinal studies measuring the sustained impact of AI systems on financial performance, compliance outcomes, and workforce efficiency. Much of the available research focuses on pilot programs, vendor case studies, or short-term impacts, leaving a need for rigorous, long-term evaluations across diverse healthcare settings (Soroush et al., 2024).

The noticeable lack of comparative effectiveness studies for these technologies is another important gap in the current evidence. There is a lack of systematic comparisons between human coders, AI-only tools, and hybrid human-AI models in real-world environments. These comparative studies would be valuable in quantifying the added value of AI over traditional methods, identifying when and where human oversight is most critical, and guiding best practices for task allocation (Russell et al., 2024; Root & Samtani, 2025). Without such

benchmarks, organizations face uncertainty in selecting the most efficient and compliant RCM strategies.

There is a gap in the evidence on understanding the implications of AI-driven billing processes on the patient experience. In fact, few studies explore the patient-centered implications of AI-driven billing and RCM processes. Research is needed to assess how AI-based errors impact patient satisfaction, billing disputes, or care-seeking behavior, particularly among vulnerable or underserved populations. As AI tools influence cost estimates, benefit verifications, and payment options, understanding the downstream effects on patient trust and equity in financial communication becomes increasingly important (Kristiansen et al., 2022; Chaturvedi et al., 2024).

Finally, there is a lack of evaluation research on AI implementation in RCM that contributes to the need for standardized metrics and reporting guidelines for evaluating AI performance in healthcare billing and coding. Metrics such as AI coding accuracy rates, audit pass rates, error types, denial prevention efficacy, and compliance audit outcomes should be consistently defined and published across studies to promote shared learning and continuous improvement (McCormack, 2024a; Nazi & Peng, 2024). The absence of such standards limits generalizability and the ability to replicate successful implementations across settings, further reinforcing the need for cross-disciplinary research collaborations and regulatory guidance.

The Human-AI Integration Framework (HAF) in Revenue Cycle Management

This section presents a detailed description of the HAF framework used for this project. The section begins with a brief overview of the framework, a discussion on how the framework evolved, and an examination of the framework's major components. The section concludes with

an overview of how the framework has been used in previous studies and an application of HAF to this project.

HAF is a conceptual model developed to understand and improve how humans and AI systems can work collaboratively within complex decision-making environments. The framework focuses on how task allocation, interaction quality, and mutual trust between humans and machines influence overall system performance. In healthcare RCM, HAF provides a framework for examining how human coders and AI tools co-perform in error-prone, high-volume tasks like medical coding and DRG validation. HAF provides insights into the conditions necessary to promote effective collaboration to include 1) explainability, 2) trust calibration, 3) feedback loops, and 4) data strategy alignment (Kim et al., 2023; Vasconcelos et al., 2023).

HAF has roots in multiple disciplines, including human factors engineering, human-computer interaction, and organizational systems design. The framework evolved in response to the growing deployment of AI technologies in domains that require human oversight, such as healthcare, aviation, finance, and defense. Early iterations emphasized trust, transparency, and system usability (Shneiderman, 2020). Later versions of the framework incorporated more dynamic models of teaming, acknowledging that effective AI integration depends on contextual sensitivity and human adaptability (Wang et al., 2019).

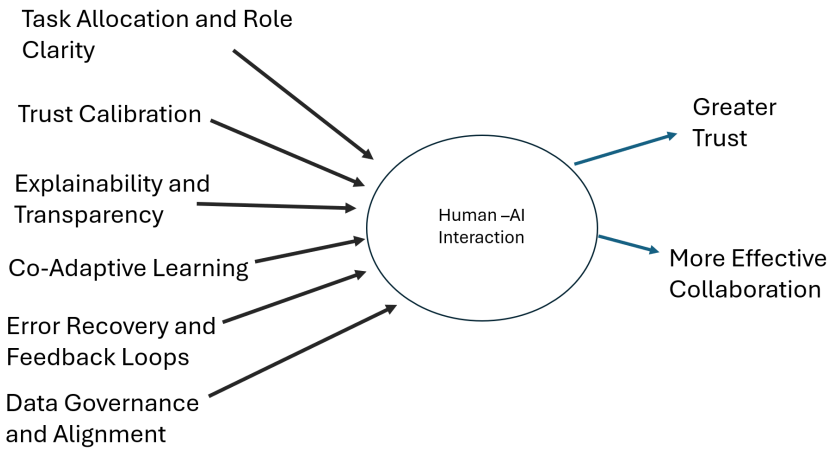
HAF gained prominence in response to concerns about automation bias and black-box algorithms in healthcare, especially in applications where patient safety and regulatory compliance are paramount (Laux et al., 2023). The need to balance algorithmic precision with human intuition and ethical reasoning led to empirical and conceptual research using HAF as a vital tool for AI system design and evaluation in high-stakes environments (Ma et al., 2023).

Major Components of the Human-AI Integration Framework.

The Human–AI Integration Framework (HAF) comprises several interdependent components that facilitate reliable, transparent, and collaborative decision-making between AI systems and human professionals. As shown in Figure 1, these components function as facilitators of trust and performance in settings such as healthcare RCM, where AI is increasingly used in DRG validation. The task allocation and role clarity component stresses the importance of clearly and appropriately dividing responsibilities between AI tools and human coders and auditors. Tasks best suited to AI, such as identifying high-volume, structured data patterns, can be automated. On the other hand, human coders and auditors could retain oversight for complex or ambiguous cases. An appropriate, clearly defined division of roles can enhance workflow efficiency, prevent duplication of work, and reduce error rates (Sankaran et al., 2022).

Figure 1

Graphical Representation of HAF Components



Trust calibration is another essential component and refers to the degree of alignment between an appropriate level of user trust in the AI system compared to actual system performance. If trust is too low, users may ignore accurate AI suggestions. Conversely, if trust is

too high, users may over-rely on flawed recommendations. Both ends of the trust-performance continuum can pose risks to patient safety, regular compliance, and process efficiency in a complex environment like healthcare (Zhang et al., 2020). Proper calibration is aided by transparent and explainable AI outputs that allow users to understand how decisions are made and to interrogate the logic behind recommendations. Explainability can be enhanced through clear rationales, confidence scores, and interactive interfaces that enable users to take corrective action when needed (Kim et al., 2023). Explainability features are particularly critical in RCM workflows where understanding how a DRG was assigned is key to compliance and reimbursement accuracy.

Co-adaptive learning in the HAF framework is a dynamic process in which both human professionals and AI systems evolve in response to each other's behaviors. For example, coders/auditors might adjust their workflows based on patterns they observe in AI performance. Subsequently, the AI system is periodically retrained or fine-tuned based on coder and auditor feedback and error corrections. Mutual adaptation strengthens system resilience, supports continual improvement, and promotes sustainable integration over time (Ma et al., 2023). In healthcare RCM environments, such learning cycles can lead to more responsive systems that adjust to institutional practices or emerging coding standards.

The error recovery and data governance component completes the HAF structure. AI-integrated systems must provide structured feedback loops that enable timely identification and correction of errors related to data inputs, decision logic, or interpretive outputs (Vössing et al., 2022). Feedback mechanisms, such as coder flags or audit triggers, can inform both human users and AI developers. Finally, data governance and alignment are critical for building trust and

fostering enhanced collaboration. It is important that AI systems are trained and operated on accurate, well-structured, and ethically sourced data (Aldoseri, et al., 2023).

The Human-AI Framework in Recent Literature.

Recent empirical studies support the utility of the HAF framework in understanding AI performance and workflow processes. Ma et al. (2023) demonstrated that users who were given their historical correctness scores and AI confidence levels were better able to judge when to trust AI outputs. Improved trust calibration improved decision quality and reduced overreliance. Vasconcelos et al. (2023) showed that providing users with context-sensitive explanations, which contributed to improved explainability and transparency of AI processes, enhanced user ability to identify and override incorrect AI recommendations. In addition, Kim et al. (2023) found that different types of explanations (e.g., saliency maps, feature-based rationales, and confidence indicators, etc.) must be tailored to user roles and domain expertise to be most effective

Recent theoretical articles reinforce the importance of the purposeful and strategic development of transparency features to enhance trust and system performance. Excessive information can overwhelm users, while insufficient detail reduces trust and usability (Zhang et al., 2020; Vössing et al., 2022). Sankaran et al. (2022) proposed a modeling framework to quantitatively assess the collaborative performance of human-AI teams in healthcare and finance, including task efficiency, error propagation, and decision-making quality. Sheinderman (2020), using ethical and system-design perspectives, advocated for human-centered AI systems that embed usability and auditability into the design process. These empirical and theoretical studies support the insight that trust, explainability, and data alignment are foundational to successful human-AI teaming.

Application of the Framework for this Dissertation Project.

This dissertation employs the Human-AI Integration Framework to examine how medical coders and auditors interact with a Natural Language Processing (NLP)/AI-powered DRG validation tool within revenue cycle workflows. The study addresses high-volume, high-error episodes of care. It uses organizational audit data to identify common error patterns, understand their impact on performance metrics, and analyze contributing factors using the HAF framework as presented in Table 1.

Table 1

Alignment of the Human-AI Interaction Framework with this Project

HAF Component	Alignment with this Project
Task Allocation	Identifying where in the DRG workflow AI performs reliability versus where human review is necessary.
Trust Calibration	Investigating whether coders defer too much or too little to AI recommendations. Understanding the impact of current trust-performance alignment on coding accuracy and reimbursement timelines.
Explainability	Assessing whether coders have sufficient understanding of AI-generated outputs to appropriately intervene if necessary.
Feedback Loops	Exploring whether error data is systematically used to improve both human processes and AI performance.
Data Strategy	Understanding whether issues of poor data quality (e.g., fragmented data, incomplete data, insufficient coding guidelines, etc.) contribute to AI errors.

Ultimately, the framework provides a structured, evidence-based lens for interpreting the findings and making practical recommendations to improve the integration of AI in RCM environments. It ensures that the project is not only technically grounded but also sensitive to the organizational, ethical, and human dynamics that shape healthcare delivery.

Summary

This applied doctoral project analyzes the operational and financial impact of implementing an artificial intelligence (AI)-driven Diagnosis-Related Group (DRG) validation audit tool into healthcare revenue cycle management (RCM). In Section 1, the project's foundation was established through the problem statement, purpose, research questions, theoretical framework, needs assessment, and the study's significance. By focusing on issues related to DRG assignment errors, claims denials and fees, and compliance risks, the research fills a key gap in the literature on how best to adopt and optimize AI-powered tools in healthcare financial operations. This is the set of challenges that hinder organizations.

This investigation has a compelling rationale, given that the literature reports that DRG-related coding errors in inpatient claims have been estimated to account for 7% to 30% of total inpatient claims (Kilanko, 2023; Montroy et al., 2023). These errors lead to revenue leakage, regulatory risk, and administrative costs. The literature acknowledges the potential of AI and Natural Language Processing (NLP) technologies to improve coding accuracy and process efficiency; however, their adoption is not ubiquitous.

While the body of existing research will no doubt expand, divergences suggest formidable operational barriers, including algorithmic bias, data integrity, workforce resistance, and limited interoperability with legacy systems (Nazi & Peng, 2024; McCormack, 2024a; Root

& Samtani, 2025). To contextualize the challenges and opportunities of AI integration, the study is grounded in the Human-AI Teaming Framework that explores workflow coordination, trust in automation, and organizational responses to coding errors. By integrating structured analysis of audit artifacts with qualitative stakeholder interviews, the study provides a comprehensive perspective on AI adoption in revenue cycle operations.

In summary, section one presented a rationale and significance for the project by synthesizing industry challenges, literature gaps, and organizational needs. This sets the stage for section two, describing the methodological approach, including participant selection, data sources, instruments, and analytic strategies. This comprehensive design aims to provide actionable insights for healthcare administrators, policymakers, and technology developers to achieve and sustain AI-driven solutions in healthcare RCM and ensure their maximum effectiveness.

Section 2: Method and Design

This section presents the research design and methodological approach used to explore the dissertation research problem. The problem and purpose statements are restated here for alignment with the overall study design. Although the promise of AI-powered solutions to transform healthcare revenue cycle management (RCM) is immense, multiple pitfalls continue to hinder both their efficacy and adoption. Among the most critical challenges are poorly trained, inadequately governed, and insufficiently tailored AI-based medical coding processes (Peng et al., 2021; Liu et al., 2022; Nazi & Peng, 2024; Futterman et al., 2023). These inadequacies often result in unpaid claims, delayed payments, and an excessive administrative burden, collectively undermining healthcare organizations' financial performance and operational capabilities (Seery & Wieczorek, 2022; Eramo, 2024b). The failure to customize AI systems to address healthcare-specific business processes further exacerbates these issues, creating inefficiencies that ripple across supply chains.

Regulatory challenges also play a significant role in impeding AI adoption in healthcare RCM. Concerns about algorithmic bias and a lack of transparency in decision-making processes erode stakeholder trust, making it difficult for organizations to embrace these technologies (Nazi & Peng, 2024; Peng, et al., 2021). The absence of standardized or authoritative implementation frameworks further complicates the landscape, leaving organizations to navigate an uncoordinated and often fragmented regulatory environment (Nazi & Peng, 2024; Weber, 2024b). These challenges are particularly acute for smaller healthcare organizations, which often lack the financial and technological resources to adapt. As a result, many of these institutions persist with outdated and inefficient processes, which may contribute to disparities in healthcare

RCM and limit their ability to compete with larger, more resource-rich organizations (Aldoseri et al., 2023; Soroush et al., 2024; Gabel et al., 2024b).

This qualitative descriptive study examines the scale coverage and perceived impact of a natural language processing (NLP)-based Diagnosis-Related Group (DRG) validation audit solution in healthcare revenue cycle management (RCM). It examines how this NLP-enhanced tool can affect DRG assignment accuracy, decrease the denial of claims, and tackle financial waste that can degrade organizational performance. Through this emphasis, our goal is to build a deeper, context-dependent understanding of how the tool impacts financial and operational practices in the RCM.

The research progressed in two stages. First, a retrospective validation audit of 400 de-identified inpatient cases was conducted to assess the accuracy of DRG assignments and the coding integrity of a proprietary AI-driven DRG validation software. The audit served as an important source of descriptive contextual data identifying the nature and frequency of DRG assignment errors. The second stage consisted of semi-structured interviews with key revenue cycle stakeholders to explore their experiences with and perceptions of AI-generated errors in DRG validation. Thematic analysis, guided by the Human-AI Interaction Framework, examined how roles, trust, explainability, and feedback processes influenced the identification and handling of such errors within organizational workflows. Three overall project questions guided the development of the research design and methods:

PQ1:What are the most common types of errors identified in NLP/AI-driven DRG validation for high-volume and high-error episodes of care?

PQ2:How do NLP/AI-driven DRG validation errors impact key performance indicators in revenue cycle management, such as claim denial rates and reimbursement timelines?

PQ3:What are the underlying factors contributing to the inaccuracies of NLP/AI-based DRG validation in revenue cycle workflows?

This chapter describes the rationale for choosing a qualitative research design, the specific research tradition guiding the inquiry, and the strategies for ensuring trustworthiness, ethical rigor, and alignment with the study's goals. The methodology reflected the exploratory nature of the project questions and the need to capture in-depth perspectives from stakeholders working with AI-driven DRG validation systems in real-world healthcare environments. The remainder of this section includes research design, population and sampling procedures, instrumentation and data collection, data analysis plan, issues of trustworthiness, ethical considerations, and a summary of the methodology.

Methodology and Design

This study used a qualitative descriptive design in which a structured artifact analysis (i.e., the DRG validation audit) was used to surface patterns or anomalies that were then explored more deeply through qualitative inquiry with key stakeholders (Guetterman et al., 2015). In the context of this project, the DRG audit served as a contextual artifact to explore how errors generated by NLP/AI-based DRG tools influenced hospital revenue cycle management. Qualitative description is an appropriate method when the goal is to obtain detailed accounts of events, experiences, or processes from the perspective of those directly involved, without requiring abstraction to high-level theory (Sandelowski, 2000). Given the applied nature of this dissertation and its emphasis on capturing operational insight and organizational response, this design enabled the collection of rich, practice-based data that informed implementation strategies, decision-making, and system improvement.

The rationale for selecting a qualitative approach stemmed from the exploratory nature of the project questions (Creswell & Poth, 2018). Quantitative approaches were insufficient to address the depth and contextual complexity required to answer these questions given the limited availability of standardized measures for NLP-related error types and their consequences within diverse healthcare settings (Patton, 2015). The study prioritized participants' interpretations and aimed to generate insights to improve AI performance in RCM processes rather than generalizing broad findings.

A qualitative descriptive design was aligned with the goals of healthcare implementation research (Bradshaw et al., 2017). This project sought to understand the lived experiences of revenue cycle, clinical documentation, and IT professionals to uncover system-level barriers and facilitators to process performance (Colorafi & Evans, 2016). This study sought to uncover how AI-generated coding errors were experienced in context and how organizational routines evolved in response to them, making a descriptive analysis of stakeholder perspectives essential.

Alternative Methodologies

Several research methodologies were considered for this project but were not selected. A quantitative survey design would have involved developing a structured questionnaire to collect numerical data from a large number of RCM professionals about the frequency, impact, or causes of DRG validation errors. This approach allowed for statistical analysis of trends and correlations and could offer generalizable insights across multiple settings (Creswell & Creswell, 2018). Although this is a strong methodology for identifying patterns or correlations, the quantitative survey design does not allow for a deep understanding of elements central to this project, including contextual factors that may contribute to coding errors, stakeholder perceptions, or the interaction between human and AI systems (Patton, 2015). In addition,

quantitative surveys require pre-structured responses that can limit the richness and nuance needed to explore how RCM professionals interpret and respond to AI-generated errors (Creswell & Poth, 2018). Given the exploratory nature of the study, a quantitative survey was determined to be ineffective for addressing project goals.

A case study approach would have allowed the researcher to conduct an in-depth investigation of a healthcare organization using AI-enabled DRG validation systems. Case study design is especially suitable for exploring complex phenomena within real-life contexts (Yin, 2018). Although the case study method could provide rich, contextualized insight, the aim of these methods was to explore a phenomenon holistically. In contrast, the aim of this project was focused more narrowly on stakeholder experiences with AI-related DRG validation errors rather than organizational implementation of AI technologies.

A grounded theory qualitative design was also considered for this project. Grounded theory is used to generate new theories or conceptual frameworks grounded in data collected from participants (Charmaz, 2014). The method involves systematic data coding and constant comparative methods to develop a theoretical model. The aim of this project was not to develop a new theoretical model but to examine a specific phenomenon through the lens of an existing conceptual framework to guide inquiry and data interpretation.

The alternative designs discussed above offered unique strengths and analytical capabilities but were not aligned with the specific aims and focus of this project. The study's emphasis on identifying common DRG validation errors, understanding stakeholder experiences, and assessing the impact of the errors on revenue cycle performance was best suited to a qualitative descriptive design. This exploratory design offered the flexibility, accessibility, and

practical relevance necessary to achieve project goals without overextending toward theory generation or broader generalization (Sandelowski, 2000).

Population and Sample

The unit of analysis for this study was the hospital inpatient revenue cycle management (RCM) function, with a focus on processes involving AI-enabled DRG validation tools. The population of interest included hospital-based RCM professionals with direct experience using natural language processing (NLP)-enabled tools in clinical coding, DRG validation, or claims auditing. Although the exact size of this population was not known, industry adoption of AI in revenue cycle functions has expanded significantly in recent years among medium to large healthcare systems (Gabel et al., 2024b). The population was appropriate for addressing the project goals of understanding the nature and context of DRG validation errors in real-world settings. RCM professionals with direct experience in using these tools were best positioned to assess the technological and human interaction factors that may contribute to or prevent these errors.

A purposive sample of 10 to 15 participants was recruited. Eligible participants included professionals in roles such as inpatient coding, clinical documentation improvement (CDI), medical auditing, and RCM program management. Participants were required to have at least three years of experience in hospital RCM and direct familiarity with NLP-enabled DRG validation tools. This sample size was consistent with recommendations for achieving saturation in qualitative descriptive and exploratory research designs with participants who have highly specialized expertise and shared experiential domains (Hennink et al., 2017; Sandelowski, 2000).

Recruitment occurred through professional associations (e.g., American Health Information Management Association [AHIMA], Healthcare Financial Management Association

[HFMA]), personal referrals within the researcher's professional networks, and outreach to clinical coders and auditors affiliated with health IT vendors or implementation partners. Participants were invited via email with a short project summary, eligibility criteria, and informed consent form. Additional information about the recruitment procedures was contained in the Project Procedures section below. The interviews were conducted virtually and recorded for transcription and analysis. This recruitment approach was designed to ensure both relevance and diversity in participant experiences while remaining feasible for the project timeline and scope.

Materials and Instrumentation

This project used two instruments for data collection: (1) a structured audit protocol for conducting DRG validation assessments and (2) a semi-structured interview guide developed using the Human–AI Interaction Framework. The two instruments enabled the triangulation of findings from observed system errors and stakeholder perspectives. The instruments were aligned with the qualitative descriptive research design for this project.

DRG Validation Audit Protocol

The DRG validation audit instrument for this project was presented in Appendix B. This section discussed the data collection instrument used in the DRG validation audit. A more detailed discussion of the data collection methods was presented in the Data Collection and Analysis section below. The instrument was a structured protocol developed by the investigator to systematically assess the accuracy and financial implications of NLP-enabled DRG coding decisions. The instrument included fields for capturing ICD-10 code assignments, derived DRG codes, audited DRG outcomes, and payment variance associated with errors. Additional fields allowed classification of the error type (e.g., diagnosis omission, miscode, sequencing error) and

root-cause assessment (e.g., ambiguous documentation, AI misinterpretation, or system training flaw).

The audit procedures discussed in Appendix B were grounded in coding compliance literature and aligned with auditing standards established by the American Health Information Management Association (AHIMA) and CMS DRG validation guidance. Adherence to these standards strengthened the content validity of the audit. In addition, the audit's case tracking spreadsheet, an example of which was presented in Table B-1 in Appendix B, was reviewed by a certified coding specialist and a clinical documentation improvement (CDI) professional to ensure alignment with industry standards. Interrater reliability was strengthened through the use of a single experienced reviewer for consistency, with internal checks for coding accuracy.

Semi-Structured Interview Guide

Appendix C contains the semi-structured interview guide developed by the investigator based on the Human–AI Interaction Framework (Zhang, Liao, & Bellamy, 2020; Kim et al., 2023). The guide contains open-ended questions designed to elicit stakeholder insights into the sources, consequences, and system-level contributors to DRG validation errors. In addition, the guide contains prompts targeting key dimensions of the framework to include role allocation, trust calibration, explainability, co-adaptation, and error recovery mechanisms (CITE).

Alignment with Project Questions

The interview questions were intentionally aligned with the project's three guiding research questions to ensure data collection would meaningfully address the project goals. Specifically, interview questions 2 and 3 sought to capture respondents' experiences with common DRG validation errors and to understand how these errors were detected during the

revenue cycle workflow. These questions were directly linked to PQ1. Interview questions 4 and 5 sought to explore how these errors influenced key revenue cycle performance metrics such as claim denials and reimbursement timelines, as well as organizational responses to these challenges. These questions were directly aligned with PQ2. Finally, interview questions 6 and 7 investigated respondent perceptions regarding the underlying technical, organizational, and contextual factors contributing to DRG validation inaccuracies. These questions were aligned with PQ3. Additional interview questions on trust, explainability, and system improvement sought to obtain further insight into broader Human–AI interaction dynamics that shaped error recognition, mitigation, and learning within revenue cycle teams. This alignment ensured conceptual integrity across the problem, purpose, and method aspects of the project.

Alignment with Human-AI Interaction Framework

The interview questions were designed to reflect core components of the Human–AI Interaction Framework (Zhang et al., 2020). Alignment of interview questions with the framework allowed for the data collection process to capture how human professionals engaged with, adapted to, and shaped the use of AI-enabled DRG validation tools (see Appendix C). Several interview questions (specifically questions 7, 8, and 9) explored trust calibration by asking participants how much confidence they placed in AI-generated outputs and under what conditions that trust was strengthened or weakened. Questions 6, 7, 9, and 10 examined the transparency and explainability of the system by probing participants' understanding of how AI decisions were made and how discrepancies were addressed. In addition, participants were asked to describe how the AI system and human coders interacted over time, how errors were communicated, and whether program improvement activities were integrated. These questions allowed for a comprehensive exploration of the framework concepts of co-adaptive learning and

feedback loops. Each interview question included multiple prompts that allowed the interviewer to surface organizational conditions for effective Human–AI teaming, such as data quality, governance, and training. The interview protocol operationalized the framework's key concepts and enabled a rich analysis of both technical and human-centered elements of system performance (Zhang et al., 2020).

The proposed instruments aligned with the project's purpose and qualitative design. The DRG validation audit protocol enabled structured analysis of AI output and identification of error patterns and financial impacts. The semi-structured interview guide complemented this data with professional insight into workflow integration, system limitations, and contextual enablers or barriers to accurate AI integration. Together, the instruments provided a rich understanding of both technical and organizational factors contributing to AI-enabled DRG validation errors. Findings from the study informed future system design and governance.

Operational Definitions of Variables

This study followed a qualitative descriptive design. Traditional operational definitions of variables were not applicable, as the focus was on rich description of participant perspectives rather than measurement of predefined constructs (Sandelowski, 2000). The study explored patterns and perspectives that emerged from stakeholder interviews and were guided by the Human–AI Interaction Framework. Constructs of interest, such as trust calibration, explainability, and human–AI teaming, were examined through thematic analysis rather than variable measurement.

Project Goals and Objectives

This section is not applicable to this project. A qualitative descriptive design does not include SMART objectives or quantifiable performance metrics. This study is guided by

exploratory project questions, focuses on thematic analysis of participant insights, and is consistent with accepted qualitative research practices (Sandelowski, 2000; Lincoln & Guba, 1985).

Metrics or Performance Measurements

This subsection is not applicable due to the qualitative descriptive design used in the study. The purpose of this study is to understand the nature, impact, and contributing factors to AI-enabled DRG validation errors through interviews and artifact analysis. The project will not assess the effectiveness of an intervention using numeric values. As such, performance measurement in the traditional sense of objective metrics is not appropriate. The rigor of this study is ensured through methodological coherence, reflexivity, and trustworthiness criteria such as credibility, transferability, dependability, and confirmability (Lincoln & Guba, 1985; Elo et al., 2014). These criteria serve as the qualitative equivalent to validity and reliability in quantitative work but do not equate to measurable performance metrics (Bradshaw et al., 2017).

Project Procedures

This study was conducted in six phases, as reflected in Table 6 below. This section provides an overview of the activities involved in each stage, along with an estimated duration (in months). A more detailed discussion of data collection and analysis procedures was provided in the subsequent section of the manuscript. The actual start and completion dates of each phase may have overlapped. For example, the DRG validation audit and participant recruitment occurred simultaneously. In addition, the data collection and analysis phases began before the recruitment phase was completed.

Table 2

Proposed Project Phases, Activities, and Duration

Phase	Activity	Anticipated Duration
1) DRG Validation Audit	a. Pull sample cases from data source b. Run cases through NLP-DRG validation software c. Cases reviewed independently by professional coders d. Comparison of coding outcomes and completion of tracking spreadsheet	2-3 months to conduct audit for 400 cases
2) Recruitment	e. Participant recruitment f. Eligibility determination g. Informed consent h. Schedule for Interview	1 month to schedule 10-15 total interviews
3) Data Collection	i. Conduct interviews j. Prepare transcripts k. Obtain participant review of and approval of transcript accuracy l. Data storage	1-2 months to obtain 10-15 validated interview transcripts
4) Data Analysis	m. Transcript review n. Initial deductive coding o. Inductive coding p. Thematic analysis q. Data saturation table	1-2 months to conduct analysis and achieve data saturation
5) Develop Findings, Implications and Recommendations	r. Comparison of findings with the evidence on Human-AI Interaction and AI-usage in RCM s. Development of practice and policy recommendations	1 month for completed Findings, Implications, and Recommendations section
6) Dissemination of Findings	t. Dissertation manuscript u. Oral defense v. Academic and Professional Conferences and Publications	1 month for development of final dissertation manuscript and oral presentation 12 months for publication in academic or professional journal

DRG Validation Audit

Audit procedures were discussed earlier in this section. A detailed discussion of the procedures, a section on sample cases, and an example of the results tracking spreadsheet are also contained in Appendix B. The results of the audit will serve as a data artifact for subsequent interviews and qualitative analysis.

Recruitment

As discussed previously, eligibility criteria for participants in this study were: (1) professionals involved in one or more of the following revenue cycle activities: inpatient coding, clinical documentation improvement, medical and claims auditing, and RCM program management; (2) participants were required to have at least three years of experience in hospital RCM; and (3) direct familiarity with NLP-enabled DRG validation tools. A purposive sampling technique was employed, beginning with the researcher's professional network of coding and CDI professionals and expanding to membership directories of the American Health Information Management Association (AHIMA) and the Healthcare Financial Management Association (HFMA) as needed. A total sample of 10 to 15 participants was recruited. Interested individuals were provided with an informed consent form detailing the study's purpose, confidentiality measures, and the voluntary nature of participation. The recruitment email and the informed consent document were presented in Appendices D and E, respectively. The researcher ensured each participant was eligible before beginning the interview. The recruitment phase lasted approximately one month.

Data Collection

A more detailed discussion of data collection and analysis procedures is contained in the following section. The data collection phase commenced upon scheduling the initial interview

for eligible participants. Interviews were conducted virtually using secure video conferencing software (e.g., Zoom) to ensure confidentiality. Each interview was expected to last 60–90 minutes. All interviews were recorded in video and audio formats and transcribed for data analysis. Participants were provided with a copy of the interview transcript to review and validate its accuracy prior to data analysis. Each participant was assigned a unique identifier to ensure their responses remained confidential. Any personally identifiable information (PII) was removed during transcription. All interview recordings and transcripts were stored on a password-protected computer.

Data Analysis

Qualitative data analysis employed both deductive and inductive coding methods (Fife & Gossner, 2024). The researcher reviewed each transcript to identify significant statements related to the project questions. An initial deductive coding scheme based on the Human-AI Framework (HAF) was presented in Appendix F. Interview statements were grouped into themes associated with the initial coding scheme. An inductive coding approach was also employed to identify emergent themes not initially covered by the HAF framework. Data analysis began at the time data collection started and ended when saturation occurred. Data saturation was determined when the analysis failed to uncover new themes among subsequent interviewees.

Development of Findings, Implications, and Recommendations

The systematic analysis of rich qualitative data was summarized into themes that reflected the lived experiences of the project participants and provided valuable insights and recommendations for healthcare organizations and for the field of healthcare administration. The implications of the findings were compared to the existing literature on the Human-AI

Integration Framework (HAF), NLP-enabled DRG validation errors, and RCM performance. The implications led to specific, actionable practice recommendations for enhanced human-AI integration and improved revenue cycle performance. In addition, the findings allowed for broader recommendations for the field of healthcare administration regarding AI integration in healthcare operational functions.

Dissemination of Findings

The findings, implications, and recommendations were documented in a comprehensive dissertation manuscript in accordance with National University guidelines and were presented at an open oral defense. To maximize the project's impact, feedback from the oral defense and the manuscript was used to develop conference presentations and publications in professional and academic journals focused on healthcare technology, AI in healthcare, revenue cycle management, or healthcare management.

Data Collection and Analysis

This section presents the data collection and analysis procedures for the qualitative descriptive study. The section begins with an introduction to the data source used for developing the DRG validation audit artifact. A discussion of audit procedures is contained in Appendix B. An overview of participant recruitment, data collection, and analysis procedures, including major tasks and the anticipated timeline, was presented in the project procedures section above. Data analysis followed a systematic and iterative process for this project.

Data Source for DRG Validation Audit

The data source for conducting the DRG validation audit consisted of de-identified inpatient data drawn from the Medical Information Mart for Intensive Care (MIMIC-IV) database, version 2.2 (Johnson et al., 2023). MIMIC-IV is a publicly accessible, de-identified database developed by the MIT Laboratory for Computational Physiology in collaboration with Beth Israel Deaconess Medical Center. The database contains comprehensive clinical data for over 70,000 unique intensive care unit (ICU) admissions from 2008 to 2019, encompassing a wide array of patient demographics, diagnoses, procedures, laboratory values, and billing information (Johnson et al., 2023). This data source was appropriate for the project's purpose because it offered high-volume, high-fidelity clinical and billing data representative of inpatient encounters and was contextually relevant for examining the accuracy and consequences of AI-enabled DRG validation systems. The nature of the MIMIC-IV dataset aligned well with the project goals of understanding the errors and financial risks introduced by NLP-based DRG assignment tools.

A stratified random sample of 400 inpatient cases was drawn from the MIMIC-IV dataset. Because MIMIC-IV data are de-identified and publicly available, recruitment of human participants for this part of the project was not required. The sample was selected to ensure coverage across major medical DRG categories (e.g., cardiovascular, respiratory, gastrointestinal, neurologic) to support the exploration of error types and financial implications across diverse clinical conditions. These sampled records were processed through proprietary in-house NLP software used for DRG validation. After DRG assignment, a manual audit was conducted to assess whether the AI-based coding system produced accurate DRG classifications. Identified discrepancies were then analyzed to estimate the associated economic consequences.

It was important to note that the DRG validation audit did not constitute a traditional quantitative analytical design involving inferential statistics, hypothesis testing, or group comparisons. As such, a power analysis was not applicable (Zhou et al., 2022). Results from the audit represented a structured quality assurance activity aimed at surfacing error patterns to inform the interview questions and qualitative analysis aspects of the project. The sample size of 400 records was consistent with published guidance for detecting meaningful error patterns in medical record reviews and ensured analytic sufficiency without incurring unnecessary processing burden (Zhou et al., 2022). Additional information regarding auditing procedures and results tracking was presented in Appendix B.

Qualitative Data Collection and Analysis

The goal of this project was to understand why NLP-enabled DRG validation errors occur and how human–AI interaction can be enhanced to improve revenue cycle performance. To achieve this goal, the researcher interviewed 10–15 professionals about their experiences with NLP-enabled DRG validation software, coding errors, and the program's workflow. Sample selection and eligibility were presented earlier. This section discusses the steps involved in data preparation and review, the use of deductive and inductive coding methods, triangulation efforts, and the role of the researcher in this process.

Audio recordings of all interviews were transcribed, and the researcher carefully reviewed and compared each transcript against the video and/or audio file to confirm accuracy. A copy of the transcript was provided to each participant to ensure accuracy. After review, transcripts were then anonymized, and any personally identifiable information was removed.

Deductive and Inductive Coding

NVivo, a qualitative analysis software (Lumivero, 2020), was used for thematic analysis to discern underlying organization and themes relevant to the Human-AI Integration Framework (HAF), the conceptual underpinning for understanding user interaction, trust calibration, and system explainability (Vasconcelos et al., 2023; Kim et al., 2023). The advantages of using qualitative data analysis software included maintaining the data digitally and conducting multiple types of coding as required (Maher et al., 2018). The initial deductive coding scheme based on the HAF framework was presented in Appendix F. After applying the deductive codes to the transcripts, the researcher adopted an inductive approach to conduct a second review. During this review, the researcher sought unexpected ideas, experiences, and nuances that emerged from the data. Any novel concepts that emerged were assigned new codes, thereby refining the initial coding scheme. The two-tiered coding strategy, deductive followed by inductive, ensured that the analysis remained aligned with the project's theoretical framework while also allowing participants' voices to be included and potentially reshaping the direction of the analysis.

Thematic Summary and Data Saturation

The coding results of each interview were recorded on a data analysis summary worksheet. An example of this worksheet was presented in Appendix G. The summary tool was designed to track the identification and repetition of themes within the data. The summary tool identified the number of themes in each interview, and the number of new themes introduced that had not appeared in previous interviews. Data analysis began when data collection began and ended when saturation was reached. Data saturation was achieved when the analysis failed to uncover new themes among subsequent interviewees.

Triangulation Efforts

Although this project relied primarily on interview data, triangulation was achieved through multiple strategies to enhance trustworthiness. First, the researcher used field notes from the interviews to provide contextual depth and to confirm interpretations. These notes captured nonverbal cues, emotional tone, or environmental factors influencing the conversation. Second, member checking (Fernandez et al., 2021) involved returning preliminary themes or summaries to select participants, inviting their feedback and adjustments to ensure their perspectives were accurately interpreted. Third, the DRG validation audit results served as a data artifact to triangulate with qualitative insights gathered from interviews. Using the data artifact as one triangulation tool allowed the investigator to compare observed error patterns with stakeholder perceptions and experiences. Comparing interview data with audit results enhanced the credibility of the findings by validating themes across independent data sources (Patton, 2015). Finally, the investigator regularly debriefed findings with dissertation committee members to identify and challenge any assumptions or interpretations of the data and to reduce the opportunity for researcher bias.

Role of the Scholar-Practitioner

The researcher was a scholar-practitioner and conducted all phases of the project, including recruitment, data collection, and analysis. The researcher was an experienced coding and revenue cycle professional with direct experience working with NLP-enabled DRG validation software. As such, the researcher was uniquely positioned to understand the participants' experiences and perceptions. Similarly, the researcher's own perceptions and experiences may have introduced personal biases or assumptions that required careful management. To mitigate potential bias, the researcher engaged in continuous reflection by

maintaining a journal to document personal responses, preconceptions, and any emotional reactions that arose during participant interviews, coding of interview transcripts, and analysis of the data (Berger, 2015). Through routine reflection, the researcher was better able to distinguish between their own personal experiences and those of interview participants (Finlay, 2002).

Assumptions , Limitations, and Delimitations

Assumptions

Assumptions are elements of the study accepted as true or accurate without direct evidence or verification (Marshall & Rossman, 2016). For this project, several foundational assumptions are inherent and necessary for guiding the inquiry process and are consistent with those made in other validations of AI and in audit accuracy studies (Montroy et al., 2023; Vasconcelos et al., 2023). First, it is assumed that the revenue cycle professionals participating in the project will possess valuable experiential knowledge and insights regarding the implementation and performance of AI-enabled DRG validation tools, which can be accessed through thoughtfully constructed interviews (Bradshaw et al., 2017). The second assumption is that participants will be willing to share honest, accurate, and reflective accounts of their experiences and perceptions, and that these accounts are trustworthy representations of their professional realities (Patton, 2015). Third, it is assumed that the source for the inpatient cases used to conduct the validation audit accurately represents real-world clinical documentation, diagnostic coding, and discharge status. Though anonymized, the MIMIC-IV dataset is derived from real ICU patients and is a commonly employed dataset in clinical informatics research published in peer-reviewed venues (Johnson et al., 2021).

Limitations

Limitations in the context of qualitative descriptive research designs are potential weaknesses or constraints beyond the researcher's control (Colorafi & Evans, 2016). Qualitative descriptive research designs are useful for capturing rich, detailed accounts of participants' experiences. The research design is inherently subject to certain limitations. A common limitation is reliance on participant self-reporting, which can introduce recall bias, social desirability bias, or selective memory, potentially distorting data accuracy (Bradshaw, Atkinson, & Doody, 2017). Second, the generalizability of the study's findings is limited because the interview data are context-specific and derived from a small, non-random sample (Colorafi & Evans, 2016). Third, the interpretive nature of qualitative analysis means that the researcher's own assumptions and positionality can influence data interpretation even with strategies in place to enhance trustworthiness (Vaismoradi, Turunen, & Bondas, 2013). Finally, and specific to this study exploring stakeholder perceptions of NLP-enabled DRG validation tools, there is also a risk that participants may not have a full understanding of the technical underpinnings of AI systems, which could limit the depth of some responses. Despite these limitations, the project's design offers a valuable opportunity for developing an initial understanding of an underexplored phenomenon: Human-AI interaction in DRG validation and its impact on revenue cycle performance. Findings could generate key insights for inquiry or system refinement.

Delimitations

Delimitations are the intentional boundaries set by the researcher to define the scope of the study, and reflect choices made to ensure feasibility, relevance, and alignment with the research purpose (Simon, 2011). Delimitations are inherent to both the research design and the need to craft a feasible dissertation project that can be completed within the university timelines

(Bloomberg & Volpe, 2019). In this study, several delimitations shape the project's design and execution. First, the study is limited to professionals with experience in hospital revenue cycle management (RCM) systems that use natural language processing (NLP)-enabled diagnosis-related group (DRG) validation tools. This choice narrows the participant pool but ensures that insights are grounded in direct and relevant user experience. Second, the study excludes outpatient or non-hospital-based settings. This delimitation may limit the transferability of findings, but it is necessary to maintain focus on the project's goal of understanding coding errors in high-volume inpatient coding environments. Third, the study adopts a qualitative descriptive design and does not attempt to quantify relationships or test hypotheses. The descriptive design is appropriate given the exploratory nature of the project questions (Bradshaw, Atkinson, & Doody, 2017). These delimitations reflect the researcher's aim to produce actionable, context-sensitive insights while balancing practical constraints such as time, resource availability, and access to data and participants. By clearly defining these boundaries, the study maintains conceptual clarity and methodological coherence (Bloomberg & Volpe, 2019).

Ethical Assurances

This project underwent full review and received approval from the Institutional Review Board (IRB) at National University prior to the initiation of any data collection activities. In alignment with ethical research standards outlined by the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979), the study adhered to the principles of respect for persons, beneficence, and justice. The study posed minimal risk to participants. Participants provided their professional perspectives and experiences working with NLP-DRG validation systems and were not asked to share sensitive or personal information.

Priority was given to maintaining confidentiality, obtaining informed consent, and ensuring voluntary participation. Each participant was provided with an IRB-approved informed consent form (Appendix E) that outlined the purpose of the study, their right to withdraw at any time, and the steps taken to protect their identity and responses (Orb, Eisenhauer, & Wynaden, 2001). No personally identifiable information was collected during interviews, and participant identities were coded using pseudonyms. In accordance with best practices in qualitative research ethics, any identifying references that emerged in transcripts were removed during transcription and member-checking processes (Kaiser, 2009).

Confidentiality and secure data management were paramount. All electronic data, including Zoom recordings, transcripts, audit documents, and coding summaries, were stored on password-protected, encrypted devices accessible only to the researcher. Any printed documents or notes were kept in a locked file cabinet within a secure, private location. In line with National University IRB policy, data are retained for 3 years following the completion of the study, after which they will be permanently deleted or shredded. The researcher completed the University's required human subjects research ethics training.

The Role of the Scholar-Practitioner

The researcher brings over 20 years of industry experience in health information management, medical coding, DRG validation, and mid-revenue cycle leadership to this project. While this background provides valuable context, it may also introduce biases in the interpretation of participant responses or audit results. To mitigate this risk, the researcher maintained a reflexive journal to document assumptions, positionality, and potential reactions throughout data collection and analysis (Berger, 2015). Regular reflexivity, coupled with

triangulation between qualitative interviews and the structured audit artifact, enhanced analytic transparency and trustworthiness. Additionally, the use of member-checking and audit trails further reduced the potential influence of researcher bias on study findings (Lincoln & Guba, 1985). Finally, the researcher regularly discussed data collection, coding, and thematic analysis with members of the dissertation committee as another method to control for potential bias.

Summary

This applied doctoral project utilized a qualitative descriptive design to investigate the effect of artificial intelligence-facilitated natural language processing (NLP) tools on DRG validation workflows in hospital revenue cycle management (RCM). The research was situated within the Human-AI Integration Framework (HAF), which takes a systemic approach to the interaction between human practitioners and AI systems. This project investigated the characteristics of DRG validation errors associated with the use of NLP/AI tools, including contributing organizational and technical factors, and proposed process improvements necessary to enable consistent, dependable integration of AI solutions into RCM processes.

The second section presented a two-stage research process. Phase one consisted of a retrospective audit of 400 de-identified inpatient cases, conducted using a structured DRG validation process. This audit served as a descriptive baseline to provide context and identify frequent errors, such as diagnostic omissions, sequencing errors, and AI misclassification. The audit results informed the qualitative analysis and guided the interview protocol in the second phase of the research. The second phase consisted of semi-structured interviews with ten to fifteen coding, auditing, CDI, and revenue cycle leadership professionals. These participants, all with practical experience with NLP-enabled tools, provided insight into how AI-generated errors are detected, interpreted, and managed in practice. The interview protocol (Appendix C) was

directly mapped to the project questions and to key HAF constructs, including trust calibration, task allocation, explainability, and feedback processes.

A qualitative descriptive design was well-suited for this inquiry, as the project focused on portraying operational realities, stakeholder perspectives, and practical implementation experiences. The audit results were not treated as outcome measures; rather, they served as context-rich stimuli for exploring human-AI interaction within revenue cycle workflows. The integration of descriptive audit findings with stakeholder experiences provided a structured approach for identifying barriers and facilitators to the successful use of AI in RCM.

By avoiding a mixed-methods or more quantitatively oriented structure, the study remained closely aligned with its conceptual framework and research questions. Rather than assessing causal relationships, the study focused on examining contextual factors, patterns of relationships, and organizational responses. The findings from this work informed actionable, human-centered recommendations to improve the accuracy of DRG validation and to integrate AI tools into real-world healthcare revenue cycle operations. This section provided the foundation for Section Three, which reported and discussed empirical findings interpreted through the HAF lens and presented implications and recommendations for health administrators, technology developers, and policymakers.

Section 3: Findings, Implications, and Recommendations

The goal of this applied doctoral project was to investigate the nature, frequency, and operational impact of errors in AI-generated natural language processing tools used for Diagnosis-Related Group validation workflows, as well as the human and organizational influences involved in their detection and management. Following the Human–AI Integration Framework, the present study uses organized artifact analysis, comprising both audited inpatient encounters and qualitative interviews with stakeholders, to generate data-driven findings that help guide human–AI collaboration strategies in healthcare revenue cycle operations. The research findings of this project are shown in different parts of the paper. The synthetic case examples employed in artifact analysis are first presented, and the development of the new examples is explained. The coding audit and the thematic analysis of the stakeholder interviews are subsequently presented. Finally, findings are discussed in terms of the three project questions and implications for healthcare revenue cycle practice and administration.

The study's project design initially suggested using the MIMIC-IV clinical database as the primary source of inpatient encounter data for artifact analysis. The goal was to build a case-level analytical dataset to identify common error patterns in artificial intelligence-enabled DRG validation workflows. However, during project implementation, technical constraints arose due to the modular structure of the MIMIC-IV dataset. The database organizes clinical data across multiple tables linked by the identifiers subject_id, hadm_id, and stay_id; variations in identifier availability and table structure across dataset releases complicated the process of building a reproducible case-level analytic file (Johnson et al., 2022; Johnson et al., 2023; Johnson et al., 2024).

These limitations posed risks of duplicate observations or incomplete encounter records, as the database attempts to consolidate whole inpatient cases across non-joining records and many-to-many table relationships. These restrictions were of great concern for reproducibility and data integrity in coding audits. Since the project's purpose was to assess patterns of AI-related discrepancies in DRG validation at the case level, the failure to construct complete encounter records from the external dataset was considered a methodological limitation. To address this limitation while maintaining the project's analytic objectives, a synthetic case corpus was developed to simulate inpatient encounter documentation. The synthetic cases were designed to reflect realistic clinical documentation patterns, coding scenarios, and DRG validation opportunities observed in healthcare revenue cycle environments. Each case included structured clinical documentation elements sufficient to perform a DRG validation audit and evaluate potential discrepancies in AI-generated recommendations.

Synthetic health records have been used extensively in health informatics applications for methodological testing and system evaluation (Zhou et al., 2022; Peng et al., 2021). With synthetic datasets, researchers can develop replicable case scenarios while avoiding issues related to patient privacy or sensitive health data. For instance, platforms like Synthea offer fully synthetic electronic health records intended specifically for research and software testing (Walonoski et al., 2018). Likewise, the Centers for Medicare & Medicaid Services has made synthetic claims datasets available, such as the Data Entrepreneurs' Synthetic Public Use File, to facilitate the analysis and implementation of applications using data without revealing identifiable patient information (Centers for Medicare & Medicaid Services [CMS], 2013; CMS, 2024). Synthetic health data applications are described in reviews for use in algorithm testing,

training environments, and health information technology development (Gonzales et al., 2023; Rujas et al., 2025).

The synthetic cases in this effort were used as controlled analytical artifacts to systematically examine AI-related coding discrepancies. The structured design of the cases enabled evaluation of coding scenarios, documentation complexity, and DRG validation opportunities of the audit sample in a consistent fashion. This methodology maintained the analytical objective of identifying error patterns while increasing transparency and reproducibility in case construction.

Forty synthetic inpatient cases were generated for inclusion in the coding audit. The cases were carefully designed from different clinical scenarios in acute care hospital settings, including multiple different MDCs, different degrees of clinical complexity, and medical and surgical DRG classifications. Not a random population of patients, the case corpus would include an overview of clinical conditions and coding patterns that may create difficulties when coding DRGs. This approach is consistent with the project's aim of highlighting patterns of issues in the AI-related inpatient cases, which can be tricky, especially within a situation where we need to document nuances and coding decisions to properly allocate DRG (and so also the project in general), but with respect to their treatment and implementation.

The distribution and characteristics of the synthetic case corpus are summarized in Appendix G. For each major diagnostic category, clinical scenario, DRG type, complication or comorbidity level, and primary coding risk, the distribution and characteristics are summarized. The cases were adapted based on known coding problems, including selection of major diagnosis criteria, CC and MCC capture, detection of procedural complexity, and need for

clinical validation. The patient observations and documentation methods were also considered, based on clinical guidelines, inpatient diagnosis history, and clinical history, to create an environment in which patients could be trained to use AI-based DRG validation tools.

To ensure real-world workflows are incorporated into the design and delivery of these cases, we adopted the standard inpatient documentation format: history and physical notes, progress notes, lab results, and discharge summaries. Our synthetic cases can serve as tools for us and are transparent to real-world cases. The real coding scenarios that the synthetic scenarios cover (simplistic and complex coding cases) could have been very useful, as we were hoping to understand, in practice, how the AI process and coding algorithms work.

Results of the Audit

Overall Error Rate

A structured audit was carried out on 40 synthetic inpatient case examples designed to model the documentation process used in a DRG validation workflow. Each case was manually tested to determine whether the results from the AI-simulated inpatient test set matched those derived manually using coding techniques, automated level CDI, and documentation standards.

In the 40 cases reviewed, 12 (30%) AI-generated recommendations were detected compared with manual diagnosis. There were twelve discrepancies across all cases considered in the audit, all of which could be attributed to a single variance. These findings demonstrate that AI-based DRG validation worked well in most tests, but only about one-third of cases could be manually tested and verified.

Distribution of Errors by Clinical Category

Not all errors were distributed evenly among clinical categories. Rather, multiple Major Diagnostic Categories were considered to have discrepancies. The greatest concentration of discrepancies occurred in cases classified as diseases and disorders of the respiratory system, which had the most discrepant cases and large financial variance. Contributions were also noticed in MDCs associated with infectious and parasitic diseases, systemic or unspecified sites, diseases and disorders of the female reproductive system, and chest pain.

Respiratory system cases correspond to a $5,343 variance per discrepant case, and infectious and parasitic disease cases correspond to a $4,739 variance per discrepant case. Most cases involve the female reproductive system, with a $6,105 variance per discrepant case. Chest pain cases show relatively low variance at $815.67. The findings indicate that differences were most pronounced in clinically difficult contexts, where documentation complexity and relationships in diagnosis affect a DRG. There appears to be a clustering trend in how AI-generated discrepancies in DRG validation may play out when people suffer from an acute respiratory problem, have sepsis, or have multiple potential comorbidities that need to be related to the data and evaluated in the context of each of these problems.

Structured Error Types

A root cause analysis was done for each discrepancy identified from our audit. AI-generated recommendations, together with manual coding selection, were tested in the RCA, with differences against ICD-10-CM coding guidelines and against DRG assignment and CDI query specifications measured. Every discrepancy was categorized and sorted by the key factor that created it.

We have identified four main structured errors:

1. Missed or incorrect query opportunities (n = 5).
2. Sepsis-related query errors (n = 4).
3. Missed complication or comorbidity capture (n = 1).

4. Principal diagnosis resequencing errors (n = 2).

Missed or incorrect query opportunities were the most frequent source of discrepancy. These occurred when documentation indicated a clinical condition but did not meet established clinical validation criteria. Sepsis-related query errors were the second most common and correlated with complexity in clinical indicator characteristics, such as changes in regulatory guidance for sepsis documentation and coding. These also demonstrated the highest mean financial variance. Principal diagnosis resequencing errors occurred when AI-generated recommendations were not consistent with the proper sequencing of diagnoses based on admission circumstances and coding guidelines. Missed complications or comorbidities were less frequent but remained clinically relevant because of their potential impact on DRG severity classification.

Financial Variance Observations

The financial variance between the AI-generated DRG recommendation and the manually assigned DRG was estimated as the difference between the reimbursed amount and the manually assigned DRG. Variance was calculated between the DRG relative weight differences based on a fixed base rate. The mean financial variance across all cases in the entire sample was $4,174, with a median of $1,685. The sepsis-related query error had a particularly high variance from the median, with a median of $7,685.25 per case. On the one hand, missed or incorrect queries led to a variance of $3,606, while on the other hand, resequencing errors for the main diagnosis accounted for only $1,317. While the variance between individual case-level financial experiences was quite different, the variance across cases has an important impact on operations.

We can use manual checks to verify the assignment of DRG based on the data in this report to meet payer requirements.

Summary of Audit Findings

All AI-based DRG validation tools correctly classified most cases, yet thirty percent required at least one manual correction. Most discrepancies were likely due to clinical problems involving respiratory disease or sepsis documentation, not those without manual correction. Most discrepancies are tied to the interpretation of clinical documentation, not just to simple code selection. In this synthetic audit of forty cases, the large number of cases requiring manual correction and the concentration of discrepancies in clinically complex scenarios were supported by qualitative interview results about the importance of human oversight and the limitations of DRG-first AI logic.

Results of Thematic Analysis

To further enhance the audit findings, semi-structured interviews were conducted with ten practitioners working in revenue cycle management, including coding, clinical documentation improvement, auditing, and revenue cycle leadership. The purpose of these interviews was to gather practitioner perspectives on the causes, detection, and operational implications of AI-enabled validation errors. Data from interviews were analyzed thematically, deductively, using a framework rooted in the Human-AI Integration Framework.

The thematic analysis used a mix of deductive and inductive approaches. We were aware that we would develop the first coding methodology based on the nine constructs of the Human-AI Integration Framework, which will be presented in Section 3. We developed four other inductive themes from the data: structural system limitations, rule-driven logic, ethical and bias

issues, and the need for change management and organizational readiness at the beginning and the end. Thirteen codes were selected from the 10 interviews overall. Table 3 shows the distributions of these constructs across all interviews. The matrix clearly identifies core Human-AI Integration Framework constructs and supports the finding that thematic saturation is an outcome of interview eight.

Table 3

Distribution of Deductive and Inductive Themes Across Interviews

Theme / Construct	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	Total
Role and Task Allocation	X	X	X	X	X	X	X	X	X	X	10
Trust Calibration	X	X	X	X	X	X	X	X	X	X	10
Transparency and Explainability	X	X	X	X	X	X	X	X	X	X	10
Feedback Loops and Error Recovery	X	X	X	X	X	X	X	X	X	X	10
Co-Adaptive Learning	X	X	X	X	X	X	X	X	X	X	10
Data Quality and Governance	X	X	X	X	X	X	X	X	X	X	10
Workflow Integration	X	X	X	X	X	X	X	X	X	X	10
Impact on Performance Metrics	X	X	X	X	X	X	X	X	X	X	10
User Experience and Usability	X	X	X	X	X	X	X	X	X	X	10
Structural System Limitations	X	X	X	X	4						
Rule-Based Logic Constraints	X	X	2								
Ethical and Bias Considerations	X	X	2								
Change Management and Organizational Readiness											

Theme Frequency and Distribution

Of the AI-generated recommendations we asked people to discuss, the most common error was false positives, with systems recognizing diagnoses or query opportunities that were not clinically supported or not suitable for coding. Participants found that errors are time-consuming and require review before coding implementation. We believe the findings of AI-

generated recommendations that relied on keyword detection are consistent with a rule-based logic environment and transparency and explainability concepts in the Human–AI Integration Framework.

Our researchers also reported difficulties with interpretational issues of context in clinical documentation. Natural language processing tools for identifying keywords to help suggest the correct keyword or phrase, and where and why in terms of clinical context, presented to investigators were described to be biased in terms of identifying language and context to interpret data points, and the algorithms used had no evidence of context in clinical patient presentations that we asked for diagnosis as we used them as well. This is most apparent in sepsis, hyponatremia, and congestive heart failure, where there are documentation nuances that influence coding. This is clearly a problem in identifying transparency and explainability, as well as in understanding how AI systems process the associations among clinical indicators.

A low degree of documentation quality and variation was also cited as a contributing factor to AI-related discrepancies, the authors also stated. Based on how physician documentation is practiced across providers and organizations, phrasing and dictation mistakes, as well as template use, influence it. Several respondents stated that AI tools are built on structured, consistent documentation, which is not aligned with clinical practice. A well-integrated system of data quality and governance was observed within the Human–AI Integration Framework.

Differences by Stakeholder Perspective

Most participants reported the same results, but they made different choices regarding professional involvement. Coding and auditing practitioners were more concerned with query

quality, sequencing reasoning, and coding standards. Closer to the heart for clinical documentation improvement professionals was the importance of documentation and the performance of systems. Revenue cycle leaders explained their business models for AI-related discrepancies: denial risk, staffing requirements, and workflow efficiency.

While the focus varied across our research groups at that time, we all recognized the importance of having a human review of AI-related recommendations as they emerged and of reviewing them with human input. They all agree that the AI outputs cannot be implemented in isolation, and that we need humans around to monitor compliance with their coding standards and ensure regulatory compliance. This finding is another reason to continue asking us to implement human-in-the-loop testing in the automation enabled by AI-enabled DRG validation tools.

Data Saturation

Data saturation was reached in our thematic analysis. Repeated themes persisted across interviews, and no new thematic categories emerged in later interviews. The last interviews were mostly consistent with the findings from the previous ones, and therefore, we confirmed the existence of the top conceptual categories related to the project question. Given this pattern, we would not expect additional interviews to reveal any new themes beyond those already identified.

Summary of Thematic Findings

All in all, the thematic analysis uncovered the influence of technical system capabilities and organizational practices on the performance of DRG-enabled AI validation tools. Although

they acknowledged the potential to improve the efficiency of AI-driven processes in case identification, those who answered were clear and pointed out the importance of validating document quality, providing quality feedback, and decision-making from a more systematic perspective to improve the quality of such systems. These results are very qualitative, given the audit results, and indicate that specific factors also affect revenue cycle workflows.

Audit and thematic analysis also gave complementary perspectives on the performance and operational benefits of AI-based DRG validation tools. The audit provided a structured analysis of errors and their root causes, and stakeholder interviews described how errors are detected, interpreted, and managed within practice workflows for these processes. Collectively, they reflect the general technical and organizational context of AI-based DRG validation, and we then provide a quantitative and operational response to the project question at hand.

Findings by Project Question(s)

PQ1. What are the most common types of errors identified in NLP/AI-driven DRG validation for high-volume and high-error episodes of care?

The combined findings of the audit and stakeholder interviews show that several recurring error types characterize the AI-enabled DRG validation discrepancies. As the audit illustrated, errors were localized to a subset of clinically complex cases and involved the contextual interpretation of documentation (rather than merely code selection). Thematic data from the interviews also contributed to an understanding of the operational circumstances in which these discrepancies occur.

Across both data sources, false-positive identification was the most frequently cited error pattern. Participants repeatedly observed that the AI systems often flagged diagnoses or query

opportunities that were not clinically supported when examined in a full clinical context. These results concur with the coding audit results, which found that a meaningful proportion of discrepancies involved inappropriate or unnecessary query opportunities generated by the system.

An additional common error scenario was the context-bound interpretation problems of natural language processing (NLP) engines. Participants in the interviews reported situations in which the system identified keywords or separated documentation details, but did not identify the clinical associations necessary to make coding decisions with precision. It was found that the contextual interpretation was particularly problematic in sepsis, hyponatremia, and respiratory failure, where documentation nuance significantly influences DRG assignment.

The audit results also included errors in the capture of complications and comorbidities, as well as in the sequencing of principal diagnoses. These inconsistencies were typically observed when AI-generated recommendations failed to align with established coding guidelines for appropriately sequencing diagnoses or with the clinical significance required for complication or comorbidity designation. While much less common than false-positive identification, these types of errors carry significant consequences for the classification of case severity and degree of reimbursement.

The results from this project contribute to the existing literature on AI-enabled coding tools by raising awareness of the types of discrepancies that are more likely to occur in real-world revenue cycle scenarios. Previous research shows that NLP-based systems can improve documentation capture and case assignment prioritization (Futterman et al., 2023), but the present analysis reveals that context remains a major challenge. The findings support the role of

a human touch in DRG validation workflows and reinforce the Human-AI Integration Framework's focus on a well-defined division of roles by identifying the roles and responsibilities of automated systems and human experts.

Through the Human-AI Integration Framework, the findings indicate that existing AI-driven DRG validation systems perform better at case identification and prioritization than at autonomous decision-making. The recurring error types identified in this study underscore the need for human reviewers to interpret data contextually, code according to the guidelines used, and validate that system recommendations adhere to user best interests before reaching final coding decisions.

PQ2. What operational and financial impacts are associated with errors in AI-enabled DRG validation workflows?

Research findings and thematic analysis of the audit suggest that errors in AI-powered DRG validation processes have both operational and financial impacts on healthcare revenue cycle management. Although the audit demonstrated measurable financial variances associated with certain discrepancy types, an important issue emerged during stakeholder interviews: the burden of monitoring and fixing AI-recommended approaches.

The assessment showed that variations in query generation, principal diagnosis sequencing, and comorbidity capture may affect DRG assignment, thereby further impacting reimbursement outcomes. For clinically complex conditions, such as those with sepsis documentation or severity classification, the potential financial variance would have been larger because these conditions shape case severity and DRG categorization. Findings in this context

align with previous research, which found that documentation specificity and diagnosis sequencing were significant contributors to revenue cycle performance (Futterman et al., 2023).

Despite financial inconsistencies observed in some cases, participants invariably indicated that the operational context of AI-based discrepancies extends beyond changes in reimbursement. The respondents in the interviews said that manual verification of AI-generated recommendations is now an integral part of their workflow. Reviewers frequently need to read the entire clinical documentation and determine whether a suggested diagnosis, query opportunity, or sequencing change is appropriate in accordance with known coding and query guidelines. This added review process increases the workload of coding, CDI, and auditing teams and may prolong case processing timelines. Participants also stated that repeated exposure to false-positive recommendations can erode reviewers' trust in AI outputs. This is likely to erode the efficiency gains initially expected from automation if reviewers presume that a large share of system-generated alerts will need to be cleaned up. Several participants discussed tasks in which reviewers would need to screen many AI-generated suggestions before determining which are valid both clinically and procedurally.

At the Human-AI Integration Framework level, these results indicate the importance of achieving an appropriate balance between system efficiency and appropriate human supervision. It focuses on the importance of clearly defined task allocation, transparent system behavior, and appropriate calibration of user trust for achieving successful human-AI cooperation. The issues in the operation described above indicate that AI-based DRG validation tools currently operate most effectively as case identification and prioritization mechanisms rather than as autonomous decision-making systems.

The findings add to the literature by indicating that the relevance of AI-enabled revenue cycle technologies cannot be reduced to financial return on investment. Review workload, reviewer confidence, and workflow integration are operational considerations as well to see whether the final efficacy of AI systems is beneficial or problematic for revenue cycle performance. Hence, an organization with AI-enabled DRG validation tools in place should use financial and operational metrics when determining its system's performance.

PQ3**.** What organizational and workflow-related factors contribute to the occurrence or persistence of errors in AI-enabled DRG validation workflows?

This thematic analysis identified numerous organizational and workflow factors that influence the occurrence and persistence of discrepancies in AI-enabled DRG validation systems. Although the audit reported on specific error types and their associated causes, interview results were very informative about how organizational practices, documentation patterns, and governance processes influence the operational performance of these technologies.

Variation in clinical documentation practices was identified as one of the most frequently cited factors contributing to the variation. As participants stated, frequently, physician documentation also differs in terms of phrasing, structure, and clinical specificity. Due to the significant dependency on text pattern recognition, disparities in documentation might affect the interpretation of clinical concepts by the AI system. Participants described cases in which incomplete documentation, ambiguous terminology, or template-driven phrasing led to system-generated recommendations that required manual correction. These findings imply that documentation quality and standardization are pivotal in determining the accuracy of AI-enabled DRG validation tools.

The second aspect was the transparency and effectiveness of the feedback processes used to enhance system performance. Several participants indicated that discrepancies discovered during coding or audit reviews were sometimes reported to system developers or vendors; however, end users did not see their involvement in the consideration of feedback during system update development. This lack of visibility sometimes led to confusion about whether previously identified errors were corrected. Participants remarked that structured feedback mechanisms and communication around system upgrades could both enhance system performance and user trust.

Participants further stressed that clear roles and responsibilities should be defined in human-AI collaborative workflows. Coding professionals, clinical documentation improvement specialists, and revenue cycle leaders all use AI-enabled tools in slightly different ways, and varying role responsibilities may affect which discrepancies are detected and remedied. The interviewees' responses showed that human reviewers emerge as the final decision-makers in coding determinations, reinforcing the need for human oversight in regulatory and compliance-sensitive environments.

Trust calibration was subsequently identified as another organizational driver of system usage. Participants described different kinds of confidence in AI-generated recommendations, with users reporting varying degrees of comfort with the algorithm's advice, depending on case complexity and prior experience with the system's accuracy. In environments where false-positive recommendations were perceived as frequent, reviewers showed greater skepticism toward system alerts and more manual validation. Such findings echo previous research suggesting that user trust in AI systems needs to be appropriately calibrated for effective human–AI collaboration (Sahni & Carrus, 2023).

Utilizing the Human-AI Integration Framework, it is evident that organizational readiness, governance structures, and workflow design are the foundational determinants of the outcomes of AI-enabled revenue cycle technologies. Such a theory stresses that, to integrate AI systems successfully, it is not enough to have technical capability; the system's functionality must also align with user roles. The theme of this study was that AI-enabled DRG validation systems work best when supported by strong documentation practices, transparent feedback loops, and clearly defined human oversight mechanisms.

Collectively, their findings underscore the need to understand AI implementation as a socio-technical process rather than a purely technological intervention. The interpretability and implementation of AI-produced outputs in practice are predominantly shaped by organizational factors related to documentation governance, workflow integration, and trust calibration.

Evaluation of Outcomes

This section does not apply to this applied project. This study used a qualitative descriptive design to understand patterns of AI-enabled DRG validation errors and stakeholders' experiences. The project did not develop or implement an intervention with predefined performance targets, nor did it establish SMART objectives or quantitative benchmarks (Sandelowski, 2000). The findings from this qualitative descriptive study were synthesized earlier. Implications and recommendations for coding professionals, vendors, and revenue cycle governance will be discussed in detail below.

Action Plan

The "Action Plan" subsection does not apply to this applied project because no discrete intervention or change program was designed, implemented, or evaluated (Colorafi & Evans, 2016). The goal of this qualitative descriptive study was to characterize patterns of AI-enabled DRG validation errors and stakeholder experiences, and to generate practice-oriented insights to address the errors. The practical implications of the study findings are discussed below.

Implications and Recommendations for Practice

Through this applied doctoral project, we have collected a wide range of practice insights for healthcare organizations on the use of artificial intelligence-enabled DRG validation tools within revenue cycle operations. Even though the findings were based on a defined audit sample and stakeholder interview group, the patterns observed have broader implications for organizations considering AI technologies for coding, clinical documentation improvement, and revenue integrity workflows. The following recommendations are based on the project findings. They are interpreted through the lens of the Human-AI Integration Framework, which emphasizes the necessity of strong collaboration between automated systems and human experts.

Recommendation 1: Formalize Human-in-the-Loop Governance Structures

Some of the most salient findings from the audit and stakeholder interviews were the continued application of human review processes to validate AI-generated coding recommendations. Participants are also keen that any automated outputs should be reviewed for compatibility with clinical documentation, coding guidelines, and regulatory requirements prior

to their deployment. These results suggest that AI-enabled DRG validation tools work best as decision-support systems rather than as complete coding automation.

As such, healthcare organizations adopting AI-assisted revenue cycle technologies should formalize human-in-the-loop governance structures as a core aspect of system deployment. Specific governance models should clearly define reviewer roles, escalation paths, and quality assurance processes to ensure that automated recommendations are assessed uniformly and adhere to accepted coding standards.

Consistent with the literature, the authors concur that AI systems in healthcare work best when integrated with structured human oversight frameworks (Sahni & Carrus, 2023). Additionally, the Human-AI Integration Framework emphasizes the need for appropriate task assignment between human experts and automated systems to ensure accuracy and regulatory compliance. By establishing formal governance mechanisms, systems will be more reliable, and users will begin to trust AI-driven coding workflows more.

Recommendation 2: Strengthen Documentation Standardization to Support NLP Performance

Documentation variability was also identified as a key contributor to AI discrepancies through thematic analysis findings. Participants commonly mentioned inconsistencies in physician documentation patterns, which vary significantly across providers and organizations, thereby affecting NLP systems' analysis of clinical language. Where documentation is vague or unclear, AI-generated recommendations may require additional human interpretation.

If healthcare organizations are to enhance DRG validation performance through innovative AI technologies, they should focus on documentation quality and standardization

initiatives as part of their overall AI-driven implementation strategy. Clinical documentation improvement programs can help this effort along by providing providers with focused education on documentation clarity, diagnostic specificity, and the clinical relationships required for correct coding.

In fact, previous studies have shown that the quality and consistency of source documentation are critical elements for the performance of NLP-driven systems (Futterman et al., 2023). The results of this study add to the validity of this connection by demonstrating how documentation variability can contribute to AI-generated coding recommendations. Healthcare institutions can improve coding accuracy and enhance the efficiency of AI-assisted technologies by strengthening documentation governance best practices.

Recommendation 3: Establish Transparent Feedback Loops for AI System Improvement

Throughout the interviews, another concept that was raised was the importance of structured feedback systems to enhance system efficiency. The respondents mentioned that inconsistencies found during the manual review process are sometimes reported to vendors or system designers, but that how these rectifications affect system updates is not always apparent to end users.

Thus, healthcare firms should develop clear feedback processes that allow coding and CDI professionals, as well as auditors, to report AI-related variations in a uniform, structured manner. This feedback might encompass everything from identifying errors to mechanisms for reporting issues to system developers, and from communication regarding system updates or algorithm adjustments.

The principles of the Human-AI Integration Framework for both human users and automated systems are to evolve co-adaptively during learning interactions through interaction and feedback. Transparent feedback systems help organizations gather valuable operational feedback from coding professionals who live with AI-generated outputs. Ultimately, these insights could lead to better system performance and greater alignment of algorithmic recommendations with clinical coding and CDI work practice over time.

Recommendation 4: Expand Performance Evaluation Beyond Financial Metrics

The results of this study indicate that the effectiveness of AI-enabled DRG validation approaches should not be analyzed solely in terms of financial return on investment. Although the audit identified potential reimbursement variances associated with certain error types, interviewees highlighted operational issues, including review burden, workflow disruptions, and reviewers' uncertainty about AI. Healthcare providers need to adopt more comprehensive performance assessment frameworks that integrate financial and operational measures for evaluating AI system performance. Related to reimbursement results, organizations might benefit from tracking metrics such as false-positive rates, manual review volume, query resolution time, and reviewer satisfaction. Numerous existing works have argued that integrating AI technologies should consider both technical performance and organizational workflow factors (Sahni & Carrus, 2023). By examining both financial and operational outcomes aligned with the Human-AI Integration Framework, these factors can provide an in-depth understanding of the impact of AI technology on revenue cycle performance.

Recommendation 5: Incorporate Trust Calibration and Change Management Strategies

Finally, the thematic results indicate that user trust in AI-enabled systems significantly contributes to the adoption of such technologies in coding workflows. Respondents reported varying levels of confidence in AI-generated recommendations, based on prior experiences with the system's accuracy and their perceptions of the frequency of false-positive alerts. Healthcare entities implementing AI-enabled revenue cycle technologies should therefore embrace trust calibration and change management practices in their process implementation strategies. These strategies could include structured training programs, a clear description of systems' capabilities and limitations, and ongoing opportunities for users to provide feedback on system performance.

The Human-AI Integration Framework emphasizes that trust calibration is an important factor for successful human-AI cooperation. Users who understand the pros and cons of automated systems are better equipped to interpret AI-generated results correctly. Establishing realistic expectations for system performance may thereby aid more efficient use of AI tools within revenue cycle workflows.

Implications for the Field of Health Administration

The resulting applications gained from this applied doctoral study are also useful to wider literature in the field of health administration, focusing on the application of artificial intelligence technologies in complex healthcare operations. Although most of the literature on AI adoption in healthcare research has focused on clinical decision support or diagnosis, little attention has been paid to assessing the extent to which AI-enabled tools have been used in administrative contexts, such as revenue cycle management. Hence, this study's findings offer a view of the interactions among the organizational, technical, and personnel aspects of institutional systems during AI adaptation in operational healthcare settings.

Interpreting the results through the lens of the Human-AI Integration Framework (HAF) indicates that several of its central elements are evident in participants' accounts of their experiences. We concluded that well-defined role assignments between automated systems and human experts were a major factor in the transition from coding audit to thematic analysis. AI-assisted DRG validation systems were highlighted as valuable tools for detecting anomalies or triaging cases for review. Still, ultimate coding decisions remained dependent on human interpretation of clinical data and regulatory coding standards. These findings strengthen the framework's focus on task allocation and underscore the ongoing importance of human expertise in AI-assisted decision-making.

The results also bring out the importance of trust calibration and transparency between humans and AI when collaborating. Participants also stated that their trust in AI-suggested recommendations was affected by their perceptions of the system's output accuracy and the visibility of the feedback mechanisms used to improve system performance. These findings align with the claim that trust is not fixed but develops through ongoing engagement between users and automated systems, as articulated in the Human-AI Integration Framework. As such, when applying AI technologies in revenue cycle operations, there is a critical focus on managing user trust.

Moreover, the results imply that documentation governance and workflow integration may be prominent contextual factors influencing the performance of AI-enabled coding tools. While these dynamics are found in Human-AI Integration Framework elements of data quality and system integration, the findings of this project indicate that documentation variability and organizational workflow design may be particularly relevant to administrative AI applications. It

is hoped that future studies will explore how documentation practices, coding guidelines, and operational workflows will affect the effectiveness of NLP-driven systems used in healthcare administration.

Lastly, the proposed study emphasizes the importance of viewing AI implementation as a socio-technical process that combines technological capability and organizational readiness. The successful adoption of AI technologies is predicated not only on the precision of the algorithms but also on the design of governance systems, feedback channels, and the way in which human monitoring will occur. These results support the broader health administration discourse that emphasizes the importance of collaborative work across disciplines among technology developers, clinicians, coders, and administrative leaders for the deployment of new digital health technologies.

Collectively, these findings add to the burgeoning literature looking at how artificial intelligence might be strategically embedded in healthcare administrative settings. This project demonstrates how abstract models of human-AI interactions can be applied to decision-making in complex operational settings using the Human-AI Integration**.**

Framework for Revenue Cycle Management

Recommendations for Future Research/Projects

The findings of this applied doctoral project identified various areas that add value to knowledge and knowledge of how artificial intelligence is integrated into healthcare revenue cycle operations. Although this paper reveals the varieties of discrepancies and their relationships in AI-driven DRG validation workflows, and the organizational characteristics that influence their management, many opportunities exist to extend the scope of inquiry.

Future Research on AI Performance in Revenue Cycle Operations

For one, studies can be even more useful when the case sample is larger and more diverse, as they can enhance our understanding of the quality and effectiveness of AI-based DRG validation systems. This study employed a synthetic case corpus to emulate actual documentation practice in the real world. Although the methodology was transparent and free of privacy concerns, it might also provide better insight into how AI tools function across different healthcare settings and documentation contexts. More insight could be gained from other investigations with large datasets and multi-institutional case samples.

Further studies may also investigate the operation of these AI-based coding technologies across different DRG methodologies and clinical service lines. Discrepancies resulting from the clinical complexity of disorders and the level of documentation detail were also noted in the current study. Future studies could examine whether those relationships are valid in other settings, such as outpatient coding, risk adjustment models, and other DRG classification systems.

Research on potential biases in AI systems used for healthcare administrative decision-making is another significant area of inquiry. Natural language processing models are trained on past clinical documentation and coding data and may be rooted in current documentation trends, geographic area, institutional workflows, or provider-specific language patterns (Peng et al., 2021; Liu et al., 2022). Thus, recommendations from AI can inadvertently reflect the biases of training data or documentation practices. Additional studies can examine whether AI-based DRG validation systems introduce systematic variations in coding recommendations across patient populations, clinical specialties, or institutional documentation styles. Understanding how bias

occurs in administrative AI applications is necessary to ensure these tools ultimately advance equitable, compliant healthcare operations.

Finally, an ongoing investigation must be conducted into the long-term operational implications of AI-enabled revenue cycle technologies. Although this study focused on the immediate operational and financial effects of AI discrepancies, it could be further researched how organizations adapt to these technologies over time. Longitudinal studies also could explain how user trust, workflow integration, and system performance evolve as organizations gain experience with AI-assisted revenue cycle tools.

Addressing Implementation Challenges in Future Projects

Future applied projects may also address various challenges encountered during this study. One notable challenge was integrating external clinical datasets for artifact analysis. The challenge of developing reproducible case-level analytical datasets from complex health data repositories underscores the need for careful data governance planning in designing AI-focused research projects. Detailed data preparation strategies should be incorporated into project design for large, modular datasets or publicly available clinical repositories to support future researchers. Alternatively, synthetic health record datasets may continue to serve as a valuable resource for developing and testing analytical methodologies without exposing protected health information.

Future projects should also investigate more systematic ways of gathering operational feedback from coding and CDI professionals and other revenue cycle staff. Since human reviewers are instrumental in discovering and mitigating inaccuracies generated by AI, their perspectives constitute valuable evidence for enhancement. Implementing standardized rules to

document reviewer insights could help enhance research and optimize operational system improvement initiatives.

Conclusions

This applied doctoral project investigated the types of discrepancies observed in AI-enabled DRG validation workflows and the organizational factors that shape how they are identified and processed. Human-AI collaboration operational realities have emerged as a focal area of investigation in healthcare, especially as healthcare institutions increasingly adopt AI during clinical services and in administrative applications. While the use of AI-driven automation tools can lead to greater efficiency and more accurate identification of opportunities, applying these advances to complex healthcare workflows has also introduced challenges in system interpretation, governance, and human oversight.

The project aimed to investigate the type, extent, and operational implications of AI-related discrepancies in DRG validation processes, as well as the role and actions of healthcare professionals in the application of such technology. The analysis integrated an audit of simulated inpatient cases with qualitative interviews with professionals working in the revenue cycle coding, auditing, clinical documentation improvement, and revenue cycle leadership. By combining artifact analysis with stakeholder perspectives, the project was designed to cover all technical and organizational aspects of AI-enabled DRG validation workflows.

The audit shows that although there were no discrepancies in most cases reviewed, a significant proportion of AI-generated recommendations were manually corrected. Errors were most common for false-positive query detection, difficulties in contextual interpretation, and discrepancies in diagnosis sequencing or complication capture. These results further support the idea that AI-driven DRG validation tools can be useful for detecting potential opportunities. Still,

interpretation should be a careful process that aligns with coding standards and regulators' guidance.

The analyses of stakeholder interviews further revealed the organizational dimensions affecting the performance of these technologies. A frequent theme among participants was the importance of human oversight in reviewing AI-generated recommendations and in operational considerations for false-positive alerts. Differences in physician documentation techniques, context-specific limitations of NLP systems, and the availability of efficient feedback loops were also mentioned as major factors in system performance.

The results of the audit and thematic analysis also speak to the relationship between automated systems and human expertise in administration, which is interpreted simultaneously. AI-enabled DRG validation systems currently serve more as decision-support tools for revenue cycle professionals than as replacements for human decision-making, helping identify potential discrepancies and/or prioritize review cases. This is further supported by human oversight and governance of the use of artificial intelligence in regulatory and compliance-sensitive workflows.

The results confirmed the Human-AI Integration Framework, which claims that task allocation, trust calibration, transparency, and co-adaptive learning are critical for achieving optimal collaboration between a human and an AI team. Such is the human-in-the-loop process, where coders and CDI professionals interpret clinical documentation and apply industry standards. This is once again a reminder that, though automated systems have a great opportunity to add value by identifying patterns or potential discrepancies that may otherwise escape detection, human expertise remains a core component of AI-augmented environments.

From a professional practice perspective, the findings of this study underscore the importance of using AI technologies within well-defined governance frameworks that promote

transparency, feedback, and human oversight. Healthcare institutions that are implementing AI-supported revenue cycle tools must be vigilant to keep industry practitioners engaged in verifying system outputs and to have working mechanisms to communicate discrepancies and drive improvement in system performance over time. Enhancements to documentation procedures, tighter feedback loops between users and system developers, and performance evaluation metrics beyond financial return on investment can help facilitate the successful implementation of these technologies.

The work also contributes to the wider health administration community by exemplifying the practical operational realities of the interaction between artificial intelligence technologies and an organization's existing workflows, professional expertise, and governance structures in real-world settings. Healthcare institutions will also invest heavily in digital transformation projects, raising the expectation that administrators must ensure AI approaches are applied rationally and effectively. Thus, it is critical to understand the operational realities of human-AI collaborations so that policies, governance models, and implementation strategies can be devised to help ensure innovation and adherence.

While many insights from this study address the performance and operational aspects of AI-based DRG validation systems, these findings still need to be interpreted in light of the specific project design. This study relied on a synthetic case corpus and a defined set of stakeholder interviews, and subsequent research could deepen understanding of AI performance across diverse healthcare settings. However, these findings provide important insights into how AI technologies operate in revenue cycle operations and the role of balanced automation alongside human expertise.

The gradual infusion of AI into healthcare administrative processes will be considered an emergent socio-technical change rather than a purely technological one. Achieving AI-driven revenue cycle technologies is not just about algorithmic accuracy, but about organizational readiness, documentation quality, and effective human oversight, the results of this study suggest. Given healthcare organizations' continued interest in artificial intelligence, it will become critical to balance technological capabilities with professional competency.

The process of Artificial Intelligence can work effectively alongside existing processes to improve the healthcare revenue cycle. These systems need the human-in-the-loop process to continually audit system outputs, implement regulatory standards, and calibrate automated recommendations based on the clinical and administrative conditions of patient care. Integrating AI into operational environments enables a human-AI workforce to leverage emerging technologies more effectively while maintaining accuracy, transparency, and accountability in healthcare administration.

References

Aldoseri, A., Khalifa, K. N. A., & Hamouda, A. M. (2023). Re-thinking data strategy and integration for artificial intelligence: Concepts, opportunities, and challenges. *Applied Sciences*, 13(12), 7082. https://doi.org/10.3390/app13127082

Berger, R. (2015). Now I see it, now I don't: Researcher's position and reflexivity in qualitative research. *Qualitative Research*, 15(2), 219–234. https://doi.org/10.1177/1468794112468475

Bloomberg, L. D., & Volpe, M. (2019). *Completing your qualitative dissertation: A road map from beginning to end* (4th ed.). *SAGE Publications*.

Bradshaw, C., Atkinson, S., & Doody, O. (2017). Employing a qualitative description approach in health care research. *Global Qualitative Nursing Research*, 4, 1–8. https://doi.org/10.1177/2333393617742282

Charmaz, K. (2014). Constructing grounded theory (2nd ed.). *SAGE Publications*.

Chaturvedi, A., Rao, S., & Srinivasan, N. (2024). Enhancing medical billing with NLP-driven coding systems. *Journal of Health Information Science*, 11(2), 114–128.

Colorafi, K. J., & Evans, B. (2016). Qualitative descriptive methods in health science research. *Health Environments Research & Design Journal*, 9(4), 16–25.

Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). *Sage Publications*.

Creswell, J. W., & Poth, C. N. (2018). Qualitative inquiry and research design: Choosing among five approaches (4th ed.). *SAGE Publications*

Elo, S., Kääriäinen, M., Kanste, O., Pölkki, T., Utriainen, K., & Kyngäs, H. (2014). Qualitative content analysis: A focus on trustworthiness. *SAGE Open*, *4*(1), 1–10. https://doi.org/10.1177/2158244014522633

Fife, S. T., & Gossner, J. D. (2024). Deductive qualitative analysis: Evaluating, expanding, and refining theory. *The International Journal of Qualitative Methods, 23*. https://doi.org/10.1177/16094069241244856

Finlay, L. (2002). Negotiating the swamp: The opportunity and challenge of reflexivity in research practice. *Qualitative Research*, 2(2), 209–230. https://doi.org/10.1177/146879410200200205

Futterman, I., Friedmann, H., & Haberman, S. (2023). NLP- a tool to address documentation gaps and improve revenue. *American Journal of Obstetrics and Gynecology*, *228*(1), S656–S656. https://doi.org/10.1016/j.ajog.2022.11.1104

Gabel, T. J., Monahan, K., & Sharif, K. (2024a). Automating medical coding: A review of NLP use in revenue cycle workflows. *Healthcare AI Review*, 9(1), 42–55.

Gabel, A., Rahman, M., & McCormack, A. (2024b). AI-enabled automation in revenue cycle management: Trends and ethical considerations. *Journal of Health Information Management, 38*(1), 42–51.

Guetterman, T. C., Creswell, J. W., & Kuckartz, U. (2015). Using joint displays and MAXQDA software to represent the results of mixed methods research. In U. Kuckartz & S. Rediker (Eds.), Mixed methods research: From theory to practice (pp. 145–168). *Sage*.

Hennink, M. M., Kaiser, B. N., & Marconi, V. C. (2017). Code saturation versus meaning saturation: How many interviews are enough? *Qualitative Health Research, 27*(4), 591–608. https://doi.org/10.1177/1049732316665344

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2023). MIMIC-IV (version 2.2). *PhysioNet*. https://doi.org/10.13026/s6n6-xd98

Johnson, A. E. W., Stone, D. J., Celi, L. A., & Pollard, T. J. (2021). The MIMIC Code Repository: Enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association, 28*(3), 435–439. https://doi.org/10.1093/jamia/ocaa288

Kaiser, K. (2009). Protecting respondent confidentiality in qualitative research. *Qualitative Health Research*, 19(11), 1632–1641. https://doi.org/10.1177/1049732309350879

Kim, E. A. W., Russakovsky, O., Fong, R., & Monroy-Hernández, A. (2023). "Help me help the AI": Understanding how explainability can support human-AI interaction. *ArXiv*. https://doi.org/10.1145/3544548.3581001

Kristiansen, J., Chen, M., & Bhatia, R. (2022). Predictive analytics for claim denials: ML applications in healthcare finance. *Journal of Healthcare Informatics*, 17(3), 202–214.

Laux, J., Wachter, S., & Mittelstadt, B. (2023). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. Regulation & Governance, 18(1). https://doi.org/10.1111/rego.12512

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. *Sage Publications*

Liu, T., Chen, Z., Liu, Y., & Sun, J. (2022). Evaluating clinical natural language processing model robustness across institutions. *npj Digital Medicine, 5*, 1–9. https://doi.org/10.1038/s41746-022-00623-4

Lo, J., Long, M., Wallace, R., Salaga, M., & Pestaina, K. (2025, January 27). *Claims Denials and Appeals in ACA Marketplace Plans in 2023 | KFF*. KFF.

https://www.kff.org/private-insurance/issue-brief/claims-denials-and-appeals-in-aca-marketplace-plans-in-2023/

Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., & Ma, X. (2023). Who should I trust: AI or myself? Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 759. https://doi.org/10.1145/3544548.3581058

Marshall, C., & Rossman, G. B. (2016). Designing qualitative researc*h* (6th ed.). *SAGE Publications*.

McCormack, J. (2024a). Clinical documentation improvement and NLP: Aligning care, compliance, and coding. *Healthcare Management Perspectives*, 31(1), 56–64.

McCormack, J. (2024b). Large language models and the future of physician query automation. *Revenue Cycle Strategy Quarterly*, 6(2), 15–28.

Montroy, T., Patel, R., & Brooks, C. (2023). LLMs in healthcare: Enhancing administrative efficiency through generative AI. *Medical Informatics & Technology*, 28(4), 193–210.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

Nazi, K., & Peng, Y. (2024). AI in healthcare administration: Prior authorization and beyond. *Administrative Health AI Innovations*, 12(1), 77–89.

Orb, A., Eisenhauer, L., & Wynaden, D. (2001). Ethics in qualitative research. *Journal of Nursing Scholarship*, 33(1), 93–96. https://doi.org/10.1111/j.1547-5069.2001.00093.x

Patton, M. Q. (2015). Qualitative research & evaluation methods: Integrating theory and practice (4th ed.). *Sage Publications*.

Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R. M., & Lu, Z. (2021). Transfer learning in clinical natural language processing: A systematic review. *Journal of Biomedical Informatics, 114*, 103685. https://doi.org/10.1016/j.jbi.2021.103685

Root, T., & Samtani, S. (2025). Patient engagement and financial support using AI-powered chatbots. *Journal of Applied AI in Healthcare*, 5(1), 88–104.

Russell, D. (2024). Predicting and preventing fraud in the healthcare revenue cycle: Applications of ML and DL. *Journal of Medical Financial Analytics*, 13(2), 109–120.

Russell, D., Hayes, E., & Zhang, L. (2024). Automating appeals with LLMs: Case studies in payer communications. *Healthcare Automation Reports*, 7(1), 31–46.

Sandelowski, M. (2000). Whatever happened to qualitative description? *Research in Nursing & Health*, 23(4), 334–340. https://doi.org/10.1002/1098-240X(200008)23:4<334::AID-NUR9>3.0.CO;2-G

Sankaran, G., Palomino, M. A., Knahl, M., & Siestrup, G. (2022). A modeling approach for measuring the performance of a human-AI collaborative process. *Applied Sciences*, 12(22), 11642. https://doi.org/10.3390/app122211642

Shneiderman, B. (2020). Bridging the gap between ethics and practice. *ACM Transactions on Interactive Intelligent Systems*, 10(4), 1–31. https://doi.org/10.1145/3419764

Simon, M. K. (2011). *Dissertation and scholarly research: Recipes for success* (2011 ed.). CreateSpace.

Soroush, N., Blake, T., & Nguyen, P. (2024). Artificial intelligence in healthcare financial functions: Models and outcomes. *Journal of Digital Health Innovation*, 15(1), 1–15.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-

making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), 1–38. https://doi.org/10.1145/3579605

Vössing, M., Kühl, N., Lind, M., & Satzger, G. (2022). Designing transparency for effective human-AI collaboration. *Information Systems Frontiers*, 24, 441–461. https://doi.org/10.1007/s10796-022-10284-3

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–15. https://doi.org/10.1145/3290605.3300831

Yin, R. K. (2018). Case study research and applications: *Design and methods* (6th ed.). SAGE Publications.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 295–305. https://doi.org/10.1145/3351095.3372852

Zhou, L., Sordo, M., & Sordo, M. (2022). Secondary use of EHR data for quality measurement and improvement. *In Biomedical Informatics* (pp. 459–483). Springer. https://doi.org/10.1007/978-3-030-58747-0_14

Appendix A

Comparison of Core Artificial Intelligence Technologies

Comparison of Core Artificial Intelligence Technologies Used in Healthcare Revenue Cycle Management

Technology	Purpose	Input Type	Key Applications in RCM
Natural Language Processing (NLP)	Extract and interpret unstructured human language	Unstructured text (e.g., clinical notes)	Automated coding, clinical documentation improvement, denial trend analysis
Machine Learning (ML)	Learn from structured data to make predictions and classify patterns	Structured/tabular data (e.g., billing history)	Predictive claim denials, fraud detection, authorization validation
Deep Learning (DL)	Model complex relationships in large datasets	Structured or semi-structured data	Complex pattern recognition, anomaly detection, automation of multi-step tasks
Large Language Models (LLMs)	Generate, summarize, and interpret human-like text	Massive corpora of text (multi-domain)	Appeal letter generation, chatbot support, AI-assisted physician query generation

Adapted from Chaturvedi et al. (2024); Montroy et al. (2023); McCormack (2024a, 2024b); Soroush et al. (2024); Russell et al. (2024).

Appendix B

DRG Validation Audit Procedures

DRG Validation Audit Design

A structured audit methodology will be undertaken to assess the performance of an NLP-driven DRG validation software. The audit will compare software outputs to the manual review of professional coders. The audit procedures mirror common industry DRG validation practices and include methods to validate, fix, and categorize software performance. Results of the audit will be used as a data artifact for further qualitative inquiry to understand how workflow processes and human-AI interaction could be enhanced to improve system performance.

Sampling and Case Selection

A stratified random sampling approach will be used to extract 400 inpatient case data from the MIMIC-III database. The stratification will be designed to ensure representation of a range of Major Diagnostic Categories (MDCs), high volume MS-DRGs, and DRGs that are known to have a high rate of error or denial. Exclusion criteria are outpatient visits, visits with incomplete data, or records lacking a coded DRG assignment.

Audit Preparation and Data Management

A spreadsheet for audit tracking will be created to provide a pathway through the audit process and aid in the structured collection of data. An example is presented in Table B-1 below. A spreadsheet row will represent a single case and contain the following variables: case ID, MDC, NLP assigned DRG, manually assigned DRG, principal and impactful secondary

diagnoses, procedure[s] submitted, missed impactful diagnosis or procedure code[s], DRG assignment rationale, AI error type (if applicable), and adjudication outcome.

Manual DRG Validation Procedure

All cases will be independently reviewed by a credentialed coding auditor blinded to the software's result. During validation, the typical coding references were referenced including the ICD-10-CM Official Guidelines for Coding and Reporting, CMS MS-DRG Definitions Manual, and AHA Coding Clinics. The physical auditing will involve the following activities:

- Validating the correct principal diagnosis and impactful secondary diagnoses.
- Reviewing procedures that affected DRG assignment.
- Confirming proper sequencing of diagnoses and procedures.
- Assigning a DRG based on CMS grouping logic.

Before reviewing the software-assigned DRG for comparison, the auditor will record the DRG assignment and supporting reason for the assignment.

Discrepancy Identification and Error Classification

After the DRG assignments are recorded, the results will be compared to identify concordant and discordant cases. Discrepant cases will be further categorized using a standardized error taxonomy:

- Omitted Codes: Diagnoses or procedures documented in the chart but not captured by the NLP system.
- Misclassification: Incorrect selection of the principal diagnosis or principal procedure.

- Incorrect Sequencing: Improper ordering of diagnoses or procedures that impacted DRG assignment.
- Overcoding: Assignment of unsupported diagnoses or procedures by the software.
- Undercoding: Omission of clinically supported diagnoses or procedures.
- No Finding: Absence of DRG assignment by the AI when one was warranted.

Each case will be adjudicated into one of four final categories: AI correct, auditor correct, both incorrect, or unable to determine (owing to ambiguous or inadequate clinical documentation).

Inter-Rater Reliability and Quality Control

For the purpose of enhancing the consistency and reducing the subjectivity of audits, a secondary audit was performed on ten percent of the sample (n = 40). The latter two were each reviewed for accuracy by a second credentialed auditor. All discrepancies were adjudicated by consensus agreement by a panel consisting of two expert coding practitioners and a clinical documentation integrity (CDI) specialist. Weekly calibration meetings took place to harmonize data interpretation and questions, and to record changes to the audit tool.

Table B-1.

Sample Layout of Audit Tracker

Case ID	MDC	AI DRG	Manual DRG	Final DRG	Principal Dx	Procedures	Discrepancy (Y/N)	Dollar Variance	Error Type	Final Adjudication	Auditor Notes	Rationale	Root Cause
123456	05	291	292	292	J18.9	0W993ZZ	Y	$350	Omitted Code	Auditor Correct	AI missed secondary Dx	Secondary pneumonia diagnosis not captured by AI	WHY?
123457	08	308	308	308	K35.80	0DTJ0ZZ	N	$0		AI Correct	No discrepancies noted	DRG aligned across AI and manual review	

Appendix C

Interview Guide

Introduction [interviewer]:

Thank you for agreeing to participate in this interview. As a reminder, this project explores how AI-enabled DRG validation tools perform in revenue cycle settings, the types of errors that occur, and how the errors impact financial and operational outcomes. Your insights will help to better understand how these systems are used and how they could be improved. There are no right or wrong answers. I'm interested in hearing your experiences and perspectives.

SECTION A: Background and Role Context

Q1. Can you describe your current role in the revenue cycle and how you interact with DRG validation systems?

[Follow-up prompts:]

- How long have you worked in this area?
- What types of cases or services are you most involved with?

SECTION B: Nature and Types of DRG Validation Errors (PQ1)

Hint: Aligned the conversation to HAF constructs: Explainability, Role Allocation, Error Recovery

Q2. In your experience, what are the most common types of errors you've observed in AI-enabled DRG validation?

[Follow-up prompts:]

- Can you give an example of a recent or memorable error?

- Were these errors related to the original documentation, AI interpretation, or something else?

Q3. How are errors usually detected and addressed in your workflow?

[Follow-up prompts]

- What role do human coders, CDI professionals, or auditors play in the process?
- From your experience, how clear are the boundaries between tasks assigned to the AI system and those assigned to human reviewers? Can you provide an example that illustrates what you mean?

SECTION C: Impact of Errors on RCM Performance (PQ2)

Hint: Align conversation to HAF constructs: Trust Calibration, System Performance Feedback

Q4. How do these AI-related DRG errors typically affect revenue cycle outcomes such as denial rates, reimbursement timelines, or appeals?

[Follow-up prompts]

- Are there specific financial or operational metrics affected by these errors?
- How often do errors result in revenue loss or rework?

Q5. How do you and your team respond when AI-derived coding outputs are questioned by payers or internal auditors?

[Follow-up prompts:]

- Is there a formal appeals or correction process?
- Do repeated AI errors affect the team's confidence in the system?

SECTION D: Contributing Factors to DRG Validation Errors (PQ3)

Hint: Align conversation to HAF constructs: Co-Adaptive Learning, Data Governance, Explainability

Q6. In your opinion, what contributes most to DRG validation errors in AI-enabled systems?

[Follow-up prompt]

- Are there issues with documentation quality, system training, or software updates?
- How well is the system trained for your patient population or service lines?

Q7. Are there any system-level or organizational barriers that make it difficult to improve AI accuracy?

[Follow-up prompts]

- Is the data feeding into the system accurate and well-aligned with coding needs?
- Is there feedback provided from end users back to the AI vendor or development team?

SECTION E: Trust, Explainability, and System Improvement

Hint: Align discussion to HAF constructs: Trust Calibration, Explainability, Co-Adaptation

Q8. How confident are you in the decisions or suggestions provided by the AI system?

Follow-up prompts:

- Are there situations where you trust it more or less?
- How does your trust change over time or across cases?

Q9. Does the system provide clear reasoning or explanations for its DRG recommendations? If not, how does that affect your work?

Follow-up prompts:

- Do you feel equipped to challenge or override AI outputs?
- What would help make the system more transparent?

Q10. Are there opportunities for the system to learn from user feedback or corrected outputs?

Follow-up prompts:

- How is feedback captured or shared?
- Have you seen any improvements over time based on system use?

SECTION F: Final Reflections

Q11. Based on your experience, what changes would you recommend to improve the accuracy, trust, or usability of AI in DRG validation workflows?

Closing [Interviewer]

Thank you for your time and insights. Your perspectives will help us better understand the challenges and opportunities of using AI in revenue cycle management. If you have any additional thoughts you'd like to share or if you'd like a summary of the findings when the study is complete, please let me know.

Appendix D

Recruitment Email

My name is Tiffany Reeves and I am a doctoral student at National University. I am conducting a study to gain better understanding on NLP-DRG validation errors and the impact of these errors on revenue cycle performance. The name of this project is "AI-Driven DRG Validation in Healthcare RCM: Challenges, Solutions, and the Path Forward"

I am recruiting individuals to participate in this project for a voluntary interview. Participants should meet the following criteria:

1. You have at least 3 years experience in one of the following hospital inpatient revenue cycle roles: inpatient coding, clinical documentation improvement, medical auditing, and revenue cycle program management.

2. You have direct experience in NLP-enabled DRG validation tools in the context of hospital inpatient revenue cycle management.

If you decide to participate in this project you will be asked to participate in a 1:1 online interview over Zoom for approximately 1 hour. During the interview, you will be asked questions about your experiences with NLP-enabled DRG validation tools and their impact on revenue cycle workflow and performance. Participation is voluntary and you can decide not to participate at any time.

If you are interested in participating in this project, please contact me at T.Reeves3774@o365.ncu.edu. Thank you for considering participating in this voluntary project!

Tiffany Reeves

Appendix E

Informed Consent

National University

National University IRB
9338 Lightwave Ave., San Diego, CA 92123
irb@nu.edu

Consent Form

My name is Tiffany Reeves, and I am a doctoral student at National University. I am conducting a study to gain better understanding of NLP-DRG validation errors and the impact of these errors on revenue cycle performance. The name of this project is "AI-Driven DRG Validation in Healthcare RCM: Challenges, Solutions, and the Path Forward."

You may participate in this study if you meet all the following criteria:

1. You have at least 3 years' experience in one of the following hospital inpatient revenue cycle roles: inpatient coding, clinical documentation improvement, medical auditing, and revenue cycle program management.
2. You have direct experience in Natural Language Processing (NLP)-enabled DRG validation tools in the context of hospital inpatient revenue cycle management.

I hope to include 10-15 people in this research project.

Please read this form carefully and ask any questions you may have before agreeing to take part in the study.

What you will be asked to do: If you agree to be in this study, you will be asked to do the following activities:

1. Participate in a 1:1 online interview over Zoom for approximately 1 hour.
2. Review a transcript of the interview to ensure accuracy.

During these activities, you will be asked questions about your experiences with NLP-enabled DRG validation tools and their impact on revenue cycle workflow and performance.

Risks: There are minimal foreseeable risks or discomforts associated with this research. You can skip any question you do not wish to answer, skip any activity, or stop participating at any time.

Benefits: If you participate in this study there are no direct benefits to you. However, you may learn more about AI usage in revenue cycle management processes. This study may increase the body of knowledge on designing more effective human-AI interaction to improve revenue cycle workflow and performance.

Recording: I would like to audio/video record your responses with Zoom during the interview. You can disable the video function of the online meeting platform at any time.

Confidentiality: I will keep the records of this interview and take reasonable measures to protect the security of all your personal information. In any report I make public, I will not include any information that will make it possible to identify you. All information will be stored on local computer/laptop and all data will be destroyed upon completion of this project.

Taking part is voluntary: Participation in this study is completely voluntary. You may quit at any time.

If you have questions: Please ask any questions you have now. If you have questions later, you may contact me at T.Reeves3774@o365.ncu.edu.

If you have any questions or concerns regarding your rights as a subject in this study, you may contact the Institutional Review Board (IRB) at National University via email at irb@nu.edu

Appendix F

Initial Deductive Coding for the Human-AI Integration Framework

Code Name	Definition	Associated PQ(s)	Illustrative Examples
Role and Task Allocation	Observations related to how tasks are divided between humans and AI systems in DRG validation workflows.	PQ1, PQ3	Descriptions of which elements are handled automatically by the NLP tool vs. reviewed manually; discussions about mismatch in task assignments.
Trust Calibration	Comments reflecting how users assess, question, or accept AI output based on experience or perceived reliability.	PQ1, PQ2	User expresses overreliance on AI-generated codes or reluctance to trust outputs without double-checking.
Transparency and Explainability	Statements about the system's ability to justify or explain its outputs and decisions.	PQ1, PQ3	Difficulty interpreting why the AI selected a particular DRG; absence of traceable logic in the AI recommendation.
Feedback Loops and Error Recovery	Accounts of how errors are identified, corrected, and used to improve either human performance or system learning.	PQ1, PQ2, PQ3	Lack of mechanisms for correcting errors within the system; user describes post-audit corrections without system learning.
Co-Adaptive Learning	Descriptions of how the system and users adjust their actions over time based on mutual feedback and iterative use.	PQ3	Coders discuss learning how to adapt to system quirks; system updates based on recurring patterns.
Data Quality and Governance	References to challenges or benefits related to data input accuracy, structure, interoperability, or alignment.	PQ3	Discussion of system errors due to outdated or inconsistent source documentation.
Workflow Integration	Insights about how well the AI system integrates into clinical and administrative workflows.	PQ2, PQ3	Frustrations about duplicate work, documentation burden, or delayed processing due to system design.

Code Name	Definition	Associated PQ(s)	Illustrative Examples
Impact on Performance Metrics	Statements linking AI system errors or successes to claim denials, revenue delays, or other quantifiable outcomes.	PQ2	Descriptions of how coding mismatches lead to denied claims or delayed reimbursement.
User Experience and Usability	General sentiments about ease of use, user interface, or emotional responses to the AI tool.	PQ1, PQ2, PQ3	Perceptions of the system being user-friendly or difficult; expressions of stress, trust, or confidence.

Appendix G

Synthetic Case Corpus Summary by Major Diagnostic Category and Coding Complexity

This appendix presents the synthetic inpatient case corpus developed for the DRG validation audit. The cases were designed to represent clinical complexity and documentation variability in this work and were chosen to simulate many coding scenarios commonly encountered in algorithms commonly employed for machine-defined verification of DRG validation results. This corpus comprises typical inpatient scenarios and high-risk coding situations in high-risk areas for automated verification.

Table C1. Distribution of Synthetic Cases by Clinical Category

Clinical Category	Number of Cases	Percentage of Corpus	Coding Complexity Focus
Respiratory System	10	25%	Respiratory failure, pneumonia, ventilatory support
Infectious/Sepsis	8	20%	Sepsis clinical validation, organ dysfunction
Circulatory/Cardiac	6	15%	Chest pain vs cardiac diagnosis selection
Female Reproductive	4	10%	Procedure linkage, complication coding

Digestive System	4	10%	Comorbidity capture, bleeding disorders
Neurologic	3	7.5%	Functional deficits, stroke documentation
Endocrine/Metabolic	3	7.5%	Combination coding, metabolic complications
Multisystem/Complex Chronic	2	5%	Multiple comorbidities and severity capture
Total	40	100%	

Table C2. Representative Synthetic Case Scenarios and Coding Risk

Case Category	Example Clinical Scenario	DRG Type	CC/MCC Level	Primary Coding Risk	AI Validation Challenge
Respiratory	Acute respiratory failure with pneumonia	Medical	MCC	CC/MCC capture	Clinical indicator interpretation
Respiratory	COPD exacerbation with hypercapnia	Medical	CC	Principal diagnosis selection	Documentation nuance
Sepsis	Sepsis vs SIRS with	Medical	MCC	Clinical validation	Sepsis criteria interpretation

	organ dysfunction				
Sepsis	UTI with possible sepsis	Medical	CC	Query opportunity	Condition relationship
Cardiac	Chest pain vs NSTEMI rule out	Medical	No CC	Principal diagnosis	Diagnostic ambiguity
Female Reproductive	Postoperative hemorrhage	Surgical	MCC	Complication coding	Procedure linkage
Digestive	GI bleed with anemia	Medical	CC	Comorbidity capture	Severity interpretation
Neurologic	Stroke with dysphagia	Medical	CC	Functional deficit capture	Secondary diagnosis recognition
Endocrine	Diabetic ketoacidosis with infection	Medical	MCC	Combination coding	Multi-condition logic
Multisystem	CHF with AKI and malnutrition	Medical	MCC	Multiple comorbidity capture	Hierarchical interpretation