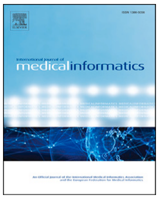
International Journal of Medical Informatics 195 (2025) 105800

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Large language models vs human for classifying clinical documents

Akram Mustafa [a], Usman Naseem [b], Mostafa Rahimi Azghadi [a,*]

[a] College of Science and Engineering, James Cook University, Townsville, 4811, QLD, Australia
[b] School of Computing, Macquarie University, Sydney, 2113, NSW, Australia

ARTICLE INFO

Keywords:
Machine learning
Clinical document improvement
Large language model
ChatGPT
Snomed
Clinical coding

ABSTRACT

Background: Accurate classification of medical records is crucial for clinical documentation, particularly when using the 10th revision of the International Classification of Diseases (ICD-10) coding system. The use of machine learning algorithms and Systematized Nomenclature of Medicine (SNOMED) mapping has shown promise in performing these classifications. However, challenges remain, particularly in reducing false negatives, where certain diagnoses are not correctly identified by either approach.
Objective: This study explores the potential of leveraging advanced large language models to improve the accuracy of ICD-10 classifications in challenging cases of medical records where machine learning and SNOMED mapping fail.
Methods: We evaluated the performance of ChatGPT 3.5 and ChatGPT 4 in classifying ICD-10 codes from discharge summaries within selected records of the Medical Information Mart for Intensive Care (MIMIC) IV dataset. These records comprised 802 discharge summaries identified as false negatives by both machine learning and SNOMED mapping methods, showing their challenging case. Each summary was assessed by ChatGPT 3.5 and 4 using a classification prompt, and the results were compared to human coder evaluations. Five human coders, with a combined experience of over 30 years, independently classified a stratified sample of 100 summaries to validate ChatGPT's performance.
Results: ChatGPT 4 demonstrated significantly improved consistency over ChatGPT 3.5, with matching results between runs ranging from 86% to 89%, compared to 57% to 67% for ChatGPT 3.5. The classification accuracy of ChatGPT 4 was variable across different ICD-10 codes. Overall, human coders performed better than ChatGPT. However, ChatGPT matched the median performance of human coders, achieving an accuracy rate of 22%.
Conclusion: This study underscores the potential of integrating advanced language models with clinical coding processes to improve documentation accuracy. ChatGPT 4 demonstrated improved consistency and comparable performance to median human coders, achieving 22% accuracy in challenging cases. Combining ChatGPT with methods like SNOMED mapping could further enhance clinical coding accuracy, particularly for complex scenarios.

1. Introduction

1.1. The problem

Discharge summaries are essential for summarizing patient information but often lack consistency and clarity, posing challenges for clinical coders. Missing or unclear details, handwritten summaries, and rushed documentation compromise accuracy. Coders frequently need to cross-reference other sources and communicate with doctors, which can be time-consuming. Improving clinical documentation would save time, as doctors currently spend more than half their working hours on medical paperwork [1], which could be streamlined for efficiency.

Large Language Models (LLMs) have the potential to revolutionize clinical document improvement by enhancing the accuracy and efficiency of medical documentation. Their advanced natural language processing capabilities allow for the automatic extraction of relevant clinical information, reducing the burden on healthcare professionals and minimizing errors [2]. Additionally, LLMs can aid in real-time decision support, providing clinicians with critical insights and recommendations based on comprehensive data analysis [3,4]. Therefore, integrating

* Corresponding author.
E-mail addresses: akram.mohdmustafa@my.jcu.edu.au (A. Mustafa), usman.naseem@mq.edu.au (U. Naseem), mostafa.rahimiazghadi@jcu.edu.au (M. Rahimi Azghadi).

https://doi.org/10.1016/j.ijmedinf.2025.105800
Received 28 November 2024; Received in revised form 16 January 2025; Accepted 19 January 2025

Available online 21 January 2025

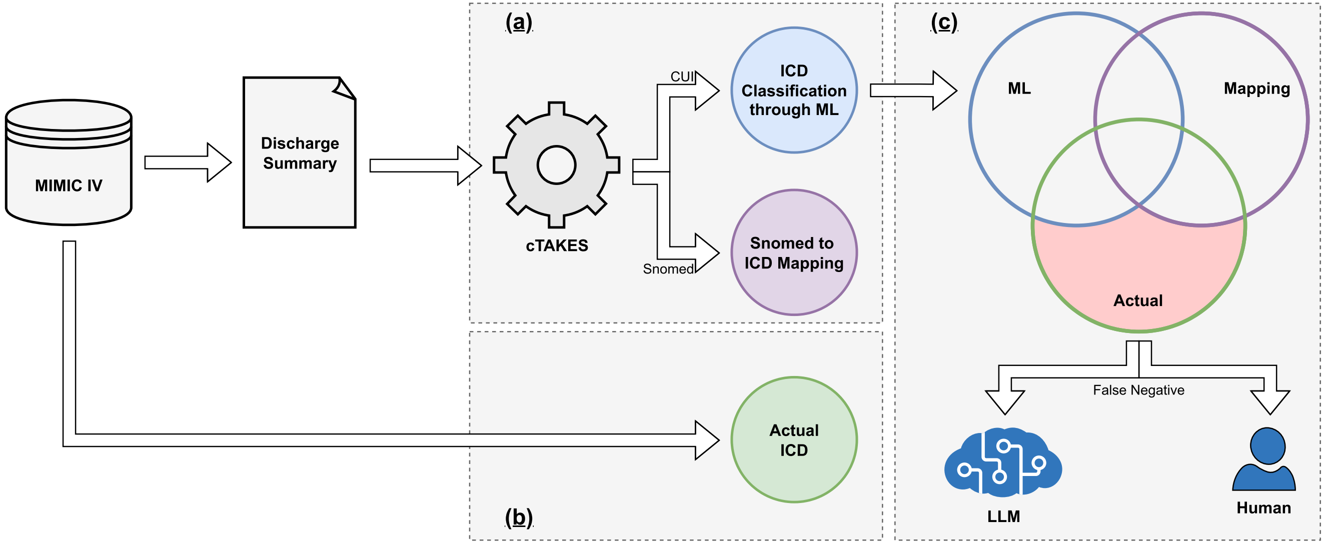


Fig. 1. The overall structure of the proposed methodology.

Large Language Models (LLMs) into Computer-Assisted Coding (CAC) systems offers a promising approach to enhancing Clinical Documentation Improvement (CDI) processes [5].

This research aims to enhance the automation of CDI by assessing the effectiveness of machine learning (ML) and large language models in processing discharge summaries. Through the approach depicted in Fig. 1, discharge summaries are analyzed using clinical Text Analysis and Knowledge Extraction System (cTAKES), to classify each summary into an ICD-10 code using ML, while also using SNOMED Mapping for each summary to an ICD-10 (see Fig. 1(a)). The methodology compares the performance of ML and SNOMED mapping and highlights cases where neither method succeeded based on the actual ICD code available from the MIMIC IV dataset (Fig. 1(b)), subsequently leveraging LLMs to tackle these difficult cases. Results are then contrasted with coding done by experienced human clinical coders (Fig. 1(c)), identifying opportunities for improving coding accuracy.

This study represents an initial step in utilizing technological advancements to optimize healthcare documentation processing. In the future, more research will focus on refining algorithms and improving the performance of LLMs in classifying clinical documents.

1.2. International classification of diseases

The International Classification of Diseases, 10th Revision (ICD-10), is a standardized system developed by the World Health Organization (WHO) for coding diseases, symptoms, and external causes of injury or diseases. ICD-10 is structured hierarchically, with the highest level providing a general description of groups of diseases. For example, the codes A00-B99 represent “Certain infectious and parasitic diseases”. At the most granular level, the code determines great details, e.g. H80.02 specifies “Otosclerosis involving the oval window, nonobliterative, left ear” [6].

ICD-10 codes offer detailed documentation of patient conditions, enabling clinical coders to convert medical reports into a standardized format. This standardization benefits healthcare providers, insurance companies, government agencies, and researchers. Hospitals use ICD-10 codes for consistent patient documentation, aiding accurate billing and reimbursement. Insurance companies rely on these codes for claims processing, and researchers use the data for epidemiological studies and healthcare analysis [7,8].

The ICD-10 serves as the foundation for this research. ICD-10 is utilized to identify primary diagnoses in each discharge summary and functions as the labeling system for the machine learning and LLM prediction process for automating clinical coding (Fig. 1(a)-(b)).

1.3. Automating clinical coding

The automation of clinical coding processes relies on the development and implementation of algorithms, encompassing both rule-based algorithms and machine learning algorithms, which use computers to streamline this task. Many research papers have been dedicated to investigating and building algorithms aimed at helping clinical coders in their duties.

Venkatarama et al. [9] conducted a study employing FasTag methodology, utilizing MIMIC III and Colorado State University (CSU) veterinary data. Their approach entailed the utilization of a Long Short Term Memory Recurrent Neural Network (LSTM-RNN) model in conjunction with MetaMap and Mapping Snomed CT and ICD-9 codes to automate medical note classification for top-level ICD-9 codes. Employing GLOVE for data representation, they achieved a commendable Macro F1 score of 74%.

Similarly, Nguyen et al. [10] explored the integration of SNOMED Clinical Terms (CT) to ICD-10 mapping to enhance Computer Assisted Coding systems. Leveraging data from the Gold Coast Hospital and Health Services (GCHHS), encompassing over half a million patient encounters and 8.5 million medical notes, their method incorporated patient-specific information such as age, gender, and length of stay to enhance classification accuracy for 34 ICD codes, yielding a sensitivity of 54.1%.

Further contributions to the field involve the application of machine learning techniques to classify diagnoses from discharge summaries and medical notes. Gehrmann et al. [11] conducted a comparative analysis between rule-based models and deep learning machine learning techniques, utilizing approximately 1610 discharge summaries from MIMIC III across 10 different diagnoses. Their findings indicated the superiority of the Convolutional Neural Network (CNN) model over other algorithms.

The introduction of automation has significantly eased the workload for clinical coders by providing prompt suggestions and analytics derived from discharge summaries. By simply inputting plain text into the system, clinical coders can obtain comprehensive details pertinent to patient diagnoses and treatments. Notable applications such as the 3M encoder [12] utilize rule-based algorithms, guiding clinical coders through a series of inquiries until a diagnosis is reached. Conversely, systems like DeepMed’s Code Doctor [5] leverage machine learning algorithms to analyze discharge summaries and propose diagnoses, empowering clinical coders to select appropriate options aligned with patient conditions.

1.4. SNOMED

SNOMED is a globally utilized healthcare terminology characterized by its leading role in clinical encoding, automated indexing based on the Systematized Nomenclature of Medicine, and comprehensive coverage of human and veterinary medicine [13–15]. It is crucial for various medical domains [16] as it offers hierarchical information and cross-references [17], aiding in the conversion of patient records [18]. Its integration enhances data quality [19] and represents clinical concepts effectively [20], yet requires manual mapping for legacy data [21].

SNOMED CT codes are employed in the clinical coding process, during which they are converted into ICD-10 codes. These ICD-10 codes are subsequently utilized for document classification. Here, we compare SNOMED-driven ICD-10 codes with those of machine learning predictions, to find challenging documents for LLM processing (see Fig. 1).

1.5. Clinical text analysis and knowledge extraction system (cTAKES)

One of the tools utilized for extracting information from medical reports and discharge summaries is cTAKES. It is an advanced Natural Language Processing (NLP) tool designed to annotate medical documents. It leverages various medical dictionaries, including SNOMED CT, to ensure comprehensive and accurate extraction of clinical concepts [22]. Moreover, cTAKES enhances the granularity of the extracted information by distinguishing between affirmed and negated terms within the text [23]. This capability is crucial for understanding the context and clinical relevance of the information, thereby improving the overall accuracy and utility of the extracted data in medical informatics.

This research analyses discharge summaries using cTAKES to extract Concept Unique Identifiers (CUIs). These CUIs are employed as features for a machine learning model designed to classify each summary. The model's classifications are subsequently compared with the outcomes of a SNOMED mapping process (Fig. 1(a)).

1.6. Large language models (LLMs)

In recent years, the field of natural language processing has seen significant advancements, particularly with the development of large language models. These models, characterized by their vast size and complexity, are capable of understanding and generating human-like text based on vast datasets they are trained on. LLMs such as OpenAI's GPT-3.5 and GPT-4 and Google Gemini have demonstrated remarkable proficiency in tasks ranging from text generation to complex question answering, summarization, and translation [24,25].

In the healthcare domain, LLMs have shown immense potential in several areas. Firstly, they can enhance clinical decision-making by providing evidence-based recommendations [26]. For example, LLMs can assist physicians by quickly analyzing medical literature and suggesting treatment options based on the latest research findings. Secondly, LLMs can improve patient care through personalized health management. By analyzing patient data, these models can generate individualized health plans and monitor patient progress, thereby enabling more targeted and effective treatments [27].

In this paper, discharge summaries that are not correctly classified by either SNOMED to ICD-10 mapping or using a machine learning classification model, are subjected to a second round of analysis using an LLM. The LLM's assessments are then compared to the expertise of human clinical coders to determine the model's accuracy in dealing with these challenging summaries (Fig. 1(c)).

2. Methods

2.1. Experimental settings

MIMIC IV is an extensive, publicly accessible dataset created and maintained by the Massachusetts Institute of Technology (MIT) [28,29]. This dataset is a critical resource for research in critical care medicine, encompassing 331,794 discharge summaries from 145,915 patients. To safeguard patient privacy, the dataset ensures that all identifiable information, including patient names, dates of birth, admission dates, and addresses, is deidentified [30,31]. MIMIC IV includes patient records coded using two versions of the ICD coding system: ICD-9 and ICD-10.

For this study, we focus exclusively on ICD-10 codes. The dataset reveals that patients were diagnosed with a total of 16,156 different ICD-10 codes at the most granular level, providing a comprehensive overview of patient diagnoses [30,31]. To streamline the analysis, we have chosen to utilize ICD-10 Level 3 [32] codes in this research. ICD-10 Level 3 codes offer a more aggregated classification while still maintaining a significant level of diagnostic detail. Examples of these ICD-10 Level 3 codes are Mycetoma which can be coded as B47, and H80 for Otosclerosis. This selection results in a total of 1,648 distinct ICD-10 Level 3 codes. This approach enables a more coherent and manageable analysis. Furthermore, to maintain focus and address issues related to the infrequency of certain diagnoses, this study is limited to the top five Level 3 diseases based on their prevalence.

Analyzing all available discharge ICD-10 summaries, Disorders of lipoprotein metabolism and other lipidemias (ICD-10: E78) appeared in 49,310 cases, with only 9 cases identified as the primary diagnosis. Similarly, Long term drug therapy (ICD-10: Z79) was documented in 40,393 cases but never as the primary diagnosis. Table 1 shows the top 5 case counts and their respective primary diagnosis cases. In contrast, Sepsis (ICD-10: A41) was recorded in 7,430 cases overall but was the primary diagnosis in 4,830 of those cases. Similarly, Myocardial Infarction (ICD-10: I21) was identified in 5,735 cases in total, with 2,722 of these being primary diagnoses. Table 2 illustrates the most frequently occurring diagnoses and their top primary diagnoses respectively. In this study, we focus on the documents summarized in Table 2.

Focusing on these most frequently occurring ICD-10 primary codes ensures that this research addresses conditions with the highest impact on patient care. Additionally, narrowing the scope to the top five codes enables a more thorough and detailed analysis. This focused approach also ensures that the research is conducted with greater accuracy and depth. Moreover, handling a smaller set of codes makes the data more manageable and reduces the complexity of the analysis, allowing for clearer and more concise conclusions.

2.2. Data preparation using cTAKES and SNOMED

The discharge summaries of the cohort of 14,338 primary diagnosis cases, as shown in Table 2, were used. An initial analysis of the text data revealed an average word count of 1,984 words per report. Upon processing these summaries through cTAKES, the verbose clinical language was distilled into a more structured format, both in CUIs and SNOMED codes (see Fig. 1(a)). Furthermore, cTAKES identifies both negated and affirmed terms within the text. This feature is crucial for accurate clinical interpretation. This dual coding capability ensures that the clinical context of the data is preserved, preventing potential misinterpretations that could arise from the mere presence of a term without its contextual qualifiers.

2.3. Diagnoses classification through machine learning

As shown in Fig. 1(a), the CUI data extracted from the cTAKES process serves as the feature set for the machine learning process. To prepare the cTAKES-generated CUIs for machine learning, the Bag of Words (BOW) feature extraction method was used. In this method, negated terms can be handled in two ways. The first approach treats a negated term as a negative count of the term; for instance, “No signs of C-difficile” would assign a value of -1 to “C0343386,” while an affirmed occurrence of C-difficile counts as +1. However, this method has a drawback: if the term appears in both positive and negative forms, they cancel each other out, resulting in a count of zero as if the term never existed

Table 1
Top 5 case count and their respective primary diagnosis in ICD-10.

Diagnosis	ICD-10 Code	Total Cases	Primary Diagnosis Cases
Disorders of lipoprotein metabolism and other lipidemias	E78	49,310	9
Essential hypertension	I10	43,574	76
Long term drug therapy	Z79	40,393	0
Personal history of other diseases and conditions	Z87	40,255	0
Place of occurrence of the external cause	Y92	35,297	0

Table 2
Top 5 primary diagnoses cases and their respective total cases in ICD-10.

Diagnosis	ICD-10 Code	Total Cases	Primary Diagnosis Cases
Sepsis	A41	7,430	4,830
Myocardial infarction	I21	5,735	2,722
Other medical care	Z51	6,919	2,370
Chronic ischaemic heart disease	I25	38,157	2,302
Hypertensive heart and renal disease	I13	8,366	2,114

in the discharge summary. Therefore, we opted for a second approach, which treats negated and affirmed terms as distinct features. In this approach, the negated “C0343386” becomes a unique feature, labeled as “!C0343386”, while the affirmed “C0343386” remains unchanged. BOW then recognizes them as two distinct terms.

We utilized multiple machine learning algorithms, including Decision Tree, Gaussian Naive Bayes (Gaussian NB), K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest, to determine which provided the optimal classification performance. The F1 score, which balances precision and recall to provide a single metric of accuracy [33], was used to measure performance across these algorithms.

The experimental dataset consisted of discharge summaries with primary diagnosis cases, as shown in Table 2. The dataset was split into five smaller datasets, one for each ICD-10 code. For example, one dataset contained only records related to A41, another for I21, and so on. To ensure balanced datasets, an equal number of randomly selected records from the full MIMIC IV discharge summary dataset, excluding those listed in Table 2, were incorporated into each of the five target ICD-10 codes in our dataset. This method produced balanced datasets comprising both positive cases (patients diagnosed with the target ICD-10 code) and negative cases (patients without that diagnosis).

Next, each dataset was transformed by assigning a binary label (1 or 0) to each record, where 1 indicated a discharge summary classified as diagnosed with the specific ICD-10 code, and 0 indicated it was not diagnosed. We utilized a 10-fold cross-validation for each model, running the algorithm ten times. This method ensured robust validation and minimized overfitting.

Fig. 2 illustrates the F1 score results for each algorithm across the five different ICD-10 classes. Random Forest emerged as the best-performing algorithm, with an F1 score of 85.83% for class A41 and reaching more than 93% for classes I13, I21, I25, and Z51. Therefore, we employed the Random Forest algorithm as the main machine learning method in our methodology shown in Fig. 1 to perform binary classification on discharge summaries. The primary objective was to determine whether each discharge summary could be classified according to specific ICD-10 codes.

Table 3 summarizes the number of records, along with the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) counts, as well as the F1 score for each dataset.

2.4. SNOMED mapping

As shown in Fig. 1(a), we used SNOMED codes extracted via cTAKES and converted them into ICD-10 codes using the SNOMED International mapping tool [34]. This process leveraged a mapping file containing 238,278 SNOMED to ICD-10 code pairs. To align with our research, we simplified the detailed ICD-10 codes (e.g., A41.9) to Level 3 codes (e.g., A41). From the complete mappings, only 717 SNOMED codes matched the ICD-10 codes relevant to our study: A41, I13, I21, I25, and Z51.

Table 3
Classification Performance Metrics for ICD-10 Code Datasets.

ICD-10 Code	Balanced Total	TP	TN	FP	FN	F1 Score
A41	9,660	4,172	4,110	720	658	85.83%
I13	4,228	1,981	1,964	150	133	93.33%
I21	5,444	2,563	2,579	143	159	94.44%
I25	4,604	2,131	2,255	47	171	95.13%
Z51	4,740	2,154	2,279	91	216	93.35%

Table 4
Detailed Performance Metrics of SNOMED to ICD-10 Mapping Process.

ICD-10 Code	Balanced Total	TP	TN	FP	FN	F1 Score
A41	9,660	3,797	4,475	355	1,033	84.55%
I13	4,228	277	2,106	8	1,837	23.09%
I21	5,444	1,875	2,622	100	847	79.84%
I25	4,604	2,117	1,970	332	185	89.12%
Z51	4,740	6	2,353	17	2,364	0.50%

Similar to the machine learning process described in Section 2.3, the accuracy of the mapping process was assessed using the F1 score. The mapping accuracy varied significantly across different ICD-10 codes. Specifically, the SNOMED to ICD-10 mapping demonstrated high accuracy for the ICD-10 code I25, achieving an F1 score of 89.12%. Conversely, the mapping accuracy was notably lower for other ICD-10 codes. The mapping for Z51 resulted in a particularly low F1 score of 0.50%, indicating significant challenges and substantial inaccuracies in accurately converting codes in this category. Table 4 provides a detailed breakdown of the mapping process, illustrating the performance metrics for each of the five ICD-10 codes under consideration.

To establish baselines for both the machine learning classification and SNOMED mapping processes, we evaluated basic methods for comparison. For machine learning, a baseline model was implemented using Logistic Regression, a method commonly used in related research as a benchmark [35,36], achieving an average F1 score of 91.32%. This baseline underscores the performance improvements achieved by more advanced models, particularly Random Forest, which consistently outperformed Logistic Regression. For SNOMED mapping, the baseline involved a direct one to one mapping using the SNOMED International

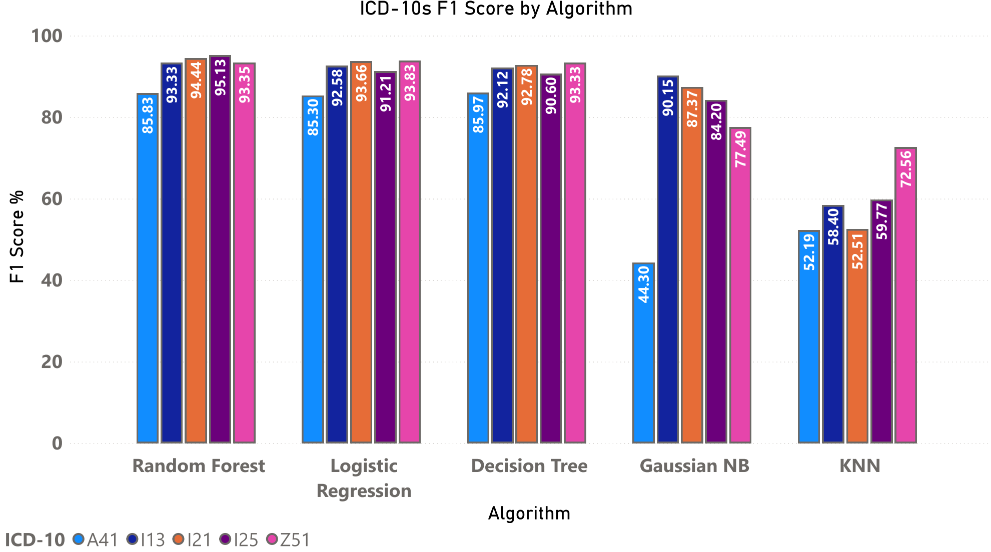


Fig. 2. Performance of various machine learning algorithms in diagnosing various records.

mapping tool without any additional validation or refinement. This approach served as a control for evaluating mapping accuracy. While this baseline performed well for some codes like I25, the low accuracy observed for codes like Z51 indicates the need for additional refinement or supplementary methodologies to address challenges in mapping certain ICD-10 categories. These baselines provide critical context for understanding the effectiveness of the proposed methodologies and underscore the improvements demonstrated by the results in Fig. 2 and Tables 3 and 4.

3. Results

3.1. SNOMED mapping vs machine learning classification

With the implementation of two methods, Machine Learning and SNOMED Mapping, we can further analyze performance across the two approaches (see Fig. 1(c)). An effective way to illustrate this analysis is through Venn diagrams. Fig. 3 presents these diagrams for the five different codes used in our experiments. For example, in the A41 diagnoses, 3,397 discharge summaries were successfully identified as positives by both methods. However, 258 A41-positive discharge summaries were not correctly identified by either method. The analysis reveals difficult cases where both ML and mapping methods fail to classify ICD-10 codes correctly, as illustrated in Fig. 1(b)-(c).

Fig. 3 illustrates that a total of 802 discharge summaries were not correctly classified as positive ICD-10 codes by either method. Specifically, the false negatives include 258 for A41, 126 for I13, 107 for I21, 98 for I25, and 213 for Z51. Addressing these false negatives is crucial for improving the overall reliability of clinical documentation and ensuring that ICD-10 classifications are as accurate as possible.

3.2. Using LLMs for challenging code classification

We investigated the use of LLMs in classifying the identified challenging cases. This measures their potential to streamline clinical document classification when used in conjunction with mapping and machine learning methods to address their limitations in handling complex cases. To that end, a stratified random sample of 100 discharge summaries from the aforementioned 802 false negative cases was selected, ensuring that the ratio of the five different ICD-10 codes was maintained. This sample included 32 summaries for A41 from the overall 160, 16 for I13, 12 for I21, 12 for I25, and 28 for Z51.

Each of these 100 discharge summaries was then submitted to ChatGPT 3.5 and 4 with the following prompt:

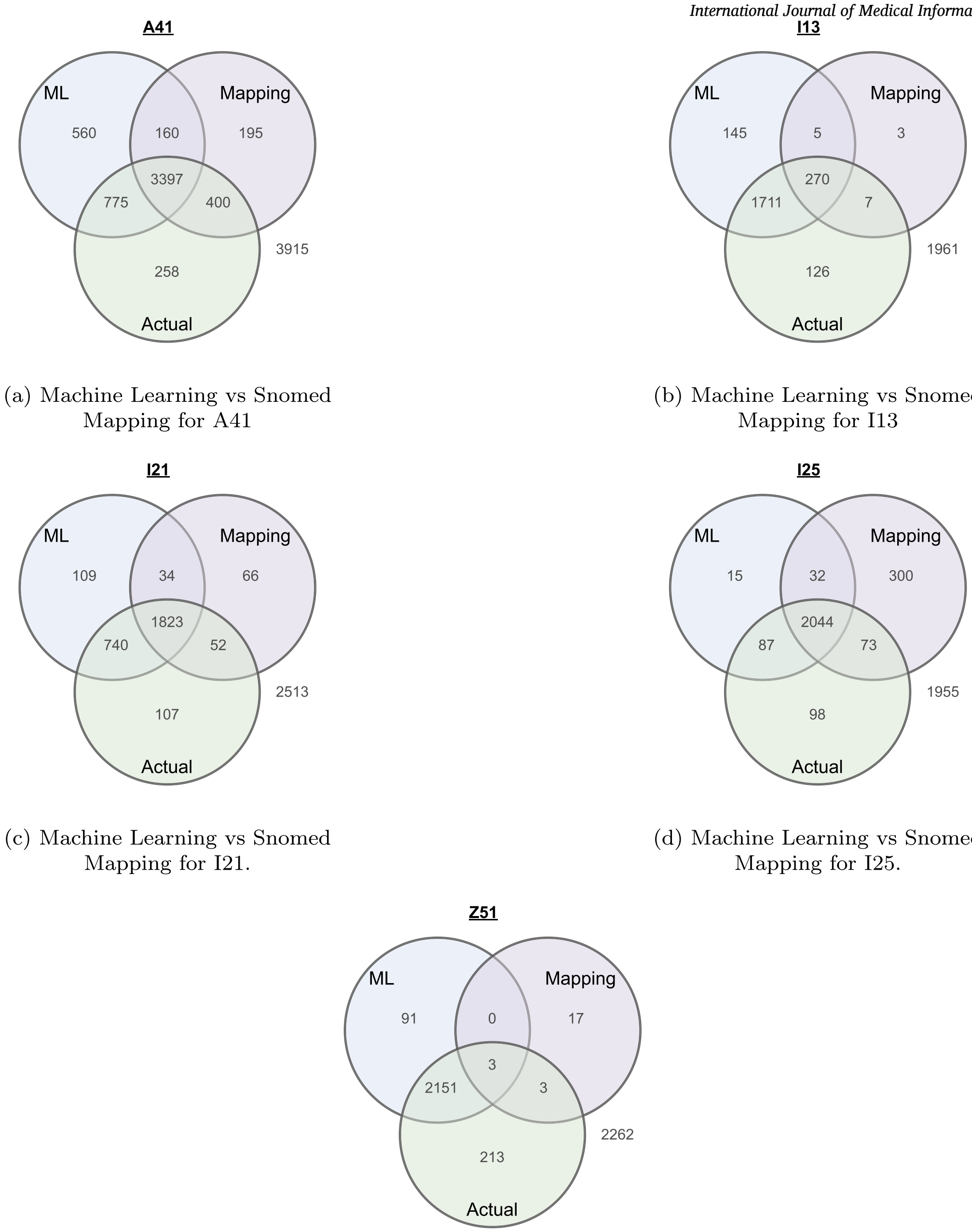
“In the role of a clinical coder, would you categorize this discharge summary as [ICD-10 Code] or not? Answer with yes or no, then explain your answer. [Discharge Summary Text]”

To ensure an unbiased comparison between the performance of ChatGPT and human coders, we utilized the exact same instruction/prompt text for both human and ChatGPT. The prompt was designed as an open-text classification task, maintaining consistency in its structure and content across both evaluation methods. The medical reports were presented in their original form, without any modifications, preprocessing, or simplification of the text. This approach ensured that the classification task reflected real-world complexities in handling unprocessed discharge summaries.

The prompt directly posed a classification question to ChatGPT, requiring it to determine whether the report text corresponded to a specific ICD-10 code. It was intentionally crafted without providing explicit instructions on how to perform ICD-10 coding, thereby simulating a zero-shot learning scenario. This decision was made to evaluate ChatGPT’s inherent ability to classify medical text based solely on its pre-trained knowledge and understanding of the task. No additional iterative refinements or fine tuning were made to the prompt during the evaluation process to ensure consistency and replicability of results. This design choice highlights the robustness of the methodology and provides a clear benchmark for comparing ChatGPT capabilities with human expertise in clinical text classification.

The anticipated outcome for utilizing ChatGPT was to obtain straightforward “yes” or “no” responses, each accompanied by an explanatory rationale. These responses were intended to either corroborate or refute the false negative classifications provided by the machine learning algorithm and SNOMED mapping. The rationale was crucial for understanding the decision-making process and identifying any potential discrepancies between the methods.

To address the trust issues with ChatGPT 3.5 due to its randomness, we assessed the consistency of its classifications. Specifically, we counted the reports where all runs of ChatGPT 3.5 consistently identified a report as positive for the ICD-10 code A41. For a report to be considered positively classified, it must be identified as A41 in all three runs of ChatGPT 3.5; if even one run fails to classify it as A41, the result is deemed negative. Table 5 compares the number of reports that all runs of ChatGPT 3.5 agreed to classify as positive against those all runs of ChatGPT 4 classified as positive. The data clearly demonstrates that multiple runs of ChatGPT, even across different versions such as 3.5 and 4, can yield more consistent results.



(a) Machine Learning vs Snomed Mapping for A41

(b) Machine Learning vs Snomed Mapping for I13

(c) Machine Learning vs Snomed Mapping for I21.

(d) Machine Learning vs Snomed Mapping for I25.

(e) Machine Learning vs Snomed Mapping for Z51.

Fig. 3. Comparison of ML and Snomed mapping in classifying the five target ICD-10 codes benchmarked against the actual ICD-10 code in the MIMIC dataset.

Table 5
Comparison of ChatGPT 3.5 and ChatGPT 4.

ICD-10 Code	Total Reports	ChatGPT 3.5	ChatGPT 4
A41	32	2	2
I13	16	1	1
I21	12	2	3
I25	12	2	2
Z51	28	0	8

Between ChatGPT 3.5 and 4, the latter demonstrated better performance, successfully identifying a subset of discharge summaries across these challenging ICD-10 codes where the two other techniques failed. Fig. 4 shows that ChatGPT 4 correctly classified 2 out of 32 discharge summaries for ICD-10 code A41, achieving an accuracy of 6.25%, and 25% accuracy for codes I13, I21, and I25. For ICD-10 code Z51, ChatGPT 4 accurately classified 10 out of 28 discharge summaries, representing an accuracy of 35.71%. Out of a total of 100 discharge summaries, ChatGPT-4 successfully classified 22, resulting in an accuracy rate of 22%.

In general, ChatGPT 4 demonstrates significant improvements over ChatGPT 3.5 in classifying medical report diagnoses due to advancements in its architecture, training data, and fine tuning. The larger number of parameters and sophisticated optimization techniques in ChatGPT 4 enable it to better capture complex patterns and medical nuances present in diagnostic texts [37]. Additionally, ChatGPT 4 benefits from training on a more extensive and higher quality dataset, which likely includes a broader range of domain specific knowledge, allowing it to generalize more effectively and produce accurate classifications.

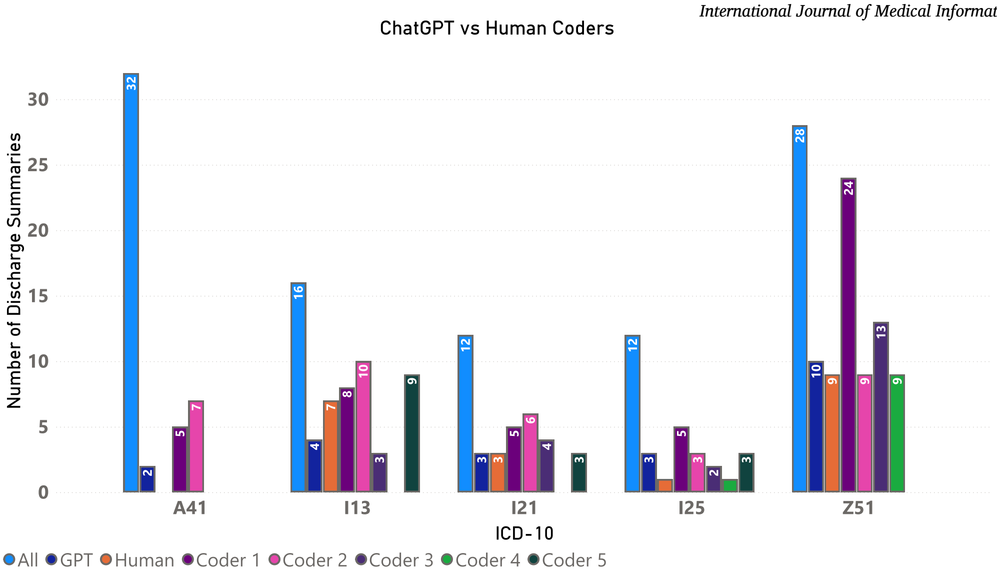
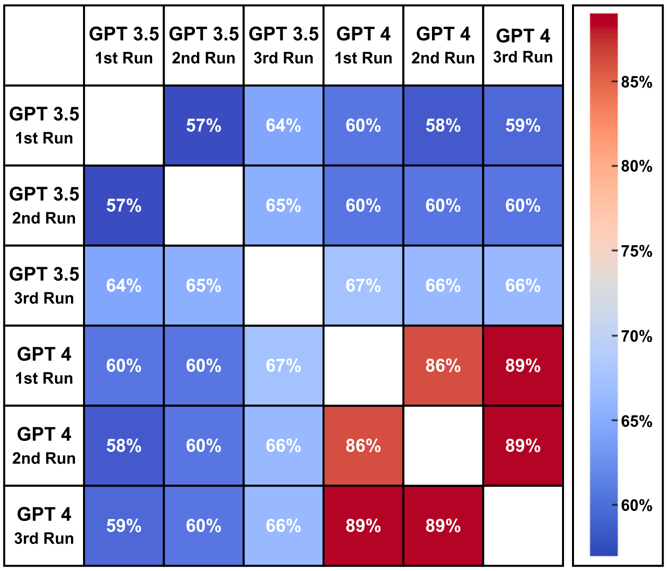


Fig. 4. ChatGPT 4 vs five human coders and a majority vote of human coders, labeled as “Human”.



	GPT 3.5 1st Run	GPT 3.5 2nd Run	GPT 3.5 3rd Run	GPT 4 1st Run	GPT 4 2nd Run	GPT 4 3rd Run
GPT 3.5 1st Run		57%	64%	60%	58%	59%
GPT 3.5 2nd Run	57%		65%	60%	60%	60%
GPT 3.5 3rd Run	64%	65%		67%	66%	66%
GPT 4 1st Run	60%	60%	67%		86%	89%
GPT 4 2nd Run	58%	60%	66%	86%		89%
GPT 4 3rd Run	59%	60%	66%	89%	89%	

Fig. 5. Consistency of Discharge Summary Classifications Across Different Runs of ChatGPT.

3.3. ChatGPT variation analysis

To measure the robustness of ChatGPT, we ran 3 runs of our experiment on all 100 documents using both ChatGPT 3.5 and 4. This resulted in matching responses between 57% and 67% for ChatGPT 3.5. In this context, “matching” means that the LLM classified a record in the same correct or incorrect category across the different runs. The low matching values indicate near-chance classification performance for ChatGPT 3.5.

ChatGPT 4 demonstrated significantly less variability, with matching results ranging from 86% to 89%. This increased consistency highlights the improvements in performance between versions of the language model.

Fig. 5 details the percentage of discharge summaries with results matching across different runs of ChatGPT for all 100 test summaries that were complex to classify by machine learning, SNOMED mapping and even human coding.

3.4. Human validation

To thoroughly understand ChatGPT’s capabilities and the complexities involved in classifying diagnoses based solely on discharge summaries, the 100 randomly selected discharge summaries classified by ChatGPT in the previous step were evaluated by five volunteer clinical coders. The coders were asked to answer the same question posed to ChatGPT.

For consistency, the five clinical coders were provided solely with the discharge summaries, without any additional case information. These independent coders collectively possess over 30 years of clinical coding experience.

As Fig. 4 shows, the results varied among the coders. In addition, this figure demonstrates a majority vote system, labeled as “Human”. A discharge summary was considered correctly identified if three or more out of the five human coders categorized it accurately.

For A41, which was identified as the most challenging for ChatGPT, most coders, three out of five, also failed to correctly identify any of the A41 cases. Furthermore, the majority human vote case completely failed. This could primarily be attributed to the limited information available in the discharge summaries, a lack of access to additional clinical documents, the absence of a CDI process, and the lack of direct communication with the attending physicians.

In the randomly selected set of 100 reports, ChatGPT achieved an accuracy of only 22%, which aligns with the median accuracy observed among human coders. Human Coder 1 demonstrated the highest accuracy at 47%, followed by Coder 2 at 35%, Coder 3 at 22%, Coder 5 at 15%, and Coder 4 with 10%, resulting in a median accuracy of 22% among the human coders. Across all diagnoses, at least two human coders consistently achieved accuracy scores equal to or greater than ChatGPT’s performance. Specifically, for codes A41 and Z51, two clinical coders outperformed ChatGPT in ICD classification accuracy, while three coders did so for codes I13 and I21, and one coder surpassed ChatGPT for code I25. This is all while, in only one case, the human coders’ majority vote, outperformed ChatGPT, showing the complexity of the task and the variability among human coders.

Analysis of the F1 scores reveals an interesting trend: ChatGPT’s F1 scores are closely aligned with the average F1 scores of the five human coders. For instance, for A41, the average F1 score of the five human coders is 12.59%, closely matching ChatGPT’s F1 score of 11.76%. Similarly, for ICD-10 codes I13, I21, and I25, ChatGPT consistently achieves F1 scores of approximately 40%, while the averages for the five human coders are 49.43%, 43.10%, and 36.55%, respectively. For Z51, the av-

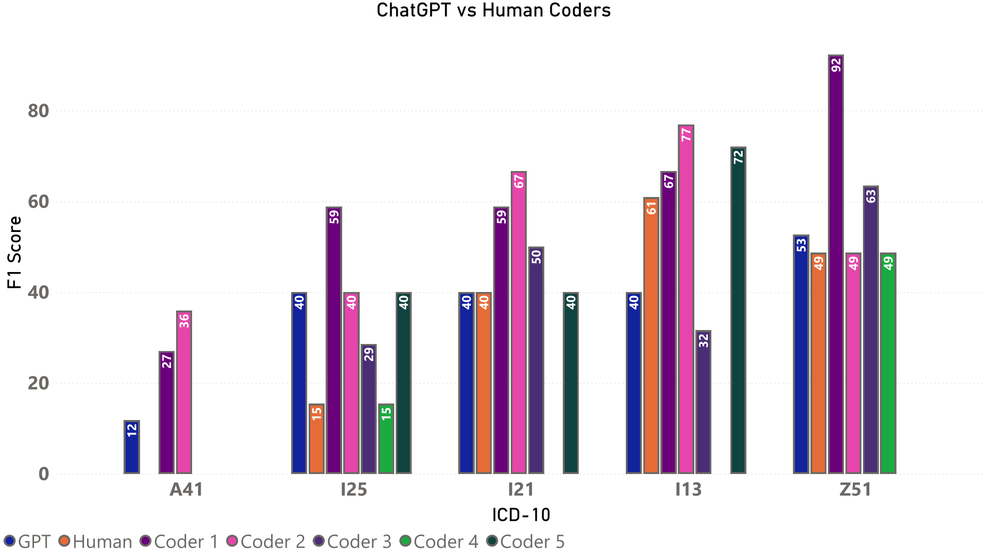


Fig. 6. F1 Score of ChatGPT 4 vs five human coders and a majority vote of human coders, labeled as “Human”.

Table 6
Comparison of Methodologies in Clinical Coding Studies using ML, SNOMED and LLMs.

Study	Methodology	Comparison to Our Approach
Fung et al. [38]	Explored SNOMED mapping as a tool to assist clinical coders, improving coding time and accuracy.	Our approach automates clinical coding with SNOMED mapping integrated with machine learning (ML) models, achieving higher accuracy. We also use the outcome for processing challenging cases using LLMs.
Boyle et al. [39]	Used LLMs for classifying clinical notes via three techniques: code-based, description-based, tree search.	Our research combines LLMs with ML and SNOMED mapping to address challenges that standalone systems face.
Tomo et al. [40]	Compared ChatGPT 3.5 and 4 for diagnosing 37 dental cases described by medical experts.	Our study targets ICD-10 diagnoses from 100 general discharge summaries, using a hybrid methodology integrating LLMs, ML, and SNOMED mapping.
Soroush et al. [41]	Assessed multiple LLMs using a comprehensive dataset from Mount Sinai Health System’s electronic health records.	Our study restricted coders’ access to additional information to match LLM model input, ensuring a fairer comparison and demonstrating real-world applicability.
This study	Integrates LLMs, ML, and SNOMED mapping for automated ICD-10 coding from discharge summaries.	First study to combine these three methods for automated clinical coding, offering a novel approach not yet reported in clinical coding literature.

erage F1 score among the coders is 50.60%, with ChatGPT achieving a comparable F1 score of 52.63%. Fig. 6 illustrates these comparisons between ChatGPT and human coders’ F1 scores.

In general, the accuracy appears to be low due to the fact that the reports classified by ChatGPT and humans represent challenging cases, particularly those identified as False Negatives by both Machine Learning and SNOMED Mapping.

4. Comparison to related work

Advancements in clinical coding have seen significant strides with the application of machine learning, LLMs and structured medical terminologies like SNOMED. However, the methodologies and applications vary across studies (see Table 4), presenting opportunities for innovation through integration. This section positions our work within the broader research landscape, highlighting key distinctions and contributions (Table 6).

The study by Fung et al. [38] demonstrated the utility of SNOMED mapping as a tool to assist clinical coders, highlighting improvements in both the time coders spend on the task and the accuracy they can achieve. It positioned SNOMED as a supporting mechanism within the human-led clinical coding process. In contrast, our research utilizes SNOMED codes as an automated clinical coding technique. Using this approach, we also automatically identify challenging documents to be processed by LLMs. This approach highlights SNOMED’s potential beyond its traditional role, achieving enhanced accuracy through seamless integration with LLMs, which have demonstrated significant promise.

The study by Boyle et al. [39] focused on LLMs as the primary tool for classifying clinical notes using three distinct techniques: code-based, description-based, and tree search. While their work highlighted the effectiveness of LLMs, our approach uniquely combines LLMs with both machine learning and SNOMED mapping, creating a hybrid methodology capable of addressing classification challenges that standalone systems may struggle to identify or resolve.

In another study conducted by Tomo et al. [40], they compared ChatGPT 3.5 and ChatGPT 4 for diagnosing 37 dental cases described by medical experts, focusing on dental-specific scenarios. In contrast, our research targets more complex ICD-10 diagnoses within discharge summaries. Furthermore, our methodology goes beyond the use of LLMs by integrating them with machine learning and SNOMED mapping, creating an innovative framework not previously reported in the clinical coding literature.

Moreover, Soroush et al. [41] assessed various LLMs, including ChatGPT 3.5, ChatGPT 4, Gemini, and Llama, using a dataset provided by the Mount Sinai Health System’s electronic health record. Their research benefited from clinical coders having access to comprehensive patient records, including radiology reports, lab results, and direct communication with physicians. In contrast, our study restricted coders’ access to additional information to match the input available to the LLM models. This approach ensured a fair comparison and benchmarking, while also emphasizing the real-world applicability of our integrated methodology in resource-constrained clinical coding environments.

Finally, our approach of integrating three methods (ML, SNOMED mapping, and LLMs) has not been previously reported. This integration provides a foundation for further exploration and the development of automated computer-assisted coding systems that unify these techniques.

5. Discussion and conclusion

This study underscores the importance of accurately classifying diagnoses from discharge summaries, a task fraught with challenges due to the complexities of medical language and the inherent limitations of traditional classification methods. By integrating advanced techniques such as machine learning algorithms, specifically the Random Forest model, with SNOMED mapping, we demonstrated significant improvements in classification accuracy, particularly for certain codes. The Random Forest algorithm achieved notable F1 scores, highlighting its effectiveness in processing complex medical data.

However, the analysis also revealed substantial gaps in performance, especially in cases where both machine learning and SNOMED mapping failed to classify certain diagnoses correctly. The examination of these false negatives, which accounted for a considerable portion of misclassifications, points to the necessity for ongoing refinement in both the mapping processes and the underlying algorithms.

The exploration of LLMs such as ChatGPT presents a promising avenue for addressing these challenges. While initial results showed that ChatGPT was able to classify some difficult cases, its performance varied across different ICD-10 codes, suggesting that further training and enhancements are required to improve its accuracy and reliability.

To ensure the safe and effective use of AI in clinical coding, it is essential to prioritize transparency, accountability, and fairness while avoiding biases to ensure equitable outcomes across diverse patient groups. Informed consent and strict privacy protections should safeguard sensitive health data, and AI systems must integrate seamlessly into clinical workflows, with clinicians and coders receiving appropriate training [42,43]. Compliance with healthcare regulations is crucial, as is ongoing monitoring and improvement of AI systems [44,45]. However, adopting AI tools like ChatGPT in clinical workflows raises concerns, including inaccurate outputs from incomplete data, lack of transparency, and data security risks, particularly in cloud-based deployments. Regulatory compliance adds further complexity to their integration, underscoring the importance of cautious implementation and continuous oversight [46]. To mitigate these risks rigorous validation and testing must be conducted to ensure accuracy and reliability across diverse scenarios. Enhancing transparency through explainable AI methods can build trust by allowing clinicians to understand the model's reasoning. Robust data security measures, such as encryption and access controls, are essential to protect sensitive patient information, with on-premise deployments preferred for highly confidential data. Additionally, adherence to regulatory standards is critical to ensure compliance and safe integration into clinical environments.

When comparing the performance of LLMs to that of human coders, notable insights emerged. Human coders, with their extensive clinical knowledge and experience, outperformed ChatGPT in several instances, particularly for challenging diagnoses like A41 and Z51. This highlights the ongoing importance of human expertise in clinical coding, especially in complex scenarios where nuanced understanding is crucial. Despite this, LLMs demonstrated potential in identifying cases that traditional methods struggled with, suggesting that an integrated approach could leverage the strengths of both human coders and LLMs.

In a set of 100 randomly selected clinical reports, ChatGPT achieved 22% accuracy, mirroring the median performance of human coders. Human Coder 1 led with 47%, followed by Coder 2 at 35%, while others varied, with at least two coders consistently matching or surpassing ChatGPT's accuracy across all diagnoses. Interestingly, ChatGPT's performance, while lower overall, showed promise in identifying cases where traditional methods had difficulty, suggesting that LLMs could complement human coding, particularly in reducing inconsistencies and improving efficiency.

This analysis demonstrates the complexity and variability inherent in clinical coding tasks, where human expertise is often necessary to navigate nuanced diagnostic information. Nonetheless, integrating LLMs like ChatGPT could offer significant advantages. A hybrid approach, leveraging LLM's speed and ability to flag difficult cases, combined with human oversight for more intricate scenarios, may enhance coding accuracy and streamline the process.

ChatGPT has the potential to play a significant role in clinical coding applications, such as Computer Assisted Coding systems. However, it still faces practical limitations. Consistent errors observed during its application suggest that fine-tuning ChatGPT or utilizing N-shot techniques could enhance its performance in classifying clinical notes.

Future work could combine ChatGPT, ML, and SNOMED mapping. Such a model would integrate ChatGPT's natural language understanding with ML's predictive capabilities and SNOMED's standardized medical terminology, creating a powerful tool for clinical workflows. This approach would enhance the accuracy of medical text interpretation, ensure interoperability through standardized coding, and improve efficiency by automating labor-intensive tasks like diagnosis coding and classification, ultimately supporting better clinical decision-making.

Furthermore, this research paves the way for further advancements in ICD classification and the enhancement of computer assisted coding systems. By demonstrating promising results, it opens up opportunities to extend these methods to other large language models such as Llama, Gemini, and Med-PaLM 2. These next generation models could further improve the accuracy and efficiency of automated medical coding, ultimately streamlining clinical documentation and patient data management.

Overall, our findings support a multifaceted approach to clinical coding that leverages the strengths of machine learning, code mapping, and LLMs to enhance diagnostic accuracy and, consequently, clinical documentation improvement. Future research should prioritize the development of integrated systems that utilize these methodologies, combined with comprehensive training datasets, to improve classification capabilities and minimize misinterpretations in clinical documentation. This could lead to improved consistency, reduced coder and medical professional fatigue, and better overall clinical documentation quality, ultimately benefiting healthcare systems.

CRediT authorship contribution statement

Akram Mustafa: Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Usman Naseem:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Mostafa Rahimi Azghadi:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to extend their deepest gratitude to several individuals who have contributed significantly to the completion of this research. We would like to acknowledge Dr Ameena Afrath, Amal Adbelmawla, Jyothi Chilaka, Fathimathul Habiya Riyas, and Mutahera Begum for their diligent work on the clinical coding of the 100 discharge summaries used in this research. Their meticulous attention and commitment to accuracy were crucial for the success of this project. We are also profoundly grateful to Dr Suhair AlShanteer for her invaluable assistance in forming the questions posed to the Clinical Coders and the LLM.

References

[1] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare, Nat. Med. 25 (1) (2019) 24–29.

[2] M. Karabacak, K. Margetis, Embracing large language models for medical applications: opportunities and challenges, Cureus 15 (5) (2023).

[3] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (1) (2019) 44–56.

[4] S. Kraus, I. Castellanos, D. Toddenroth, H.-U. Prokosch, T. Burkle, Integrating arden-syntax-based clinical decision support with extended presentation formats into a commercial patient data management system, J. Clin. Monit. Comput. 28 (2014) 465–473.

[5] DeepMed, Code doctor: intelligent computer assisted coding system, https://deepmed.ca, 2023. (Accessed 21 October 2024).

[6] M. Möller, D. Sonntag, P. Ernst, Modeling the international classification of diseases (icd-10) in owl, in: International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management, Springer, 2010, pp. 226–240.

[7] N. England, National Clinical Coding Standards icd-10, 5th edition, Nature Medicine, 2023.

[8] I.T. Adeleke, O.O. Ajayi, A.B. Jimoh, A.A. Adebisi, S.A. Omokanye, M.K. Jegede, Current clinical coding practices and implementation of icd-10 in Africa: a survey of nigerian hospitals, Am. J. Health Res. 3 (1–1) (2015) 38–46.

[9] G.R. Venkataraman, A.L. Pineda, O.J. Bear Don't Walk IV, A.M. Zehnder, S. Ayyar, R.L. Page, C.D. Bustamante, M.A. Rivas, Fastag: automatic text classification of unstructured medical narratives, PLoS ONE 15 (6) (2020) e0234647.

[10] A.N. Nguyen, D. Truran, M. Kemp, B. Koopman, D. Conlan, J. O'Dwyer, M. Zhang, S. Karimi, H. Hassanzadeh, M.J. Lawley, et al., Computer-Assisted Diagnostic Coding: Effectiveness of an nlp-Based Approach Using Snomed ct to icd-10 Mappings, AMIA Annual Symposium Proceedings, vol. 2018, American Medical Informatics Association, 2018, p. 807.

[11] S. Gehrmann, F. Dernoncourt, Y. Li, E.T. Carlson, J.T. Wu, J. Welt, J. Foote Jr, E.T. Moseley, D.W. Grant, P.D. Tyler, et al., Comparing rule-based and deep learning models for patient phenotyping, arXiv preprint, arXiv:1703.08705, 2017.

[12] M.H.I. Systems, Medical coding solutions, https://www.3m.com, 2024. (Accessed 21 October 2024).

[13] Y. Wang, M. Halper, H. Min, Y. Perl, Y. Chen, K.A. Spackman, Structural methodologies for auditing snomed, J. Biomed. Inform. 40 (5) (2007) 561–581.

[14] F. Wingert, An indexing system for snomed, Methods Inf. Med. 25 (01) (1986) 22–30.

[15] D.J. Rothwell, Snomed-based knowledge representation, Methods Inf. Med. 34 (01/02) (1995) 209–213.

[16] J.J. Berman, G.W. Moore, Snomed-encoded surgical pathology databases: a tool for epidemiologic investigation, Mod. Pathol. 9 (9) (1996) 944–950.

[17] D.J. Rothwell, R. Cote, Managing information with snomed: understanding the model, in: Proceedings of the AMIA Annual Fall Symposium, American Medical Informatics Association, 1996, p. 80.

[18] Y. Lussier, D. Rothwell, R. Cote, The snomed model: a knowledge source for the controlled terminology of the computerized patient record, Methods Inf. Med. 37 (02) (1998) 161–164.

[19] R. Vuokko, A. Vakkuri, S. Palojoki, Systematized nomenclature of medicine–clinical terminology (snomed ct) clinical use cases in the context of electronic health record systems: systematic literature review, JMIR Med. Inform. 11 (2023) e43750.

[20] R.L. Richesson, J.E. Andrews, J.P. Krischer, Use of snomed ct to represent clinical research data: a semantic characterization of data items on case report forms in vasculitis research, J. Am. Med. Inform. Assoc. 13 (5) (2006) 536–546.

[21] P.M. Nadkarni, J.A. Darer, Migrating existing clinical content from icd-9 to snomed, J. Am. Med. Inform. Assoc. 17 (5) (2010) 602–607.

[22] M.G. Kersloot, F. Lau, A. Abu-Hanna, D.L. Arts, R. Cornet, Automated snomed ct concept and attribute relationship detection through a web-based implementation of ctakes, J. Biomed. Semant. 10 (2019) 1–13.

[23] A. Mustafa, M. Rahimi Azghadi, Automated machine learning for healthcare and clinical notes analysis, Computers 10 (2) (2021) 24.

[24] B. Min, H. Ross, E. Sulem, A.P.B. Veyseh, T.H. Nguyen, O. Sainz, E. Agirre, I. Heintz, D. Roth, Recent advances in natural language processing via large pre-trained language models: a survey, ACM Comput. Surv. 56 (2) (2023) 1–40.

[25] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774, 2023.

[26] Q. Jin, Z. Wang, C.S. Floudas, F. Chen, C. Gong, D. Bracken-Clarke, E. Xue, Y. Yang, J. Sun, Z. Lu, Matching Patients to Clinical Trials with Large Language Models, arXiv 2023.

[27] M. Sallam, Chatgpt Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns, Healthcare, vol. 11, MDPI, 2023, p. 887.

[28] C. Meng, L. Trinh, N. Xu, J. Enouen, Y. Liu, Interpretability and fairness evaluation of deep learning models on mimic-iv dataset, Sci. Rep. 12 (1) (2022) 7166.

[29] M. Gupta, B. Gallamoza, N. Cutrona, P. Dhakal, R. Poulain, R. Beheshti, An extensive data processing pipeline for mimic-iv, in: Machine Learning for Health, PMLR, 2022, pp. 311–325.

[30] A.e.a. Johnson, Mimic-iv-note: deidentified free-text clinical notes (version 2.2), in: PhysioNet, 2023.

[31] A.E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T.J. Pollard, S. Hao, B. Moody, B. Gow, et al., Mimic-iv, a freely accessible electronic health record dataset, Sci. Data 10 (1) (2023) 1.

[32] A. Mustafa, M. Rahimi Azghadi, Clustered automated machine learning (caml) model for clinical coding multi-label classification, Int. J. Mach. Learn. Cybern. (2024) 1–23.

[33] N.R. Kolukula, S. Puli, C. Babi, R.P. Kalapala, G. Ongole, V.M.K. Chinta, Processing of clinical notes for efficient diagnosis with feedback attention–based bilstm, Med. Biol. Eng. Comput. (2024) 1–16.

[34] SNOMED International, SNOMED CT, https://www.snomed.org, 2024. (Accessed 21 October 2024).

[35] S. Dubois, N. Romano, Learning effective embeddings from medical notes, arXiv preprint arXiv:1705.07025, 2017.

[36] J. Huang, C. Osorio, L.W. Sy, An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes, Comput. Methods Programs Biomed. 177 (2019) 141–153.

[37] D. Banik, N. Pati, A. Sharma, Systematic exploration and in-depth analysis of chatgpt architectures progression, Artif. Intell. Rev. 57 (10) (2024) 257.

[38] K.W. Fung, J. Xu, S.T. Rosenbloom, J.R. Campbell, Using snomed ct-encoded problems to improve icd-10-cm coding—a randomized controlled experiment, Int. J. Med. Inform. 126 (2019) 19–25.

[39] J.S. Boyle, A. Kascenas, P. Lok, M. Liakata, A.Q. O'Neil, Automated clinical coding using off-the-shelf large language models, arXiv preprint arXiv:2310.06552, 2023.

[40] S. Tomo, J.R. Lechien, H.S. Bueno, D.F. Cantieri-Debortoli, L.E. Simonato, Accuracy and consistency of chatgpt-3.5 and- 4 in providing differential diagnoses in oral and maxillofacial diseases: a comparative diagnostic performance analysis, Clin. Oral Invest. 28 (10) (2024) 544.

[41] A. Soroush, B.S. Glicksberg, E. Zimlichman, Y. Barash, R. Freeman, A.W. Charney, G.N. Nadkarni, E. Klang, Large language models are poor medical coders—benchmarking of medical code querying, NEJM AI 1 (5) (2024) AIdbp2300040.

[42] T. Lysaght, H.Y. Lim, V. Xafis, K.Y. Ngiam, Ai-assisted decision-making in healthcare: the application of an ethics framework for big data in health and research, Asian Bioethics Rev. 11 (2019) 299–314.

[43] A. Youssef, A.A. Nichol, N. Martinez-Martin, D.B. Larson, M. Abramoff, R.M. Wolf, D. Char, Ethical considerations in the design and conduct of clinical trials of artificial intelligence, JAMA Netw. Open 7 (9) (2024) e2432482.

[44] H. Dong, M. Falis, W. Whiteley, B. Alex, J. Matterson, S. Ji, J. Chen, H. Wu, Automated clinical coding: what, why, and where we are?, npj Digit. Med. 5 (1) (2022) 159.

[45] A.J. Thirunavukarasu, K. Elangovan, L. Gutierrez, Y. Li, I. Tan, P.A. Keane, E. Korot, D.S.W. Ting, Democratizing artificial intelligence imaging analysis with automated machine learning: tutorial, J. Med. Internet Res. 25 (2023) e49949.

[46] E. Ullah, A. Parwani, M.M. Baig, R. Singh, Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology–a recent scoping review, Diagn. Pathol. 19 (1) (2024) 43.