JMALS

Journal of Medical and Life Science

https://jmals.journals.ekb.eg/

SPBH

Clinical and Medical Coding: A New Pathway for Automation-An Updated Review

Naif Fahad Almarshadi[1], MOHAMMED MUBARAK AWAD ALHARBI[2], Mohammed Saleh Alharbi [3], Waked Ahmed Alwaked[3], Hatem Mazyad Alsayer[3], Saber Maqbel Alhussain[3], and Oawed Sultain Aeed Almotary[4]

[1]Hospital management specialist, Northern Area Armed Forces Hospital, Saudi Arabia

[2]HEALTH INFORMATICS, Northern Area Armed Forces Hospital, Saudi Arabia

[3]Health Information Technician, Northern Area Armed Forces Hospital, Saudi Arabia

[4]Pharmacy Technician, Northern Area Armed Forces Hospital, Hafar Albaten, Saudi Arabia

Corresponding author Email: Naif.f.Almarshadi@gmail.com

DOI:10.21608/jmals.2024.413890

Abstract:

Background: Clinical coding is a critical process in healthcare, involving the transformation of free-text medical records into structured codes using classification systems like ICD-10. This process ensures consistent and comparable clinical data, supporting healthcare planning, policy-making, and epidemiological research. **Aim:** This review aims to explore the evolution of automated clinical coding, evaluate the performance of state-of-the-art deep learning models, and identify key challenges and future directions for improving automated coding systems. **Methods:** The review synthesizes findings from 113 studies on automated clinical coding, focusing on the transition from rule-based symbolic AI to neural AI, particularly deep learning. It examines the performance of multi-label classification models, the integration of knowledge-based approaches, and the challenges of handling long documents, imbalanced data, and terminology changes. The review also highlights the importance of human-in-the-loop learning and explainability in automated systems. **Results:** Deep learning models, particularly transformer-based architectures like BERT, have achieved Micro-F1 scores of 58-60% on benchmark datasets like MIMIC-III. However, challenges such as handling infrequent codes, processing long documents, and incorporating symbolic reasoning persist. Hybrid approaches combining symbolic and neural AI show promise, as do knowledge-augmented deep learning methods. Studies also emphasize the need for high-quality datasets, explainability, and adaptability to new coding systems like ICD-11.

Conclusion: Automated clinical coding has made significant progress but remains a complex task requiring further research. Future directions include integrating symbolic reasoning, improving explainability, and developing more representative datasets. Collaboration between AI researchers and clinical coding experts is essential to advance the field.

Keywords: Clinical coding, automated coding, ICD-10, knowledge graphs, explainability, hybrid AI.

Received: September 23, 2024. Accepted: December 1, 2024. Published: December 19, 2024

Introduction:

Clinical coding is a crucial procedure that entails applying established categorization systems like ICD-10 (International Categorization of Diseases, Tenth Revision) to transform medical records—which are usually composed of free-text narratives by medical professionals—into organized codes. To classify patient information into the proper diagnosis and procedure codes within the ICD and OPCS (OPCS Classification of Interventions and Procedures) systems, for example, a standardized method is applied in Scotland. In the end, these codes add to the national dataset known as the Scottish Morbidity Records (SMR01), which is an essential tool for healthcare analytics and decision-making [1]. Ensuring that clinical data is consistent and comparable throughout time and across various care units is the main goal of clinical coding. The resulting national databases play a key role in assisting a number of sectors, such as improving the epidemiological understanding of a wide range of medical disorders, planning and policy formation for healthcare, and health improvement activities. As a result, the precision and dependability of these statistics are crucial. Furthermore, clinical codes are mostly utilized for billing in the US, underscoring their importance to healthcare institutions' bottom lines [2]. NHS Digital offers easily navigable slides titled "Clinical coding for non-coders" [3] for individuals looking for a basic introduction to clinical coding in the UK.

Clinical coding is not an easy assignment for human coders. Data abstraction and summarization are part of the intricate process [4]. In particular, a proficient clinical coder must choose the most accurate codes from a comprehensive classification system or ontology after interpreting a wide range of documents pertaining to a patient's episode of treatment. This choice must follow often updated coding requirements and be consistent with the context given in the various papers. For instance, the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM), which has over 68,000 diagnosis codes, is used for coding in the US [5]. Similarly, the UK's main classification system for coding is ICD-10. Textual analysis, summary, and well-defined stages for code classification are all part of the standardized manual coding process that is used to guarantee data consistency. According to the NHS Digital coding standards of 2021, these procedures are sometimes referred to as the four stages of coding: analyze, locate, assign, and verify [6, p.11]. This methodical approach reduces the possibility of making mistakes or variances that could result in poor decision-making. Standardized data collection, analysis, and application are therefore essential. The accuracy and applicability of the coding process are further guaranteed by frequent revisions to coding standards and recommendations, such as those offered by Public Health Scotland [6]. It can take months or even longer to train a skilled clinical coder in the UK's National Health Service (NHS) or elsewhere, highlighting the intricacy and level of experience needed for this position [7].

The goal behind automated clinical coding is to automate the clinical coding process by utilizing artificial intelligence (AI) techniques like machine learning and natural language processing (NLP) [8]. This method is included in the more general category of computer-assisted coding (CAC) [9]. By applying cutting-edge machine learning and natural language processing techniques to intelligently interpret the ever-increasing volume of medical data, artificial intelligence (AI) has emerged as a viable tool for altering healthcare in recent years [10]. One possible AI use that can simplify the administration and management of clinical records in medical research settings and hospitals is automated clinical coding. Research articles on automated clinical coding have significantly increased in the last few years, especially those that use deep learning, which is currently the most popular method in AI. Recent surveys have provided ample evidence of this trend [11-13]. Even with the advancements in automated clinical coding, the problem is still far from being solved. This topic has been the focus of our work for the last two years, and we have had in-depth conversations with clinicians from Scotland and the UK as well as clinical coding practitioners. We have shown both automatic and manual clinical coding, together with their possible interactions, to demonstrate the procedures involved. Our goal in this study is to provide an overview of the technological difficulties in clinical coding, especially those on deep learning, and to suggest future lines of inquiry. The intricacy of medical terminology, the requirement for context-aware coding, and the integration of automated systems with current manual procedures are some of the difficulties. By tackling these issues, we intend to forward the creation of automated clinical coding systems that are more dependable and efficient, which will ultimately improve the effectiveness and precision of healthcare data administration.

The Need for Automated Clinical Coding

A number of serious issues with manual coding procedures make automated clinical coding necessary. These difficulties show how using artificial intelligence (AI) and natural language processing (NLP) technology might enhance clinical coding's effectiveness, precision, and general quality. The main justifications for the necessity of automated clinical coding are discussed here, along with how it can overcome the drawbacks of manual coding.

1. Time-Consuming Nature of Manual Coding

Clinical coding by hand is a labor-intensive procedure by nature. For instance, in NHS Scotland, a clinical coder usually spends 7 to 8 minutes on each of the 60 cases they handle each day. Every month, a coding section with 25–30 coders handles more than 20,000 cases. In spite of this endeavor, a backlog of cases frequently exists, which may take months or even more than a year to resolve [14]. Decision-making, policy-making, and epidemiological research all depend on prompt access to vital healthcare data, which may be hampered by this delay. Because automated clinical coding streamlines the coding process, allows for quicker turnaround times, and guarantees that healthcare data is available when needed, it has the potential to drastically minimize this backlog.

2. Prone to Errors in Manual Coding

Incomplete patient data, subjectivity in diagnosis code selection, inexperience with coding, and data input errors are some of the reasons why manual coding is prone to errors [4]. The average clinical coding accuracy, according to UK studies, is about 83%, with considerable variation between investigations, ranging from 50% to 98% [15]. According to tests conducted between 2019 and 2020, Scotland's coding accuracy is comparatively good, with primary conditions reporting 92.5% accuracy for 3-digit codes and 88.8% accuracy for 4-digit codes. There is potential for improvement, nevertheless, as under-coding is seen in 20% of frequent disorders even in Scotland [16]. By reducing human error, maintaining consistency, and following coding standards, automated coding systems—especially those that use AI and NLP—can improve accuracy. The idea that computer-assisted coding (CAC) can enhance the precision, caliber, and effectiveness of manual coding is supported by a recent qualitative literature assessment [9]. Automated coding systems can assist clinical coders more effectively by incorporating AI technology, which lowers the possibility of errors and increases data reliability.

3. Improving Efficiency and Quality

The cognitive strain required of human coders, who must evaluate, condense, and categorize intricate medical information into standardized codes, limits the effectiveness of manual coding. It frequently takes months or even years to become proficient in this technique, which calls for a great deal of training and experience [7]. By managing monotonous and time-consuming activities, automated coding systems can lessen this load, freeing up human coders to concentrate on more complicated cases that need sophisticated judgment. Automated systems can also process vast amounts of data rapidly, which helps coding teams fulfill deadlines and cut down on backlogs. Automated coding can increase productivity while simultaneously enhancing the general quality of healthcare data, increasing its dependability for use in clinical decision-making, research, and policy-making.

Automated Clinical Coding a Complex Problem to Solve:

Even though automated clinical coding has several advantages, creating efficient systems is a difficult and complex process. The complexity of classification systems, the structure of clinical documentation, and socio-technical aspects are some of the factors that make automated clinical coding challenging. We go into further depth about these difficulties below.

1. Complexity of Clinical Documents:

Because of their variable structure, length, and frequent use of symbols, acronyms, and partial information, clinical documents are naturally difficult to process. Clinical papers are frequently unstructured and contain concise notations that require domain-specific knowledge to comprehend, in contrast to standardized writings like articles or social media posts. For instance, discharge summaries in the MIMIC-III dataset, a popular dataset for intensive care, typically consist of 1,500 words and contain symbols like "+" to indicate a positive test result or "?" to indicate ambiguity, as well as acronyms like "Hep C" for hepatitis C [19,20]. A thorough understanding of a patient's records, which may comprise a variety of document types like radiology reports, pathology reports, and discharge summaries, is also necessary for clinical coding. The coding process is made more difficult by the fact that these materials are not always consistent or comprehensive [8,21].

2. Dynamic and Complex Classification Systems:

Classification schemes like ICD-10-CM and ICD-11, which are dynamic and intricate, constitute the foundation of clinical coding. About 68,000 diagnosis codes make up the ICD-10-CM system, which was introduced in the US in 2015 and is five times more extensive than the ICD-9-CM system [5]. With more than 120,000 codable terms and almost 17,000 distinct codes for illnesses, injuries, and reasons of death, the more recent ICD-11 system, which went into use in early 2022, adds even more complexity. ICD-11 is far more flexible, but it is also more difficult to use because it permits code combinations to describe almost 1.6 million clinical circumstances [22, 23]. Additionally, ICD-11 brings significant revisions to diagnostic criteria, chapter organization, and diagnostic categories, especially in areas like psychiatric categorization [24]. The "Foundation Component," a semantic network that depicts a deep polyhierarchy of medical concepts, serves as the system's backbone. Code combinations are employed to express complicated patient characteristics in "post-coordination," which is made possible by this structure [25, 26]. These complex and dynamic classification schemes, which are updated frequently to consider advancements in medical knowledge and procedures, require automated coding systems to adjust [6].

3. Socio-Technical Challenges

The adoption of automated clinical coding systems poses socio-technical issues that need to be resolved in addition to technical ones. The way coders engage with AI-based technologies must be carefully considered when a national healthcare system makes the switch to a semi-automated or completely automated coding environment. To maximize the usage of accurate automatic codes, for example, how should information be given in an automated coding system so that coders can quickly spot and fix mistakes? Building trust in automated systems is also essential; programmers need to have faith in the correctness and dependability of the system. Additionally, programmers' jobs may change from manual coding to positions like coding editors or analysts, which would call for additional education and training [9]. The implementation of automated coding systems requires a comprehensive strategy that considers both technological and human considerations, as these socio-technical issues demonstrate. The drawbacks of manual coding, such as its inefficiency, error-proneness, and time commitment, could be greatly mitigated by automated clinical coding. Automated systems can increase clinical coding speed, accuracy, and quality by utilizing AI and NLP technology, which will ultimately improve healthcare data management. Nevertheless, creating such systems is a difficult undertaking that calls for resolving issues with dynamic classification schemes, the intricacy of clinical documentation, and socio-technical factors. To fully realize the potential of automated clinical coding and improve healthcare data management, these issues must be resolved.

Role of AI: Symbolic or Neural AI?

There are two main schools of thinking in artificial intelligence (AI) that have influenced the development of automated clinical coding systems: neural AI (including deep learning) and symbolic AI. Each strategy has unique benefits and drawbacks, and the decision between the two—or combining the two—has a big impact on how well automated clinical coding systems work. The historical background, advantages, and disadvantages of these strategies are examined here, and we make the case for a hybrid solution that combines the best features of each.

Symbolic AI: Knowledge-Based Approaches:

Symbolic AI, sometimes referred to as rule-based or knowledge-based AI, models human reasoning and decision-making by using logic, rules, and symbols. The goal of symbolic AI in clinical coding is to mimic the accepted procedures and standards that human coders adhere to. To link clinical text to the proper medical codes, this method entails developing specific rules and logical expressions. For instance, clinical records have been categorized into pertinent codes using regular expressions, keyword matching, and decision trees [8,13,27-29]. From the 1950s until the beginning of the 1980s, symbolic AI dominated AI research. When used in intricate, real-world situations, such as natural language understanding (NLU), it was severely limited. Scalability was the biggest obstacle: manually creating rules to deal with the enormous and complex diversity of natural language was impractical and time-consuming, particularly for assignments comprising tens of thousands of codes and their interactions [17, 18]. Rule-based approaches have proven to be highly accurate in clinical coding despite these drawbacks, especially when utilized to assist human coders. For example, keyword matching and regular expressions can detect certain terms or phrases with high accuracy, but they frequently have limited recall, which means they might overlook crucial information that does not fit the predetermined rules [30].

Neural AI: Deep Learning Approaches:

Neural AI, especially deep learning, differs from symbolic AI in that it learns intricate correlations and patterns straight from data. Artificial intelligence (AI) has been transformed by deep learning models like transformers and convolutional neural networks (CNNs), which have achieved state-of-the-art performance in a variety of applications, such as image recognition and natural language processing (NLP). To assign one or more medical codes to a given clinical document depending on its content is the aim of deep learning approaches that structure the issue as a multi-label classification problem in the context of automated clinical coding [31-34]. Compared to symbolic AI, deep learning offers a number of benefits. First, because the models acquire pertinent features straight from the training data, it does away with the requirement for feature engineering and manually created rules. This increases the adaptability and ease of applying deep learning techniques to new coding systems or datasets. Second, when compared to conventional machine learning techniques, deep learning models can perform better overall because they can capture intricate, non-linear correlations in the data [32]. Deep learning has been used more and more for automated clinical coding since about 2017, which has sparked a boom in research in this field [31, 32].

Deep learning for medical coding is becoming more and more popular, according to recent surveys and carefully selected paper collections [11,12,13,33]. Deep learning is not without its difficulties, though. Its need for vast volumes of labeled training data, which can be challenging to acquire in the clinical setting because of privacy issues and the intricacy of medical terminology, is one of its key drawbacks. Furthermore, deep learning models frequently function as "black boxes," which makes it challenging to understand their judgments or integrate domain-specific knowledge, like coding standards or ontologies. This lack of interpretability can be a serious disadvantage in the healthcare industry, where accountability and transparency are essential.

The Case for a Hybrid Approach:

Although deep learning has emerged as the most popular method for automated clinical coding, its shortcomings are increasingly being acknowledged by the necessity to use symbolic AI techniques. A promising answer is provided by a hybrid strategy that blends the advantages of neural and symbolic AI. This method uses deep learning's scalability and pattern recognition skills in conjunction with symbolic AI's interpretability and domain-specific knowledge. Using knowledge-augmented deep learning techniques, which include external knowledge sources like ontologies and knowledge graphs into deep learning models, is one new approach. For instance, embedding-based methods for integrating knowledge graphs into deep learning frameworks have been investigated in a number of publications [37-40]. By strengthening the model's comprehension of medical concepts and their connections, these techniques hope to improve the model's capacity to assign precise codes. However, the majority of previous research has concentrated on utilizing the target ontology's hierarchical structure and terminology (such as ICD-9 or ICD-10), leaving other important knowledge sources like SNOMED-CT and UMLS (Unified Medical Language System) underdeveloped [38]. Integrating coding standards and recommendations into deep learning models presents another potential. These location-specific and frequently updated standards offer a vital context for precise coding. A major problem, though, is to extract and encode this knowledge in a way that deep learning models can use. To make sure that the models are in line with practical procedures, cooperation between clinical coding specialists and AI researchers is necessary. The argument between neural and symbolic AI for automated clinical coding is about combining their advantages rather than picking one over the other. While neural AI offers scalability and the capacity to extract intricate patterns from data, symbolic AI offers interpretability and the ability to integrate domain-specific knowledge. The most promising method for overcoming the difficulties associated with automated clinical coding is a hybrid strategy that combines the two paradigms, such as knowledge-augmented deep learning. We can create systems that are accurate and transparent by utilizing the extensive knowledge found in clinical ontologies and coding guidelines, along with the strength of deep learning. This will ultimately increase the effectiveness and dependability of clinical coding in healthcare systems across the globe.

State-of-the-Art Deep Learning Models:

Clinical coding is a very difficult work that presents many difficulties for modern AI, especially in the areas of deep learning and natural language processing (NLP). Although deep learning models have made significant strides in automating clinical coding in recent years, the problem is still far from being resolved. To translate a patient's clinical notes into a collection of pertinent medical codes, state-of-the-art deep learning models for clinical coding are mostly based on multi-label classification. Nevertheless, these models have a number of drawbacks, such as trouble digesting lengthy texts, managing rare or invisible codes, and a lack of symbolic reasoning skills. We examine the present state of deep learning models, their successes, and their shortcomings below.

The Multi-Label Classification Framework:

Learning a complex, non-linear function that associates a patient's clinical notes with a collection of medical codes is the fundamental idea behind the most advanced deep learning models for clinical coding. Usually, this is presented as a multi-label classification problem, in which every clinical remark has the potential to be linked to several codes. For this objective, deep learning models have become the norm, especially those based on transformer architectures such as BERT (Bidirectional Encoder Representations from Transformers) [41-44]. The semantic content of clinical literature is captured by these models using pre-trained language representations, which are subsequently refined on particular coding tasks. Even with their success, these models continue to perform below par on benchmark datasets such as MIMIC-III (Medical Information Mart for Intensive Care III) [20]. For instance, for the entire set of 8,932 ICD-9 codes in MIMIC-III, the optimal Micro-F1 scores (a harmonic mean of precision and recall) range from 58% to 60% [45-52]. This demonstrates how challenging the problem is, even for sophisticated deep-learning models. Furthermore, despite being widely utilized, MIMIC-III has limited representativeness. The information, which was gathered between 2001 and 2012, is more than 10 years old and only includes US intensive care patients. This restricts its applicability to more varied clinical settings or other areas, like the UK.

Key Challenges in Deep Learning for Clinical Coding

While deep learning models have shown promise, they face several major challenges when applied to clinical coding:

1. **Handling Unseen, Infrequent, and Imbalanced Labels**: The unequal distribution of codes is one of the biggest problems with clinical coding. Over 50% of codes in the MIMIC-III dataset never appear at all, and about 5,000 codes appear less than ten times in the training data [37]. Deep learning models, which depend on a lot of labeled data to train efficiently, are challenged by this. Because they are unable to generalize to uncommon or unexpected scenarios, vanilla deep learning models have trouble with infrequent or unseen codes. In multi-label categorization, where there are many possible code combinations and the labels are wildly unbalanced, this problem is made worse. To tackle this problem, strategies like zero-shot learning [37,46] and meta-learning [47] have been investigated, but they are still being studied.

2. **Lack of Symbolic Reasoning Capabilities**: To arrive at the correct codes for clinical coding, coders must frequently use complicated reasoning to integrate unrelated bits of information from a patient's records. For instance, a coder may have to apply coding principles to evaluate ambiguous circumstances or reconcile contradicting information from various sources [8,43]. However, deep learning algorithms do not directly describe the reasoning process; instead, they mostly use labeled data to learn relationships between text and codes. This restricts their capacity to reason like humans, which is essential for precise coding. Deep learning models' performance and explainability may be improved by integrating symbolic reasoning, for example, by formalizing coding rules into logical expressions [29] or by utilizing knowledge graphs [37-40]. This field is yet unexplored, though.

3. **Handling Long Documents**: Clinical notes are frequently long and contain repetitive material, especially discharge summaries. Discharge summaries in MIMIC-III typically have 1,500 tokens, while some documents have 10,000 tokens or more [19]. For deep learning models, which must find pertinent information for every code in a vast amount of text, this presents a "needle-in-a-haystack" difficulty. Due to the memory-intensive nature of their self-attention processes, transformer-based models such as BERT can only process inputs of up to 512 sub-

word tokens [54]. This maximum has been raised to 4,096 tokens by recent developments like Longformer [55], TransformerXL [56], and BigBird [57], although many clinical records still require more. The process is further complicated by the fact that clinical notes sometimes contain text repetition, also known as "note bloat" [58,59]. To solve this problem, methods like text de-duplication based on similarity metrics have been put forth, although research is still ongoing.

4. Recent Advances and Future Directions

Despite these challenges, deep learning models have made significant strides in clinical coding. Recent studies have explored various techniques to improve performance, including:

- **Text Representation Learning**: Advances in text representation learning, such as pre-trained language models, have enabled models to capture the semantic meaning of clinical text more effectively [19,44].
- **Multi-Task Learning**: By training models on multiple related tasks simultaneously, multi-task learning can improve generalization and performance [41,45].
- **Zero-Shot and Meta-Learning**: These approaches aim to improve the model's ability to handle rare or unseen codes by learning from limited data [37,46,47].
- **Multi-Modal Learning**: Integrating multiple data sources, such as text and structured data, can enhance the model's understanding of patient information [48].

However, resolving the aforementioned constraints is necessary to develop a complete deep learning-based clinical coding system. Future studies should concentrate on utilizing external information sources like ontologies and coding standards, enhancing the processing of lengthy documents, and incorporating symbolic reasoning capabilities. Furthermore, creating more current and representative datasets will be essential to the field's advancement. Clinical coding automation has advanced significantly thanks to state-of-the-art deep learning models, but the work is still difficult and complex. Although transformer-based multi-label classification frameworks have produced impressive results, they are limited in their ability to scan lengthy documents, handle uncommon codes, and carry out symbolic reasoning. A variety of cutting-edge strategies, such as knowledge-augmented deep learning, enhanced text representation approaches, and the creation of more representative datasets, will be needed to meet these problems. We can get closer to developing automated clinical coding systems that are precise, dependable, and explicable by bridging the gap between deep learning and symbolic reasoning [49-55].

Potential Challenges:

The task of automated clinical coding is intricate and multidimensional, requiring the resolution of several technological, pragmatic, and domain-specific problems. Even though deep learning models have shown promise, especially those built on transformer architectures like BERT, they still have a long way to go before they can perform on par with humans. Creating high-quality datasets, managing heterogeneous and noisy data, enhancing explainability, integrating human feedback, enabling few-shot and zero-shot learning, adjusting to terminology changes, and integrating knowledge representation and reasoning are some of the major issues that must be resolved to advance automated clinical coding. We also go over the function of the industry and the necessity of customized solutions for various settings and goals [56-59].

1. Creating Gold Standard Coding Datasets

The absence of huge, excellent, and publicly accessible datasets is one of the biggest obstacles to automated clinical coding. Despite its value, the popular MIMIC-III dataset has drawn criticism for perhaps being under-coded, which means that many pertinent codes might not be present in the annotations [60-62]. This restricts the generalizability of models developed on MIMIC-III to other datasets or real-world situations.

Furthermore, MIMIC-III is unique to US critical care patients and might not accurately represent the variety of clinical notes in other countries, such as the UK, China, or Spain. More representative and varied datasets that are expert-labeled and span a wider range of clinical contexts and objectives are required to meet this challenge. For instance, datasets intended for billing reasons might concentrate on Diagnosis-Related Groups (DRGs) or Healthcare Resource Groups (HRGs), whereas datasets customized for epidemiological studies might need to link clinical notes to complex terminologies like SNOMED CT. Making such datasets available will aid in the creation of models that are more resilient and broadly applicable.

2. Coding from Heterogeneous, Incomplete, and Noisy Sources

Analyzing a variety of papers, such as laboratory findings, radiology reports, pathology reports, and discharge summaries, is usually necessary for clinical coding. Nevertheless, the majority of recent research only looks at discharge summaries, which restricts how thorough the coding procedure may be [14]. Clinical data in the real world is frequently unreliable, noisy, and presented inconsistently. Discharge summaries, for instance, can be typed or handwritten, with differing degrees of thoroughness and information. Another major obstacle is managing lengthy papers. Clinical notes are frequently longer than the 512-token limit that transformer-based algorithms like BERT can process. This limit has been raised by recent developments like Longformer and BigBird, although they are still unable to process the longest documents. Furthermore, the work is made more difficult by the abundance of duplicated information, or "note bloat," in clinical notes [58]. Though they need further study and improvement, strategies like text de-duplication and summary could be able to aid with this problem.

3. Explainability of Clinical Coding

For automated clinical coding systems, explainability is essential since coders must comprehend the decision-making process. Present-day deep learning models frequently function as "black boxes," making it challenging to decipher their results, especially those that rely on multi-label classification. Although some studies have highlighted important words or phrases in the text using attention processes [19,61,63], these highlights frequently show connections rather than causation. Future systems should use rule-based techniques and symbolic reasoning to increase explainability. These approaches can offer more transparent and comprehensible decision-making procedures. For instance, using knowledge graphs or formalizing coding guidelines into logical expressions could improve the model's capacity to justify its choices. To assess the value of explainability qualities for clinical coders, user studies are also required.

4. Human-in-the-Loop Learning with Coders' Feedback:

For automated coding systems to be implemented in practice, clinical coders' input must be incorporated. Coders can offer insightful information that can be utilized to iteratively enhance the system, such as rules, highlights, and manual corrections. The model's performance can be gradually enhanced and problems like under-coding can be addressed by human-in-the-loop learning, in which programmers actively take part in the training process [9]. Another interesting strategy is active learning, which entails choosing the most instructive samples for human annotation. The MedCATTrainer system, for instance, employs active learning to find samples that coders can examine and utilize to update the model [64]. In the same vein, users can improve the results by adding labels for mentions in SemEHR [65]. These methods show how human knowledge and machine learning can be combined to provide more accurate and dependable code.

5. Few-Shot and Zero-Shot Learning:

A major problem for multi-label classification models is that many medical codes are either rare or completely missing from the training data. Approximately 5,000 codes appear less than 10 times in the MIMIC-III dataset, and more than 50% of codes never appear [37]. Models find it challenging to generalize to uncommon or invisible codes as a result. To overcome this difficulty, few-shot and zero-shot learning strategies allow models to learn with little to no labeled input. For instance, knowledge-based strategies can aid in bridging the gap between seen and unseen codes by utilizing code descriptions, hierarchies, and linkages from ontologies [37,46,47]. When adjusting models to new or updated coding systems, like the switch from ICD-10 to ICD-11, these methods are very crucial.

6. Adaptation to Terminology Changes:

One of the biggest challenges for automated clinical coding is switching from one coding system to another, such as ICD-10 to ICD-11. Significant modifications are brought about by ICD-11, such as new diagnostic categories, post-coordination of codes, and a more intricate semantic network structure [24]. In addition to precise ontology matching and idea drift management, new paradigms like self-supervised learning, transfer learning, and meta-learning are needed to adapt models to these shifts.

7. Knowledge Representation and Reasoning in Coding:

To achieve human-like performance, automated coding systems must incorporate knowledge representation and reasoning. Symbolic reasoning, which is essential for activities like applying coding principles or reconciling contradicting information, is frequently absent from current models. The model's comprehension of clinical material and its capacity to assign precise codes can both be improved by knowledge graphs, which show the connections between medical concepts [37-40]. Another crucial step is to formalize coding guidelines into rules that can be read by machines. To increase the model's precision and explainability, for instance, rules pertaining to hypothetical situations, mutual exclusion, and the precedence of particular codes might be incorporated [29]. Furthermore, using other ontologies like SNOMED CT and UMLS might give important context for coding choices.

8. Tailoring Systems for Different Purposes and Contexts:

Automated clinical coding systems must be customized for various settings (e.g., different countries or healthcare systems) and goals (e.g., billing versus health-related research). Systems may concentrate on predicting DRGs or HRGs for billing reasons because they have fewer codes and are frequently clustered from the entire set of ICD codes [66-75]. Systems may need to favor precision over recall and offer a broader range of terminologies, like SNOMED CT or ORDO, for research purposes [76-82]. It's also necessary to consider country-specific elements like billing systems and documentation procedures. To enhance model performance, for instance, specific preprocessing methods are needed to address the "note bloat" issue in the US, where redundant information is copied and pasted into clinical notes [58].

9. Industry Collaboration and Proprietary Solutions:

Automated clinical coding is being advanced by industry organizations, who frequently work with academic institutions to create and implement solutions. For instance, to decrease documentation time and increase coding accuracy, the CogStack team is collaborating with NHS Trusts in England to include NLP technologies such as MedCAT into the Epic EHR system [83,84]. Although they offer pre-built solutions for clinical idea extraction, industry APIs like Microsoft Text Analytics for Health and Amazon Comprehend Medical frequently lack transparency and research access [85-87]. Additionally, startups like AKASA are creating deep learning-based automated coding solutions, demonstrating cutting-edge results on benchmark datasets [50,88]. These initiatives demonstrate the growing interest in automated clinical coding and its potential to revolutionize research and healthcare administration. With the potential to greatly increase the effectiveness and precision of healthcare data management, automated clinical coding is a difficult but exciting field of study. To advance the discipline, it will be essential to address the main issues mentioned above, such as creating datasets, managing noisy data, enhancing explainability, integrating knowledge representation, and incorporating human feedback. Automated coding systems that satisfy the requirements of academics, policymakers, and healthcare professionals can be created by fusing the advantages of deep learning and symbolic reasoning and customizing solutions for various uses and situations.

Overview of Medical and Clinical Coding:

The need to decrease the time-consuming nature of human chart checks and increase the effectiveness and precision of healthcare data management has motivated decades of efforts to automate clinical coding and classification procedures. Researchers have investigated a number of automated system uses throughout the years, such as automating biosurveillance, applying clinical guidelines, structuring text for clinical decision support, and selecting clinical trial patients. However, this has proven to be a difficult effort due to the intricacy of clinical writing, the unpredictability of coding tasks, and the limits of natural language processing (NLP) techniques. The development of automated clinical coding, how well these systems perform in comparison to human coders, and the main obstacles still facing the field are all examined in this systematic review.

Historical Context and Evolution of Automated Clinical Coding:

Both environmental conditions and technical improvements have influenced the development of automated methods for clinical coding and classification. Beginning in the mid-1990s, there were early attempts to automate processes like selecting clinical trial participants or following clinical recommendations. The need to improve care quality and expedite healthcare procedures motivated these initiatives. For instance, automated biosurveillance systems were timed to coincide with actual events. Anthrax exposures in 2001 and the Salt Lake City Olympics in 2002 sparked further advancements after the first such system was tested at the 1996 Atlanta Olympics. In light of the twin goals of raising healthcare quality and cutting costs, the emphasis has recently switched to automating administrative duties including keeping problem lists and reporting quality metrics.

Performance of Automated Systems: Humans vs. Machines:

One of the main topics in the literature has been the effectiveness of automated coding and classification systems. Automated systems could beat laypeople and perform at least as well as physicians in simple binary tasks, including diagnosing acute bacterial pneumonia on chest X-ray reports, according to early evaluations like those conducted by Chapman and Haug in 1999 [88]. Although direct comparisons were difficult due to human perception and contextual understanding, both systems were praised for their consistency. The limitations of automated systems in complicated reasoning tasks were highlighted by Elkins et al. in 2000, who found that although computers could perform well on binary tasks, their performance deteriorated when many parameters were involved [89]. Researchers have continuously highlighted the potential of automated systems despite these drawbacks. Chapman et al. concluded in 2003 that text processing algorithms were getting precise enough to handle actual medical issues [90]. However, Kukafka et al. noted in 2006 that complicated reasoning tasks, such as tying together different pieces of knowledge, continued to be a major issue for NLP systems [91]. Only four of the 113 papers that made up this review specifically claimed that people outperformed automated systems, whereas 26 claimed that automated systems performed on par with or better than humans. This implies that even while automated systems have advanced significantly, the task's complexity and the intended result greatly influence how well they work.

Challenges in Evaluating Automated Systems:

There are many difficulties in assessing how well automated coding and categorization systems function. One significant problem is the diversity of research approaches, which makes it challenging to compare study findings. For instance, whereas some research used routine practice as the reference standard, others developed a gold standard for comparison. Furthermore, comparisons were made more difficult by the significant variations in statistical techniques employed to assess system performance. These investigations also differed greatly in the level of intricacy of the coding and categorization schemas employed. While some systems created unique coding schemes, especially for the study, others used pre-existing classification schemes like SNOMED CT or ICD. It is challenging to draw broad judgments regarding automated systems' performance because of this unpredictability. Simpler jobs are more likely to produce positive results for automated systems, thus further research is required to associate the coding task's complexity with the study's findings.

Specific Challenges in Administrative Coding:

Examining the seven studies that concentrated on administrative coding in greater detail reveals the difficulties in automating this particular process. These studies assessed different systems using diverse document kinds and classification approaches. Although some findings were encouraging, they also showed important drawbacks. For instance, Warner [92] and Dinwoodie and Howell [93] biased their conclusions by evaluating systems exclusively in situations where they could code with confidence. Applying evaluation and management (E/M) code levels, a particularly challenging subset of codes, was inconsistent even for humans, according to Morris et al. [94], highlighting the task's difficulty. Despite identifying areas for development, Lussier et al. [95] concluded that the system was not yet prepared for production. In a similar vein, Goldstein et al. [96] and Kukafka et al. [91] presented findings that, although promising, did not show appreciable advancements over earlier systems. The most encouraging results were obtained by Pakhomov et al. [97], who reported Type B results ranging from 90% to 95% accuracy and Type A results reaching 98% accuracy. As a possible next step for incorporating automation into practical applications, these authors also suggested a tiered approach that blends automated coding with human supervision.

The Complexity of Clinical Texts:

The intricacy of clinical texts is one of the biggest obstacles to automating clinical coding. Natural language understanding (NLU) of narrative text documents is intrinsically challenging, as stated by Barrows et al. [98]. However, the difficulty is exacerbated when working with notational text documents. These texts may be fragmentary or inconsistent, frequently use terse symbols and abbreviations, and frequently lack punctuation and language. Phrases like "a [xx] y/o M w/ Hep C, HTN, CKD, a/w HTN emergency," for instance, may appear in a discharge statement and are difficult to correctly understand without domain-specific knowledge [99]. Because of this intricacy, automated systems find it challenging to attain the precision and dependability needed for practical applications.

The Need for Generalizability and Adaptability:

The challenge of generalizing and adapting NLP techniques for various purposes is another recurrent issue in literature. This restriction was brought to light by Turchin et al. [100], who pointed out that fresh sets of regular expressions frequently need to be created and verified for every unique task. Because it restricts their application across many healthcare settings and specialties, this lack of generalizability is a major obstacle to the widespread adoption of automated coding systems. Over the past few decades, automated clinical coding and classification systems have advanced significantly, although there are still issues. The difficulty of the task and the caliber of the reference standard have a significant impact on how well these systems function, even though they can do well on some tasks. The assessment and comparison of these systems are made more difficult by the diversity of research approaches and the intricacy of clinical texts. More standardized assessment techniques, enhanced generalizability of NLP tools, and a deeper comprehension of the integration of automated systems with human supervision are all necessary going forward. Automated clinical coding systems can get closer to achieving their full potential in raising the effectiveness and caliber of healthcare by tackling these issues [101].

Conclusion:

Automated clinical coding represents a transformative opportunity to enhance healthcare data management by addressing the inefficiencies and limitations of manual coding. Over the past few decades, significant progress has been made in developing AI-driven systems, particularly those leveraging deep learning and NLP. However, the journey toward fully automated clinical coding is far from complete, as several technical and practical challenges remain unresolved. One of the most pressing challenges is the complexity of clinical texts, which often contain unstructured, incomplete, and noisy information. Deep learning models, while powerful, struggle with tasks requiring symbolic reasoning, such as reconciling contradictory information or applying coding guidelines. This limitation underscores the need for hybrid approaches that combine the strengths of symbolic AI, with its interpretability and rule-based reasoning, and neural AI, with its ability to learn complex patterns from data. Knowledge-augmented deep learning, which integrates external knowledge sources like ontologies and coding guidelines, offers a promising pathway to bridge this gap. Another critical challenge is the handling of infrequent and unseen codes, which are common in clinical datasets. Few-shot and zero-shot learning techniques, along with the integration of knowledge graphs, can help models generalize to rare or new codes, improving their applicability in real-world scenarios. Additionally, the development of high-quality, expert-labeled datasets is essential to train and evaluate these models effectively. Current benchmark datasets like MIMIC-III, while valuable, are limited in scope and representativeness, highlighting the need for more diverse and up-to-date datasets. Explainability and human-in-the-loop learning are also crucial for the successful deployment of automated coding systems. Coders need to understand how decisions are made, and systems must be designed to incorporate feedback from users, enabling iterative improvements. Techniques like active learning and attention mechanisms can enhance the transparency and usability of these systems, fostering trust and adoption among healthcare professionals. Finally, the transition to new coding systems, such as ICD-11, presents both challenges and opportunities. Automated systems must be adaptable to changes in terminology and coding standards, requiring advancements in self-supervised learning, transfer learning, and ontology matching. In conclusion, while automated clinical coding has made significant strides, achieving human-level performance will require addressing these challenges through interdisciplinary collaboration, innovative hybrid approaches, and a focus on explainability and adaptability. By doing so, automated coding systems can revolutionize healthcare data management, improving efficiency, accuracy, and patient outcomes.

Conflict of interest: None

Funding: None

References:

1. Public Health Scotland. National Data Catalogue. General acute inpatient and day case - Scottish Morbidity Record (SMR01). https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=5 (2020).
2. American Academy of Professional Coders (AAPC). What is medical coding? https://www.aapc.com/medical-coding/medical-coding.aspx (2022).
3. NHS Digital. Clinical coding for non coders. https://hscic.kahootz.com/gf2.ti/f/762498/30719205.1/PPSX/-

/Coding_for_non_coders_automaticnew.ppsx (2017).

4. Enrico, C. In *Guide to Health Informatics* Ch. 24 (Taylor & Francis Group, 2015).
5. National Center for Health Statistics. International Classification of Diseases, (ICD-10-CM/PCS) transition – background. https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm (2015).
6. Public Health Scotland. Terminology Services. Scottish Clinical Coding Standards. https://www.isdscotland.org/Products-and-services/Terminology-services/Clinical-coding-guidelines/ (2022).
7. Otero Varela, L. et al. International Classification of Diseases clinical coding training: an international survey. *Health Inf. Manag*. https://doi.org/10.1177/18333583221106509 (2022)
8. Stanfill, M. H., Williams, M., Fenton, S. H., Jenders, R. A. & Hersh, W. R. A systematic literature review of automated clinical coding and classification systems. *J. Am. Med Inf. Assoc.* **17**, 646–651 (2010).
9. Campbell, S. & Giadresco, K. Computer-assisted clinical coding: a narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals. *HIM J.* **49**, 5–18 (2020).
10. Jiang, F. et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc. Neurol*. https://doi.org/10.1136/svn-2017-000101 (2017)
11. Kaur, R., Ginige, J. A. & Obst, O. AI-based ICD coding and classification approaches using discharge summaries: A systematic literature review. *Expert Syst. Appl.* 118997 https://doi.org/10.1016/j.eswa.2022.118997 (2022).
12. Ji, S., Sun, W., Dong, H., Wu, H. & Marttinen, P. A unified review of deep learning for automated medical coding. Preprint at *arXiv* http://arxiv.org/abs/2201.02797 (2022).
13. Teng, F. et al. A review on deep neural networks for ICD coding. In *IEEE Transactions on Knowledge and Data Engineering* 1–19 (IEEE, 2022)
14. Alonso, V. et al. Problems and barriers during the process of clinical coding: a Focus Group Study of coders' perceptions. *J. Med. Syst.* **44**, 62 (2020).
15. Burns, E. M. et al. Systematic review of discharge coding accuracy. *J. Public Health* **34**, 138–148 (2012).
16. Public Health Scotland. Data quality assurance. Assessment of SMR01 Data Scotland Report 2019 V1. https://beta.isdscotland.org/media/7465/assessment-of-smr01-data-scotland-report-2019-v1.pdf (2019).
17. Wooldridge, M. *The Road to Conscious Machines: The Story of AI* (Penguin UK, 2020).
18. Russell, S. J. & Norvig, P. *Artificial Intelligence: A Modern Approach, Global Edition* (Pearson, 2021).
19. Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J. & Eisenstein, J. Explainable prediction of medical codes from clinical text. In *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* 1101–1111 (Association for Computational Linguistics, 2018).
20. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).
21. Barrows Jr, R. C., Busuioc, M. & Friedman, C. Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. In *Proc. AMIA Symposium* 51 (American Medical Informatics Association, 2000).
22. World Health Organization. *ICD-11 for Mortality and Morbidity Statistics* (WHO, 2022).
23. World Health Organization. WHO's new International Classification of Diseases (ICD-11) comes into

effect. https://www.who.int/news/item/11-02-2022-who-s-new-international-classification-of-diseases-(icd-11)-comes-into-effect (2022).

24. Gaebel, W., Stricker, J. & Kerst, A. Changes from ICD-10 to ICD-11 and future directions in psychiatric classification. *Dialogues Clin. Neurosci.* **22**, 7–15 (2020).
25. Chute, C. G. The rendering of human phenotype and rare diseases in ICD-11. *J. Inherit. Metab. Dis.* **41**, 563–569 (2018).
26. World Health Organization. ICD-11 Reference Guide. 2.10 Precoordination and postcoordination. https://icdcdn.who.int/icd11referenceguide/en/html/index.html#precoordination-and-postcoordination (2022).
27. Bengio, Y., Lecun, Y. & Hinton, G. Deep learning for AI. *Commun. ACM* **64**, 58–65 (2021).
28. Dinwoodie, H. P. & Howell, R. W. Automatic disease coding: the 'fruit-machine' method in general practice. *Br. J. Prev. Soc. Med.* **27**, 59–62 (1973).
29. Farkas, R., & Szarvas, G. Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics* **9**, 1–9 (2008).
30. Zhou, L., Cheng, C., Ou, D. & Huang, H. Construction of a semi-automatic ICD-10 coding system. *BMC Med. Inform. Decis. Mak.* **20**, 1–12 (2020).
31. Shi, H., Xie, P., Hu, Z., Zhang, M. & Xing, E. P. Towards automated ICD coding using deep learning. Preprint at *arXiv* https://arxiv.org/abs/1711.04075 (2017) .
32. Karimi, S., Dai, X., Hassanzadeh, H. & Nguyen, A. Automatic diagnosis coding of radiology reports: a comparison of deep learning and conventional classification methods. in *BioNLP 2017* 328–332 (Association for Computational Linguistics, 2017).
33. acadTags. Awesome-medical-coding-NLP. https://github.com/acadTags/Awesome-medical-coding-NLP (2022).
34. Nam, J., Kim, J., Loza Mencía, E., Gurevych, I. & Fürnkranz, J. In *Machine Learning and Knowledge Discovery in Databases* (eds. Calders, T., Esposito, F., Hüllermeier, E. & Meo, R.) 437–452 (Springer, 2014).
35. Kraljevic, Z. et al. Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit. *Artif. Intelligence Med.* **117**, 102083 (2021).
36. Wiegreffe, S., Choi, E., Yan, S., Sun, J. & Eisenstein, J. Clinical concept extraction for document-level coding. In *Proc. 18th BioNLP Workshop and Shared Task* 261–272 (Association for Computational Linguistics, 2019)
37. Rios, A. & Kavuluru, R. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proc. 2018 Conference on Empirical Methods in Natural Language Processing* 3132–3142 (Association for Computational Linguistics, 2018).
38. Teng, F., Yang, W., Chen, L., Huang, L. & Xu, Q. Explainable prediction of medical codes with knowledge graphs. *Front. Bioeng. Biotechnol.* **8**, 867 (2020).
39. Xie, X., Xiong, Y., Yu, P. S. & Zhu, Y. EHR coding with multi-scale feature attention and structured knowledge graph propagation. In *Proc. 28th ACM International Conference on Information and Knowledge Management* 649–658 (ACM, 2019).
40. Cao, P. et al. Hypercore: hyperbolic and co-graph representation for automatic ICD coding. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* 3105–3114 (Association for Computational Linguistics, 2020).
41. Falis, M. et al. Ontological attention ensembles for capturing semantic concepts in ICD code prediction from clinical text. In *Proc. Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)* 168–177 (Association for Computational Linguistics, 2019).

42. Falis, M., Dong, H., Birch, A. & Alex, B. CoPHE: a count-preserving hierarchical evaluation metric in large-scale multi-label text classification. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* 907–912 (Association for Computational Linguistics, 2021).
43. Kukafka, R., Bales, M. E., Burkhardt, A. & Friedman, C. Human and automated coding of rehabilitation discharge summaries according to the international classification of functioning, disability, and health. *J. Am. Med. Inform. Assoc.* **13**, 508–515 (2006).
44. Ji, S., Hölttä, M. & Marttinen, P. Does the magic of BERT apply to medical code assignment? A quantitative study. *Computers Biol. Med.* **139**, 104998 (2021).
45. Sun, W., Ji, S., Cambria, E. & Marttinen, P. Multitask balanced and recalibrated network for medical code prediction. *ACM Trans. Intelligent Syst. Technol*. https://doi.org/10.1145/3563041 (2022)
46. Chalkidis, I. et al. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 7503–7515 (Association for Computational Linguistics, 2020).
47. Wang, R. et al. Meta-LMTC: meta-learning for large-scale multi-label text classification. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* 8633–8646 (Association for Computational Linguistics, 2021).
48. Xu, K. et al. Multimodal machine learning for automated ICD coding. In *Machine Learning for Healthcare Conference* 197–215 (PMLR, 2019).
49. Liu, Y., Cheng, H., Klopfer, R., Gormley, M. R. & Schaaf, T. Effective convolutional attention network for multi-label clinical document classification. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* 5941–5953 (Association for Computational Linguistics, 2021).
50. Kim, B. H., & Ganapathi, V. Read, attend, and code: pushing the limits of medical codes prediction from clinical notes by machines. In *Machine Learning for Healthcare Conference* 196–208 (PMLR, 2021).
51. Yuan, Z., Tan, C., & Huang, S. Code synonyms do matter: multiple synonyms matching network for automatic ICD coding. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* 808–814 (Association for Computational Linguistics, 2022).
52. Huang, C. W., Tsai, S. C., & Chen, Y. N. PLM-ICD: automatic ICD coding with pretrained language models. In *Proc. 4th Clinical Natural Language Processing Workshop* 10–20 (Association for Computational Linguistics, 2022).
53. Terminology and Classifications Delivery Service, National Health Service Digital. National Clinical Coding Standards ICD-10 5th Edition. https://classbrowser.nhs.uk/ref_books/ICD-10_2021_5th_Ed_NCCS.pdf (2021).
54. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* 4171–4186 (Association for Computational Linguistics, 2019).
55. Feucht, M., Wu, Z., Althammer, S. & Tresp, V. Description-based label attention classifier for explainable ICD-9 classification. In *Proc. Seventh Workshop on Noisy User-generated Text (W-NUT 2021)* 62–66 (Association for Computational Linguistics, 2021).
56. Yogarajan, V., Pfahringer, B., Smith, T., & Montiel, J. In *Artificial Neural Networks and Machine Learning – ICANN 2022* (eds. Pimenidis, E., Angelov, P., Jayne, C.,

Papaleonidas, A. & Aydin, M.) 209–221 (Springer Nature Switzerland, 2022).

57. Michalopoulos, G., Malyska, M., Sahar, N., Wong, A. & Chen, H. ICDBigBird: a contextual embedding model for ICD code classification. In *Proc. 21st Workshop on Biomedical Language Processing* 330–336 (Association for Computational Linguistics, 2022).
58. Liu, J., Capurro, D., Nguyen, A. & Verspoor, K. “Note Bloat” impacts deep learning-based NLP models for clinical prediction tasks. *J. Biomed. Inform*. **133**, 104149 (2022)
59. Searle, T., Ibrahim, Z., Teo, J. & Dobson, R. Estimating redundancy in clinical text. *J. Biomed. Inform.* **124**, 103938 (2021).
60. Gao, S. et al. Limitations of transformers on clinical text classification. *IEEE J. Biomed. Health Inform*. **25**, 3596–3607 (2021).
61. Dong, H., Suárez-Paniagua, V., Whiteley, W. & Wu, H. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *J. Biomed. Inform.* **116**, 103728 (2021).
62. Searle, T., Ibrahim, Z. & Dobson, R. Experimental evaluation and development of a silver-standard for the MIMIC-III clinical coding dataset. In *Proc. 19th SIGBioMed Workshop on Biomedical Language Processing* 76–85 (Association for Computational Linguistics, 2020).
63. Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M. & Elhadad, N. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*. 409-416 (2018).
64. Searle, T., Kraljevic, Z., Bendayan, R., Bean, D. & Dobson, R. MedCATTrainer: a biomedical free text annotation interface with active learning and research use case specific customisation. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations* 139–144 (Association for Computational Linguistics, 2019).
65. Wu, H. et al. SemEHR: a general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *J. Am. Med. Inform. Assoc.* **25**, 530–537 (2018).
66. Dong, H. et al. Rare disease identification from clinical notes with ontologies and weak supervision. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)* 2294–2298 (EMBC, 2021).
67. Dong, H. et al. Ontology-based and weakly supervised rare disease phenotyping from clinical notes. Preprint at http://arxiv.org/abs/2205.05656 (2022).
68. Ferreira, M. D. et al. Active learning for medical code assignment. In *Workshops from ACM Conference on Health, Inference, and Learning (CHIL) 2021.* Preprint at *arXiv* http://arxiv.org/abs/2104.05741 (2021).
69. Chen, J. et al. Knowledge-aware zero-shot learning: survey and perspective. In *Proc. Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021* 4366–4373 (IJCAI, 2021).
70. Falis, M. Blood is thicker than water, a hierarchical evaluation metric for document classification. https://www.ltg.ed.ac.uk/blood-is-thicker-than-water/ (2021).
71. Healthcare Cost and Utilization Project (HCUP). Clinical classifications software (CCS) for ICD-9-CM. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp (2017).
72. Hahn, U. & Oleynik, M. Medical information extraction in the age of deep learning. *Yearb. Med. Inform.* **29**, 208–220 (2020).
73. Falis, M., Dong, H., Birch, A. & Alex, B. Horses to zebras: ontology-guided data augmentation and synthesis for ICD-9 coding. In *Proc. 21st Workshop on Biomedical Language Processing* 389–401 (Association for Computational Linguistics, 2022).
74. DeYoung, J., Shing, H.-C., Kong, L., Winestock, C. & Shivade, C. Entity anchored ICD coding.

Accepted to American Medical Informatics Association (AMIA) 2022 Annual Symposium. Preprint at *arXiv* http://arxiv.org/abs/2208.07444 (2022).

75. Liu, J., Capurro, D., Nguyen, A. & Verspoor, K. Early prediction of diagnostic-related groups and estimation of hospital cost by processing clinical notes. *NPJ Digital Med.* **4**, 1–8 (2021).
76. Donnelly, K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud. Health Technol. Inform.* **121**, 279 (2006).
77. Vasant, D. et al. ORDO: an ontology connecting rare disease, epidemiology and genetic data. In *Bio-Ontology @ ISMB 2014*. 1-4. https://www.researchgate.net/publication/281824026_ORDO_An_Ontology_Connecting_Rare_Disease_Epidemiology_and_Genetic_Data (2014).
78. Alex, B. et al. Text mining brain imaging reports. *J. Biomed. Semant.* **10**, 1–11 (2019).
79. Ford, E., Carroll, J. A., Smith, H. E., Scott, D. & Cassell, J. A. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J. Am. Med. Inform. Assoc.* **23**, 1007–1015 (2016).
80. Rannikmäe, K. et al. Developing automated methods for disease subtyping in UK Biobank: an exemplar study on stroke. *BMC Med. Inform. Decis. Mak.* **21**, 1–9 (2021).
81. Lovelace, J., Hurley, N. C., Haimovich, A. D. & Mortazavi, B. J. Dynamically extracting outcome-specific problem lists from clinical notes with guided multi-headed attention. In *Machine Learning for Healthcare Conference* 245–270 (PMLR, 2020).
82. Rannikmäe, K. et al. Accuracy of identifying incident stroke cases from linked health care data in UK Biobank. *Neurology* **95**, e697–e707 (2020).
83. Noor, K. et al. Deployment of a free-text analytics platform at a UK National Health Service Research Hospital: CogStack at University College London Hospitals. *JMIR Med. Inform.* **10**, e38122 (2022).
84. King's College Hospital NHS Foundation Trust. CogStack wins an artificial intelligence in health and care. https://www.kch.nhs.uk/news/public/news/view/34965 (2021).
85. Amazon Web Services. ICD-10-CM linking. https://docs.aws.amazon.com/comprehend-medical/latest/dev/ontology-icd10.html (2022).
86. Azure. What is text analytics for health in Azure Cognitive Service for Language? https://docs.microsoft.com/en-us/azure/cognitive-services/language-service/text-analytics-for-health/overview?tabs=ner (2022).
87. Google Cloud. Healthcare natural language API. https://cloud.google.com/healthcare-api/docs/concepts/nlp (2022).
88. Chapman WW, Haug PJ. Comparing expert systems for identifying chest x-ray reports that support pneumonia. Proc AMIA Symp 1999:216–20
89. Elkins JS, Friedman C, Boden-Albala B, et al. Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. Comput Biomed Res 2000;33:1–10
90. Chapman WW, Cooper GF, Hanbury P, et al. Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders. J Am Med Inform Assoc 2003;10:494–503
91. Kukafka R, Bales ME, Burkhardt A, et al. Human and automated coding of rehabilitation discharge summaries according to the International Classification of Functioning, Disability, and Health. J Am Med Inform Assoc 2006;13:508–15
92. Dinwoodie HP, Howell RW. Automatic disease coding: the 'fruit-machine' method in general practice. Br J Prev Soc Med 1973;27:59.

93. Warner HR., Jr Can natural language processing aid outpatient coders? J AHIMA 2000;71:78–81; quiz 83–74.
94. Morris WC, Heinze DT, Warner HR, Jr, et al. Assessing the accuracy of an automated coding system in emergency medicine. Proc AMIA Symp 2000:595–9
95. Lussier YA, Shagina L, Friedman C. Automating ICD-9-CM encoding using medical language processing: a feasibility study. J Am Med Inform Assoc 2000:1072–2
96. Goldstein I, Arzrumtsyan A, Uzuner O. Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. AMIA Annu Symp Proc 2007:279–83
97. Pakhomov SV, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. J Am Med Inform Assoc 2006;13:516–25
98. Barrows RC, Jr, Busuioc M, Friedman C. Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. Proc AMIA Symp 2000:51–5
99. Hersh W. Evaluation of biomedical text-mining systems: lessons learned from information retrieval. Brief Bioinform 2005;6:344–56
100. Turchin A, Kolatkar NS, Grant RW, et al. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. J Am Med Inform Assoc 2006;13:691–5
101. Edin, J., Junge, A., Havtorn, J. D., Borgholt, L., Maistro, M., Ruotsalo, T., & Maaløe, L. (2023, July). Automated medical coding on MIMIC-III and MIMIC-IV: a critical review and replicability study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2572-2582).

الترميز السريري والطبي: مسار جديد للأتمتة – مراجعة محدثة

الملخص

الخلفية :يعد الترميز السريري عملية أساسية في الرعاية الصحية، حيث يتم تحويل السجلات الطبية النصية الحرة إلى رموز منظمة باستخدام أنظمة تصنيف مثل **ICD-10.** يضمن هذا التحويل اتساق البيانات السريرية وإمكانية مقارنتها، مما يدعم تخطيط الرعاية الصحية، وصياغة السياسات، والبحوث الوبائية. ومع ذلك، فإن الترميز اليدوي يستغرق وقتًا طويلاً، وعرضة للأخطاء، ويتطلب خبرة كبيرة. ظهر الترميز السريري المؤتمت، الذي يعتمد على الذكاء الاصطناعي **(AI)** ومعالجة اللغة الطبيعية **(NLP)**، كحل واعد لتحسين الكفاءة والدقة. ورغم التطورات، لا تزال هناك تحديات تتعلق بمعالجة النصوص السريرية المعقدة، والتعامل مع الرموز النادرة، ودمج الاستدلال الرمزي.

الهدف :تهدف هذه المراجعة إلى استكشاف تطور الترميز السريري المؤتمت، وتقييم أداء أحدث نماذج التعلم العميق، وتحديد التحديات الرئيسية والاتجاهات المستقبلية لتحسين أنظمة الترميز المؤتمتة.

الأساليب :تستند هذه المراجعة إلى تحليل 113 دراسة حول الترميز السريري المؤتمت، مع التركيز على الانتقال من الذكاء الاصطناعي الرمزي القائم على القواعد إلى الذكاء العصبي، خاصة التعلم العميق. كما تناقش أداء نماذج التصنيف متعددة التصنيفات، ودمج النهج القائمة على المعرفة، والتحديات المرتبطة بمعالجة الوثائق الطويلة، والبيانات غير المتوازنة، وتغير المصطلحات الطبية. كما تسلط الضوء على أهمية التعلم بمشاركة الإنسان وقابلية تفسير الأنظمة المؤتمتة.

النتائج :حققت نماذج التعلم العميق، خاصة النماذج القائمة على المحولات مثل **BERT**، درجات **Micro-F1** تتراوح بين **58-60%** على مجموعات بيانات معيارية مثل **MIMIC-III.** ومع ذلك، لا تزال التحديات قائمة، مثل التعامل مع الرموز النادرة، ومعالجة الوثائق الطويلة، ودمج الاستدلال الرمزي. تظهر الأساليب الهجينة التي تجمع بين الذكاء الاصطناعي الرمزي والعصبي نتائج واعدة، كما هو الحال مع أساليب التعلم العميق المعززة بالمعرفة. تؤكد الدراسات أيضًا الحاجة إلى مجموعات بيانات عالية الجودة، وقابلية تفسير النماذج، والتكيف مع أنظمة الترميز الجديدة مثل **ICD-11.**

الخلاصة :أحرز الترميز السريري المؤتمت تقدمًا كبيرًا، ولكنه لا يزال مهمة معقدة تتطلب المزيد من البحث. تشمل الاتجاهات المستقبلية دمج الاستدلال الرمزي، وتحسين قابلية التفسير، وتطوير مجموعات بيانات أكثر تمثيلًا. يعد التعاون بين الباحثين في الذكاء الاصطناعي وخبراء الترميز السريري ضروريًا لدفع هذا المجال إلى الأمام.

الكلمات المفتاحية :الترميز السريري، الترميز المؤتمت، التعلم العميق، معالجة اللغة الطبيعية، **ICD-10**، الرسوم البيانية المعرفية، قابلية التفسير، الذكاء الاصطناعي الهجين.