
Proceedings Paper:
Jiang, X. orcid.org/0000-0003-4255-5445, Khan, K. orcid.org/0009-0008-0588-1974, Vasantha, S.T. orcid.org/0009-0001-1935-5552 et al. (1 more author) (2025) Evidence extraction for automated medical coding: preliminary evaluation. In: NLPIR '24: Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval. NLPIR 2024: 2024 8th International Conference on Natural Language Processing and Information Retrieval, 13-15 Dec 2024, Okayama University, Japan. Association for Computing Machinery (ACM) , pp. 18-23. ISBN 9798400717383

https://doi.org/10.1145/3711542.3711580

Evidence Extraction for Automated Medical Coding: Preliminary Evaluation

Xiaorui Jiang*
The University of Sheffield
Sheffield, United Kingdom
xiaorui.jiang@sheffield.ac.uk

Kulsoom Khan
Institute of Business Administration
Karachi, Pakistan
k.khan.25375@khi.iba.edu.pk

Sumithra Thinakaran Vasantha
Covenry University
Coventry, United Kingdom
thinakaras@uni.coventry.ac.uk

Sajjad Haider
Institute of Business Administration
Karachi, Pakistan
sahaider@iba.edu.pk

ABSTRACT

Coding clinical texts in standard language such as ICD is an important but tedious and error-prone process. Automated medical coding algorithms suffer problems due to the combined the challenge of handling the significant length of clinical text, the complexity of the huge code hierarchy and the lack of interpretability to ensure user trust. Large language models (LLM) have also been proven struggling with this task in recent studies. Recent efforts have been made to annotate an evidence-supported medical coding dataset. The current study makes the first empirical investigation into how well (small) fine-tuned pretrained language models (PLM) and LLMs could identify the sentences containing medical evidence supporting the assigned codes. Hierarchical sequential sentence classification and GPT-3.5 in the zero-shot setting were tested for evidence sentence extraction. Extra evaluation was performed to investigate how evidence extraction impacts clinical coding and what implications it has towards the future generation algorithms for automated medical coding.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Computing methodologies** → Information extraction.

KEYWORDS

ICD coding, code evidence, sequential sentence classification, large language model

ACM Reference Format:
Xiaorui Jiang, Kulsoom Khan, Sumithra Thinakaran Vasantha, and Sajjad Haider. 2018. Evidence Extraction for Automated Medical Coding: Preliminary Evaluation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

*Correesponding author

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Medical coding is an important task as it provides a summary of the diagnosis and treatment a patient receives in an interoperable language across different segments of the healthcare sector. However, coding is tedious, challenging and error-prone. For example, 1/5-1/3 general practice's time is spent on documentation, mainly clinical coding [31]. The call for AI to alleviate the manual overload on clinical coding is urgent. Automated medical coding has been studied for several decades, but the state of the art is still unsatisfactory to replace professional human coders [9]. Generally, it is very hard for computational methods to overcome the specific challenges of medical coding at the same time: (1) the scarcity of training data compared to an extreme size of the label set, (2) achieving high precision and recall for all ICD codes at the same time, and (3) lack of explainability that is the key to ensure user trust. While there are a vast number of approaches to tackling the first two challenges [15], research on explainable and trustful automated medical coding is comparatively limited. López-García et al. explicitly modelled coding evidence extraction as a medical entity recognition task and observed its positive impact as a "preprocessing" step for code prediction[25]. Inspired by their results the current study will put the emphasis on the evidence extraction task, hypothesizing that proper improvement on evidence extraction may lead to game-changing impact on automated medical coding.

Different from [25], the current study will investigate into the subtask of sentence-level code-relevant evidence extraction. It is motivated by the fourth challenge of medical coding – the sparsity of code-relevant information in long clinical texts [24]. Although pretrained language models (PLMs) have made significantly improvement for ICD coding (e.g., PLM-ICD [13]), most methods struggle to handle long clinical texts. The relative sparse coding-relevant information in a few sentences (e.g., on average 14 codes versus 1311 words per document in the MIMIC datasets) makes it extremely challenging to capture such nuanced signals. Meanwhile, although large language models (LLMs) have shown impressive performances in clinical information extraction [1] and biomedical question answering [29], they have been demonstrated inferior on medical coding [6, 12] and found unable to understand medical codes [21, 30]. Both observations motivated the current paper to investigate the potentials of PLM and LLM for identifying sentences describing information about disease diagnosis and treatment as supporting evidence for ICD code assignment.

2 LITERATURE REVIEW

2.1 Deep Learning and Pretrained Language Models

The ML-Net proposed in [10] used BiRNN (Bi-directional Recursive Neural Network) for document encoding and introduced a complimentary label count prediction task. However, the label count prediction component struggled with a large label set of hierarchical structures. The latter was handled by a tree-of-sequences LSTM (Long Short-Term Memory) component in [35] to encode the ICD code relationships embedded in code hierarchy. In real-world clinical coding datasets, most medical codes are rare, the number of codes is huge, and the prediction output space is extremely sparse. To deal with these challenges, the important label attention mechanism (LAAT) was proposed in [32] to transform document representation to a series of label-specific encodings. Further, JointLAAT was proposed to perform a hierarchical prediction by first predicting the block-level code (first three digits), which is further embedded and appended to document encodings for predicting the leaf code. Compared to LAAT, the RAC architecture (Read, Attend, and Code) in [18] encodes code descriptions and aligned them to the document embedding through the label-aware attention mechanism. Joint Attention Network (JAN) further extends to use bidirectional attention and fuse the document-aware label representations (labels attends to document) and the label-aware document representation (document to label) [34].

Recently, pretrained language models (PLM) have pushed automated medical coding to the new state of the art. Ji et al. found in-domain pretrained BERT models like SciBERT [5] improved prediction, and also proposed to use a hierarchical transformer layer over text chunks to handle lengthy clinical texts [14]. The PLM-ICD method in [13] adopted a simple idea to segment, encode and pool smaller chunks of a long clinical text, and achieved the then state of the art on the MIMIC-III full test set. XR-LAT in [24] replaced the BERT component in PLM-ICD with models that can handle longer input sequences and then integrated Label Attention (LAT) to align and predict the hierarchy of labels. However, the performance of deep learning-based medical coding is still suboptimal. Text length, sparsity of code-relevant information and extreme label size together made this task profoundly challenging.

2.2 Interpretable Medical Coding

In [4], attention weights were used to signify the implicitly identified diagnostic terms aligning to assigned codes. A follow-up line of research attempted to employ information text snippets in clinical texts to assist medical coding. In [11], keyphrases were automatically extracted to form a knowledge base for coding new cases and explaining code assignment. A similar idea was to find the most informative words to each code using Word Mover's Distance [20] and use them for code assignment and explanation [19]. Recently López-García et al. framed the explainable coding problem as a dual task of both evidence snippet extraction and clinical code assignment [25]. Surprisingly, the authors observed that, compared to the multi-task technique, the hierarchical approach seemed to be prominently more successful in reducing the intrinsic complexity by treating evidence extraction and code assignment independently.

Efforts also exist on creating resources for explainable ICD coding research. For example, the CodiEsp-X dataset on the CLEF 2020 international forum contains 1000 clinical cases and 16,504 sentences in Spanish, labelled with both ICD-10 codes and expert-annotated evidence text spans [27]. A similar effort was the Cantemist-Norm shared task at IberCLEF 2020, which contains ICD-O (Oncology) codes and evidence text spans for 1301 clinical oncology cases [26]. Recently, first and only MIMIC-based dataset for explainable ICD coding MDACE was introduced, short for MIMIC Documents Annotated with Code Evidence [8]. MDACE was built on a subset of the MIMIC-III discharge summaries, including 302 inpatient and 52 outpatient charts. Several caveats exist according to the authors. For example, it was noted that professional coders are trained to provide just sufficient evidence for each code so may under-annotate. Besides, the size of annotated evidence is relatively small compared to that of assignable codes.

2.3 Large Language Models for Medical Coding

Large language models (LLM) have shown excellent abilities without requiring supervisory or training data in many tasks such as reading comprehension, summarization and translation [2]. Trained on trillions of tokens, recent LLMs such as GPT-3/4 have demonstrated strong performance on medical question answering tasks [29]. However, LLMs struggle a lot to directly predict medical codes, such as ICD in alphanumeric format, from long clinical texts. In [30], mainstream LLMs including GPT-3.5, GPT-4, Gemini Pro and LLaMA-2 70b were prompted to generate code from the corresponding code description but the performance was unsatisfactory as the models repeated the same codes for different code descriptions. A similar evaluation was done in [21] with similar conclusions. In [12], GPT-3.5 alone was found inadequate for real-world practice as it has a tendency to either overpredict or under-predict codes. However, if code descriptions are presented within the prompt, GPT-3.5 can better predict ICD codes. Innovations must be made on how to elicit LLMs' clinical knowledge and reasoning capabilities towards a useful coding agent.

Witnessing similar observations as stated above, Boyle et al. proposed to frame the ICD coding problem "as an information retrieval task" to "retrieve 'mentions' of candidate codes" (descriptions of codes) from clinical texts. To scale the method to the extreme number of assignable codes, a hierarchical approach was employed to first predict the chapters and sections/blocks and then (sub)categories and leaf codes. A claimed benefit is its ability to handle rare codes, new codes or even complete ICD diagnostic ontology revision. Similarly, Yang et al. observed that GPT-4 often predicts an excessive number of ICD codes [36]. To tackle this, they proposed a two-stage approach to first predict a code together with an evidence sentence and then use a small language model, such as LSTM, to verify the predicted codes using the collection of LLM-extracted evidence sentences, which gained significant performance improvement on few-shot and rare codes. The idea of verifier was extended in [22] into a comprehensive approach called the Multi-Agent Coding (MAC). For example, the MAC-1 method let an LLM to play different roles, as the coder to generates code and evidence sentence, as reviewer to adjust the initially assigned codes, as patient and physician to review the revised codes, and as

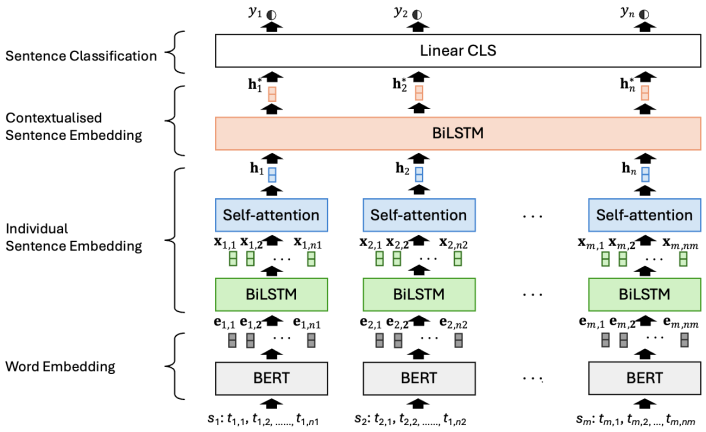

Figure 1: Hierarchical sequential sentence classification.

What are the ICD-10 codes and descriptions for the given clinical text?
Clinical text: {clinical_text}
Answer format: code: {{}}, description: {{}}

(a) Evidence-supported prompt.

What are the ICD-10 codes and descriptions for the given clinical text?
Clinical text: {clinical_text}
Answer format: code: {{}}, description: {{}}

(b) A simpler code-only prompt.

Figure 2: ICD-10 coding prompts for GPT-3.5.

adjustor to verdict on disputes. Essentially, MAC-1 applied the LLM itself for self-verification four times after initial code assignment.

3 DATASET

CodiEsp-X and Cantemist-Norm (see Sect 2.2) are originally in Spanish, though English machine translations are provided. CodieEsp-X contains 1000 case reports, on average, 411.1 words and 18.4 codes per report, and has a 3.2 times higher code density than MIMIC datasets. Cantemist-Norm is only about ICD for oncology. Therefore, the current study used MDACE [8]. It is based on a portion of the most widely used MIMIC-III dataset [17]. In total, 302 Inpatient and 52 Profee charts were annotated with 3,934 and 5,563 evidence spans respectively. It is a proper testbed as its documents are much longer than MIMIC-III's average, 19,372 and 11,116 words for Inpatient and Profee respectively.

The MDACE dataset was further processed by linking the discharge summaries to the annotation files in MDACE, segmenting each discharge summary into a sequence of sentences and adding sentence-level labels according to MDACE annotations. If a sentence has at least one annotated evidence span in MDACE, then a YES label is assigned to the corresponding sentence. The resulting dataset, coined MDACE-Sent, contains in total 74,333 sentences, with 6,192 positive (evidence) sentences and 68,141 negative (non-evidence) sentences. On average, each clinical text in MDACE has 68.36 sentences per document. The imbalance ratio is as high as about 1:11, making it a quite challenging task.

4 HIERARCHICAL DOCUMENT MODELLING FOR EVIDENCE SENTENCE CLASSIFICATION

To capture sentence meanings in the context, we chose to implement a hierarchical sentence modelling architecture similar to [16] (Figure 1). Suppose each clinical text is represented by a sequence of sentences $s_1, \ldots, s_m$, where m is the length of the text. Each sentence s_i is represented as a sequence of tokens $T_i = \{t_{i,1}, t_{i,2}, \ldots, t_{i,ni}\}$, where $t_{i,j}$ is the vocabulary index of the j-th token in the i-th sentence and ni is the length of the i-th sentence. The architecture contains four layers. The Word Embedding layer uses a domain-specific BERT model to map the all the tokens in each sentence s_i into a sequence of initial feature vectors, the word embeddings vectors $\mathbf{E}_i = \{\mathbf{e}_{i,1}, \mathbf{e}_{i,2}, \ldots, \mathbf{e}_{i,ni}\}$, where $E_i = \text{BERT}(T_i)$.

The initial word embeddings of each sentence are then summarised by the Sentence Encoding layer into the local meaning representation of each sentence. Several possible ways exist. For BERT-family models, the "[SEP]" symbol can be used for sentence representation. Max-pooing, mean-pooing or self-attention can be used. Here, to generate better sentence representations, the Sentence Encoding layer introduces a one-layer BiLSTM to further encode the tokens of each sentence s_i into the word representations $X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,ni}\}$, where $x_{i,j} = \text{BiLSTM}(\mathbf{e}_{i,j})$. Then the multi-headed self-attention component [7, 16] will summarise the word representations into a sentence representation vector $\mathbf{h}_i$. To account for the context in which sentences occur to encode sentence meanings, the Contextualised Sentence Encoding layer uses another one-layer BiLSTM (denoted as BiLSTM^c) and updates sentence representations: $\mathbf{H}_i^* = \{\mathbf{h}_1^*, \mathbf{h}_2^*, \ldots, \mathbf{h}_m^*\}$, $where \mathbf{h}_i^* = \text{BiLSTM}^c(\mathbf{h}_i)$.

Finally, the Sentence Classification layer is simply a linear (i.e., a fully connected) layer which consumes the updated sentence representations for prediction. Optionally, a conditional random field (CRF) can be added above the linear layer to better capture the inter-sentential relationships. But it was omitted in our experiments as most of the time, evidence for a code happens in just one sentence, so the labels for the positive class are unlikely correlated with each other.

5 PROMPTING LARGE LANGUAGE MODEL FOR EVIDENCE SENTENCE EXTRACTION

The current paper mainly tested the zero-shot reasoning capability of GPT-3.5. In addition to evidence extraction, GPT-3.5 was also prompted to generate the ICD-10 codes based on the identified evidence. Figure 2a shows the prompt. The main purpose was to instruct GPT-3.5 to identify which sentences contain medical evidence to support code assignment. To enhance GPT-3.5's confidence, it was also prompted to generate the description of the predicted code and the code itself from the identified evidence. On one hand, the predicted code and code description serve as

Table 1: Examples of medical evidence extraction for ICD coding by GPT-3.5.

Ex. 1:	"Text": ["He also had signed a DNR/DNI order."], "Predicted Response": "Yes, Text evidence: DNR/DNI order, Code: Z66
Ex. 2:	"Text": ["He was clear that he did not want to be dependent of disabled."], "Response": "No, Text evidence: null, Code: null"
Ex. 3:	"Text": ["Past Medical History:- DIABETES TYPE II- HYPERLIPIDEMIA- GLAUCOMA- OSTEOARTHRIPS- CAROTID STENOSIS left 60-69%, rt 50- VASOVAGAL SYNCOPE- BACK PAIN"], "Response": "Yes, Text evidence: Past Medical History:- DIABETES TYPE II- HYPERLIPIDEMIA- GLAUCOMA- OSTEOARTHRIPS- CAROTID STENOSIS left 60-69%, rt 50- VASOVAGAL SYNCOPE- BACK PAIN, Code: E11.9, E78.5, H40.9, M15.9, I65.29, R55, M54.5"

the explanation for evidence identification. On the other hand, evidence identification elicits the reasoning capability of GPT-3.5. Table 1 shows three coding examples: Ex. 1 about a correct code and its evidence, Ex. 2 without code and evidence, and Ex. 3 with some incorrect (*italic underlined*) and overpredicted or hallucinated codes (underlined). In Ex. 3, the correct code for the text evidence "OSTEOARTHRIPS" should be M19.90 (Unspecified osteoarthritis, unspecified site), rather than the output M15.9 (Polyosteoarthritis, unspecified). The code I65.29 ("Occlusion and stenosis of unspecified carotid artery") predicted by GPT-3.5 for "CAROTID STENOSIS left 60-69%, rt 50" is imprecise according to professional coders who assigned I65.23 (Occlusion and stenosis of bilateral carotid arteries). While MDACE annotators did not assign a code to the text "VASOVAGAL SYNCOPE", the code R55 (Syncope and Collapse) predicted by GPT-3.5 seems to be related.

To investigate the impact of evidence extraction has on the performance of ICD coding, we also tested a simpler setting by directly instructing GPT-3.5 to generate the ICD-10 codes, which is essentially equivalent to [12] and [21]. Figure 2b shows the simple prompt demanding GPT-3.5 to output all ICD codes it detects from the clinical text (the CLINICAL_TEXT field). Although LLMs seem to be unable to understand medical codes [30], the current study holds the belief similar to [12] that LLMs will be much smarter in generating ICD codes if code definitions are provided in the prompt. Different from [12], we prompted GPT-3.5 to generate the descriptions of the codes that match the evidence it finds. This way was proved much better than GPT-3.5 predicting codes in a "blind eyed" way.

6 EXPERIMENTAL RESULTS

6.1 Sequential Sentence Classification for Evidence Extraction

Five BERT variants were experimented, including SciBERT [5], ClinicalBERT (trained on a multi-centre EHR dataset "with a large corpus of 1.2 billion words of diverse diseases") [33], ClinicalBERT-MIMIC (trained on MIMIC-III Dataset [3], BioBERT [18], MedBERT [28]. A five-fold cross validation over the MDACE documents were done and the average performance of each model was reported. The training settings are as follows. BiLSTM's hidden size: same as the BERT variants. Self-attention dimension: 200. Number of attention heads: 15 as in [7]. Loss Function: Due to the huge imbalance focal loss was adopted [23], which was known to be more capable of handling hard cases, often samples from the minority class. Batch Size: each document was considered as a batch. Optimiser: Adam. Learning rate epoch decay: 0.9. Learning Rate: 3e-5. Dropout rate: 0.5. Epochs: a maximum of 30 epochs were trained and the best models were selected on a separate validation set (20% in the training split) according to the F1 score of the positive class (evidence sentence).

Table 2: The Performance of hierarchical sequential sentence classification on evidence extraction.

	Evidence Sentence			Non-evidence Sentence			Auc
	Prec.	Rec.	F1	Prec.	Rec.	F1	
SciBERT	**0.594**	0.652	0.572	0.964	0.937	0.951	0.795
ClinicalBERT	0.539	0.602	0.569	0.960	**0.949**	0.951	0.775
ClinicalBERT-MIMIC	0.545	0.543	0.544	0.954	0.955	**0.955**	0.749
BioBERT	0.535	**0.695**	0.604	**0.969**	0.940	0.954	**0.818**
MedBERT	0.487	0.671	0.564	0.966	0.930	0.948	0.801

Table 2 shows the results. The best recall for the positive class was slightly below 70% on the MDACE dataset (by BioBERT), but the precision was only about 53.5%. Overall, BioBERT was the best performing model. In an anticipated pipelined approach for ICD coding, the purpose of evidence extraction is to significantly reduce coding-irrelevant information while retaining most coding-relevant information, so that the standard deep learning-based methods such as PLM-ICD can achieve better performance on an "information-dense and succinct summary of medical evidence". In this sense, a precision around 50% or higher is acceptable, but the recall of positive class must be significantly increased. If we look at the negative class, however, the precision, recall and F1 values of all models are quite high. There is high potential to adjust these models for higher precision and use them to rather safely screen out the non-evidence sentences. Another potential direction is to build an ensemble of non-evidence sentence filter.

6.2 Zero-Shot Prompting of GPT-3.5 for Evidence Extraction

The GPT-3.5 model "gpt-3.5-turbo-0125" was used to generate responses under the default setting (through Azure AI API). Table 3 shows that while GPT-3.5 was able to identify about 66% evidence sentences, its precision was only 24.8%, much lower compared to finetuned PLMs shown in Sect. 6.1. Out result shows that the number of evidence sentences identified by GPT-3.5 was three times bigger than the size of real evidence sentences. Though our results demonstrated the difficulty of the medical evidence extraction subtask, there are still many opportunities. The most obvious is that the discrepancies between the results of PLMs and LLMs (GPT-3.5

Table 3: Comparative performance of GPT-3.5 on evidence extraction.

	Evidence Sentence		
	Prec.	Rec.	F1
GPT-3.5	0.248	0.659	0.361
SciBERT	**0.594**	0.652	0.572
ClinicalBERT	0.539	0.602	0.569
ClinicalBERT-MIMIC	0.545	0.543	0.544
BioBERT	0.535	**0.695**	**0.604**
MedBERT	0.487	0.671	0.564

in the current study) implied that both methods could be synergized. For example, to improve the recall, it is possible to merge the results of the methods of both categories, followed by a potential verifier, either a finetuned PLM or an LLM. Another potential way is to inject knowledge about the ICD ontology into the prompt and perform top-down decision making akin to [6]. The third potential direction is to let an LLM to verify its result by elicit more in-depth reasoning over the evidence it identified, the description of the code it assigned, its reasoning chain towards these answers etc., or to let several LLMs collaborate and debate with each other [22]. Finally, considering the large sizes for both clinical texts and assignable labels, the "information retrieval" point of view and the "retrieve and rank" paradigm sound reasonable directions for further research.

6.3 Zero-Shot Prompting of GPT-3.5 for ICD-10 Coding

Though not satisfactory, the empirical results still showed value of this non-trivial task of code-relevant evidence extraction, which we meant to promote for more research. To demonstrate this, two additional experiments were done to evaluate **how well GPT-3.5 understands evidence text and predicts codes from real evidence**. In the first experiment, GPT-3.5 was prompted with the gold evidence text to predict the code and code description (Figure 2b). We checked whether the block-level code (first three characters, in the form of XXX) matches the gold label. In the second experiment, the results in Sect. 6.2were analysed and prediction performance was evaluated by checking the block-level code (first three characters), the category-level code (plus one character after dot, in the form of XXX.Y), and leaf code (i.e., exact match). Figure 3 compares the results. The results again proved that prompting an LLM to predict the code-relevant evidence and reason about medical evidence in clinical texts does have positive impact on the accuracy. The significant performance drop from section-level code prediction to category-level code prediction may imply the boundary of knowledge or reasoning capability of GPT-3.5. However, it also implies the potential of synergizing LLMs (e.g., for block-level prediction) and smaller LMs tailored to subfields of medicine, which may be a promising rescue for LLM for automated medical coding.

7 CONCLUSIONS

The current study made an initial investigation into whether supporting evidence for ICD coding can be extracted by fine-tuned pretrained language models (PLM) and large language models (LLM)

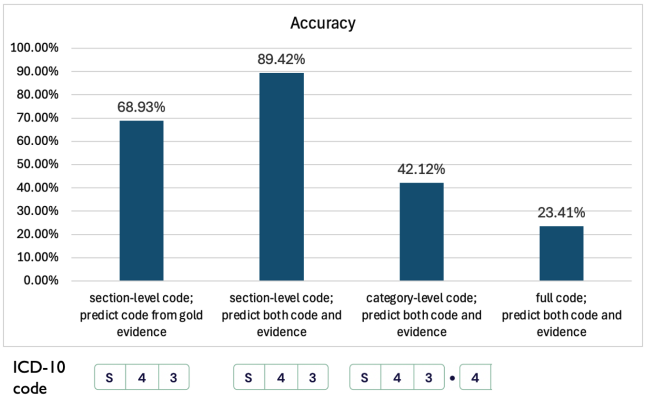

Figure 3: Performance of code prediction by GPT-3.5.

such as GPT-3.5. Empirically proved, this is a difficult task. Our best PLM recognised about 69.3% evidence on MDACE, which is still not adequate to for developing a robust codes assignment algorithm. GPT-3.5 achieved a similar recall but the precision was only less than half of that of the PLMs. In addition, the current study also makes the first investigation into the capability of GPT-3.5 in medical coding based on evidence identification. The results also corroborate with most recent studies in that the task is beyond LLM's zero-shot capability, if no complex mechanism was applied to elicit LLM's reasoning capability. There are also several positive sides. All PLMs' AUCs were not bad, implying the opportunity for tuning the trade-off between precision and recall. All PLMs seem to be much better in filtering out non-evidence sentences, so an ensemble might be a remedy solution for automated medical coding based on evidence extraction, and a good trade-off for recall may hopefully be achievable. Given extracted evidence, GPT-3.5 performed well in predicting the block part of ICD-10 codes, which signify the high potential of hierarchical coding. Finally, LLMs and (smaller) PLMs can also collaborate and correct each other.

REFERENCES

[1] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP '22)*. Association for Computational Linguistics, Kerrville, TX, USA, 1998–2022. https://doi.org/10.18653/v1/2022.emnlp-main.130

[2] Alec Radford and Jeffrey Wu and Rewon Child and David Luan and Dario Amodei and Ilya Sutskever 2019. *Language Models are Unsupervised Multitask Learners.* Retrieved Feb 16, 2025 from https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[3] Kexin Huang Jaan Altosaar and Rajesh Ranganath. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. In *Proceedings of the Workshop of the 2020 ACM Conference on Health, Inference, and Learning (CHIL '20 Workshop).* https://arxiv.org/abs/1904.05342

[4] Aitziber Atutxa, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz, and Olatz Perez de Viñaspre. 2019. Interpretable deep learning to map diagnostic texts to ICD-10 codes. *Int. J. Med. Inform.* 129 (Sept. 2019), 49–59. https://doi.org/10.1016/j.ijmedinf.2019.05.015

[5] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP '19)*. Association for Computational Linguistics, Kerrville, TX, USA, 3615–3620. https://doi.org/10.18653/v1/D19-1371

[6] Joseph Boyle, Antanas Kascenas, Pat Lok, Maria Liakata, and Alison O'Neil. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the Workshop on Deep Generative Models for Health at the the 36th International*

Conference on Neural Information Processing Systems (DG4H '23 @ NeurIPS '23). https://openreview.net/forum?id=mqnR8rGWkn

[7] Arthur Brack, Anett Hoppe, Pascal Buschermöhle, and Ralph Ewerth. 2022. Cross-domain multi-task learning for sequential sentence classification in research papers. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries (JCDL '22)*. Association for Computational Machinery, New York, NY, USA, Article No. 34, 13 pages. https://doi.org/10.1145/3529372.3530922

[8] Hua Cheng, Rana Jafari, April Russell, Russell Klopfer, Edmond Lu, Benjamin Striner, and Matthew Gormley. 2023. MDACE: MIMIC Documents Annotated with Code Evidence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL '23)*. Association for Computational Linguistics, Kerrville, TX, USA, 7534–7550. https://doi.org/10.18653/v1/2023.acl-long.416

[9] Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. 2022. Automated clinical coding: what, why, and where we are? *npj Digit. Med.* 5, Article number 159 (Oct. 2022), 8 pages. https://doi.org/10.1038/s41746-022-00705-7

[10] Jingcheng Du, Qingyu Chen, Yifan Peng, Yang Xiang, Cui Tao, and Zhiyong Lu. 2019. ML-Net: multi-label classification of biomedical texts with deep neural networks. *J. Am. Med. Inform. Assn.* 26, 11 (Nov. 2019), 1279–1285. https://doi.org/10.1093/jamia/ocz085

[11] Andres Duque, Hermenegildo Fabregat, Lourdes Araujo, and Juan Martinez-Romo. 2021. A keyphrase-based approach for interpretable ICD-10 code classification of Spanish medical reports. *Artif. Intell. Med.* 121, Article 102177 (Nov. 2021). https://doi.org/10.1016/j.artmed.2021.102177

[12] Matúš Falis, Aryo Pradipta Gema, Hang Dong, Luke Daines, Siddharth Basetti, Michael Holder, Rose S Penfold, Alexandra Birch, and Beatrice Alex. 2024. Can GPT-3.5 generate and code discharge summaries? *J. Am. Med. Inform. Assn.* 31, 10 (Nov. 2024), 2284–2293. https://doi.org/10.1093/jamia/ocae132

[13] Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. PLM-ICD: Automatic ICD Coding with Pretrained Language Models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop (ClinicalNLP '22)*. Association for Computational Linguistics, Kerrville, TX, USA, 10–20. https://doi.org/10.18653/v1/2022.clinicalnlp-1.2

[14] Shaoxiong Ji, Matti Hölttä, and Pekka Marttinen. 2021. Does the magic of BERT apply to medical code assignment? A quantitative study. *Comput. Biol. Med.* 139, Article 104998 (Dec. 2021), 7 pages. https://doi.org/10.1016/j.compbiomed.2021.104998

[15] Shaoxiong Ji, Xiaobo Li, Wei Sun, Hang Dong, Ara Taalas, Yijia Zhang, Honghan Wu, Esa Pitkänen, and Pekka Marttinen. 2024. A Unified Review of Deep Learning for Automated Medical Coding. *ACM Comput. Sur.* 56, 2, Article No. 36 (Oct. 2024), 41 pages. https://doi.org/10.1145/3664615

[16] Di Jin and Peter Szolovits. 2018. Hierarchical Neural Networks for Sequential Sentence Classification in Medical Scientific Abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP '18)*. Association for Computational Linguistics, Kerrville, TX, USA, 3100–3109. https://doi.org/10.18653/v1/d18-1349

[17] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2015. MIMIC-III, a freely accessible critical care database. *Sci. Data* 3, Article number 160035 (May 2015), 8 pages. https://doi.org/10.1038/sdata.2016.35

[18] Byung-Hak Kim and Varun Ganapathi. 2021. Read, Attend, and Code: Pushing the Limits of Medical Codes Prediction from Clinical Notes by Machines. In *Proceedings of the 6th Machine Learning for Healthcare Conference (ML4H '21)*. PMLR, 196–208. https://proceedings.mlr.press/v149/kim21a.html

[19] Amit Kumar, Suman Roy, and Sourabh Bhattacharjee. 2022. A fast unsupervised assignment of ICD codes with clinical notes through explanations. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing (SAC '22)*. Association for Computational Machinery, New York, NY, USA, 610–618. https://doi.org/10.1145/3477314.3506983

[20] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q.Weinberger. 2015. From Word Embeddings to Document Distances. In *Proceedings of the 32nd International Conference on Machine Learning (ICML '15)*. PMLR, 957–966. https://proceedings.mlr.press/v37/kusnerb15.html

[21] Simon A. Lee and Timothy Lindsey. 2024. Can Large Language Models abstract Medical Coded Language. arXiv:2403.10822 [cs.CL]

[22] Rumeng Li, Xun Wang, and Hong Yu. 2024. Exploring LLM Multi-Agents for ICD Coding. arXiv:2406.15363 [cs.CL]

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal Loss for Dense Object Detection. *IEEE T. Pattern Anal.* 42, 2 (Feb. 2018), 318–327. https://doi.org/10.1109/tpami.2018.2858826

[24] Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. 2023. Automated ICD coding using extreme multi-label long text transformer-based models. *Artif. Intell. Med.* 144, Article 104323 (Oct. 2023), 10 pages. https://doi.org/10.1016/j.jbi.2023.104323

[25] Guillermo López-García, José M. Jerez, Nuria Ribelles, Emilio Alba, and Francisco J. Veredas. 2023. Explainable clinical coding with in-domain adapted transformers. *J. Biomed. Inform.* 139, Article 102662 (March 2023), 10 pages. https://doi.org/10.1016/j.jbi.2023.104323

[26] Antonio Miranda-Escalada, Eulàlia Farréa, and Martin Krallinger. 2020. Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF '20)*. CEUR, 303–323. https://ceur-ws.org/Vol-2664/capitel_overview.pdf

[27] Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. 2020. Overview of automatic clinical coding: annotations, guidelines, and solutions for non-English clinical cases at CodiEsp track of eHealth CLEF 2020. In *Proceedings of the CLEF eHealth Evaluation Lab 2020 (CLEF '20)*. CEUR. https://ceur-ws.org/Vol-2664/capitel_overview.pdf

[28] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.* 4, Article 86 (May 2021). https://doi.org/10.1038/s41746-021-00455-y

[29] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature* 620 (Aug. 2023), 172–180. https://doi.org/10.1038/s41586-023-06291-2

[30] Ali Soroush, Benjamin S. Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W. Charney, Girish N Nadkarni, and Eyal Klang. 2024. Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying. *NEJM AI* 1, 5 (April 2024). https://doi.org/10.1056/AIdbp2300040

[31] Kaushik P. Venkatesh, Marium M. Raza, and Joseph C. Kvedar. 2023. Automating the overburdened clinical coding system: challenges and next steps. *npj Digit. Med.* 6, Article number 16 (Feb. 2023), 2 pages. https://doi.org/10.1038/s41746-023-00768-0

[32] Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for ICD coding from clinical text. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence (IJCAI '20)*. Article 461, 3335–3341 pages. https://doi.org/10.24963/ijcai.2020/461

[33] Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, Kanmin Xue, Xiaoying Li, and Ying Chen. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nat. Med.* 29 (Sept. 2023), 2633–2642. https://doi.org/10.1038/s41591-023-02552-9

[34] Yuzhou Wu, Zhigang Chen, Xin Yao, Xuechen Chen, Zeren Zhou, and Jinkai Xue. 2022. JAN: Joint Attention Networks for Automatic ICD Coding. *IEEE J. Biomed. Health* 26, 10 (July 2022), 5235–5246. https://doi.org/10.1109/JBHI.2022.3189404

[35] Pengtao Xie and Eric Xing. 2018. A Neural Architecture for Automated ICD Coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL '18)*. Association for Computational Machinery, New York, NY, USA, 1066–1076. https://doi.org/10.18653/v1/P18-1098

[36] Zhichao Yang, Sanjit Singh Batra, Joel Stremmel, and Eran Halperin. 2023. Surpassing GPT-4 Medical Coding with a Two-Stage Approach. In *Proceedings of of the 2023 Machine Learning for Health symposium (ML4H '23)*. 19 pages. https://arxiv.org/abs/2311.13735