Peer Review File

Article information: https://dx.doi.org/10.21037/jmai-23-106

Reviewer A

In this paper, Ong et al conducted a study on the efficacy of ChatGPT on providing retinal ICD coding. This is an interesting approach to a practical clinical problem, and I congratulate the authors for their work.

I consider this to be a pilot study which needs further validation on larger samples with real clinical notes across various centers and languages. However, these pilot studies are crucial to understand the potential before planning larger studies. Therefore, the work by Ong et al constitutes an important initiative.

I only have one minor point. Using chatgpt for diagnostic coding have a range of ethical and academical implications (e.g. Registry-based studies). It would improve the paper further, if potential ethical and academical challenges in the application of Chatgpt for diagnostic coding could be discussed

Response: Thank you for the kind and expert review of our work. We absolutely agree with the reviewer that it is critical for further validation, and we appreciate the reviewer for their expert perspective on this initiative. We also agree and appreciate the perspective of the reviewer that there are ethical and academic challenges in diagnostic coding when utilizing an LLM. We have added this in the paper with corresponding word changes.

Lines 245-256: There are also potential ethical and academic implications for the deployment of ChatGPT for diagnostic coding. Medical billing is a process that requires utmost professionalism and integrity from clinicians. Ensuring consistency and accuracy of ICD coding for matching clinical documentation encounters serves to fair reimbursement for providers. Mismatches in ICD coding may result in dilemmas for various individuals including the clinician, patient, insurance, and stakeholders; thus, it is important to note that the utilization of LLMs may be utilized to assist in the efficiency of a medical practice, but the clinician should review and finalize the final selection provided. Another point of discussion is that this utilization of LLMs for ICD coding may generate further discussions in the academic research and quality-improvement setting such as registry-based studies. These studies may be impacted under full-autonomous ICD coding, thus, highlighting the importance of the clinician to ensure that the ICD codes provided by the LLM match the clinical encounter.

Reviewer B

The authors present a thoughtful study to show the potential for LLMs to assist with ICD10 diagnoses.

ChatGPT is a general-purpose language model and will perform better in many different ways at this task. Simply 4.0 would likely perform better.

Using prompt engineering can give more specific responses and guide GPT to give better responses. I would love to see an example of prompts used to see if responses can be improved. Decreasing temperature or using a web search copilot (perplexity) - or creating one would drastically decrease hallucination rates. Also continuing in the same chat retains context and can influence future responses. Ideally new chats are created each time.

This is a great exploratory study but a lot is already known on how to improve the output. I would like to see this study incorporate some of the latest research in LLMs and not just be a direct usage of ChatGPT.

Response: We thank the reviewer for their kind review and their expert perspective. The reviewer provides timely perspective. We provide a discussion that expands upon broader artificial intelligence in ICD coding. We discuss deep neural networks for ICD coding and the fundamental general architectures and layers that have described in the literature. We also mention different iterations of convolutional neural networks that can be applied to these techniques and conclude that these established deep learning techniques for ICD coding may be applied to the emergence of LLM artificial intelligence. We also thank the reviewer for their very timely review to discuss advances in LLMs. On September 25, 2023, OpenAI released new capabilities for their LLM that can analyze both image and voice. Such technology can be employed in a variety of aspects in the clinical aspect, including ICD coding. We have cited and discussed this now in the paper. We also mention that future research, alongside cybersecurity research, may be utilized to further optimize this technology. We thank the reviewer for their expert and timely insight.

Lines: 281-310: ICD coding has also been an interest in other forms of artificial intelligence. Teng et al. discuss the application of deep neural networks for ICD coding (17). As ICD coding is vulnerable to human error, there has been prior research with machine learning and deep learning techniques. One of the large challenges noted in ICD coding includes distinct writing styles, non-relevant information for coding, and long documents. These variables represent challenges in consistent ICD coding, however, Teng et al. describe a general deep neural network architecture for ICD coding that employs an input layer, representation layer, feature layer, and output layer. The input layer employs multi-source data input, including external knowledge, such as the ICD-10 taxonomy, free-input that comes from the health records, and the code relationship. The technical aspects of the representation and feature are outside the scope of this discussion, however, the feature layers employ convolutional neural networks (CNNs) to extract various features from the data that are critical for output layer generation. Lastly, the output layer generates the ICD code as well as takes into account the loss function with subsequent back-propagation for further optimization.

Teng et al. discusses various CNNs and their iterations including attention-convolution and dilated convolution that can be applied to ICD coding. Ultimately, these advances in artificial intelligence applied to ICD coding may be merged with advances seen here in LLM technology to construct more accurate and efficient models.

Research in other aspects of the LLM optimization have been explored including being able to visualize/analyze images and engaging with the technology with voice. In September 2023, OpenAI announced the ability of ChatGPT to analyze voice and image input (18). This optimization of LLM input may further optimize ICD coding. Along with further validation in cybersecurity and privacy, this technique may have the potential to take in voice-based clinical encounter discussions to generate ICD codes. This would further optimize clinic workflow compared to the methods of the aforementioned study as no text input would be required. Additionally, clinicians may not document their final encounter note immediately after the clinical encounter, thus, ICD code generation with LLMs must be performed after this manual step. By employing voice-based LLMs, ICD code selections may be available immediately after the patient encounter has finished. Future research may be geared towards these optimizations in LLM technology in ICD coding as well as the cybersecurity research that must go behind these technologies to ensure safe clinical implementation.

Reviewer C

This paper investigated the performance of ChatGPT for retina ICD coding, which is interesting. The methods results are described in detail and some interesting observations were provided. I have a few comments.

Overall Response: We thank the reviewer for their expert and kind review of our work. We appreciate the comments and insights to optimize the manuscript. We are especially thankful for their expertise on providing further insights with deep learning. We have answered each comment with corresponding lines.

1. Although it is new to using LLM for ICD coding, there are a rich body of research on automated ICD coding using machine learning and deep learning methods, e.g., Teng, Fei, et al. "A review on deep neural networks for ICD coding." IEEE Transactions on Knowledge and Data Engineering 35.5 (2022): 4357-4375. It would be useful to include some discussion on these methods and how the proposed method differs from existing studies, to provide a comprehensive view of landscape in this field.

Response: Thank you for comment and very helpful insight and suggestion. We have now added a discussion about this. We highly appreciate the author for their expert insight and enriching this discussion. We have added the discussion below with corresponding lines.

Lines 281-297: ICD coding has also been an interest in other forms of artificial intelligence. Teng et al. discuss the application of deep neural networks for ICD coding

(17). As ICD coding is vulnerable to human error, there has been prior research with machine learning and deep learning techniques. One of the large challenges noted in ICD coding includes distinct writing styles, non-relevant information for coding, and long documents. These variables represent challenges in consistent ICD coding, however, Teng et al. describe a general deep neural network architecture for ICD coding that employs an input layer, representation layer, feature layer, and output layer. The input layer employs multi-source data input, including external knowledge, such as the ICD-10 taxonomy, free-input that comes from the health records, and the code relationship. The technical aspects of the representation and feature are outside the scope of this discussion, however, the feature layers employ convolutional neural networks (CNNs) to extract various features from the data that are critical for output layer generation. Lastly, the output layer generates the ICD code as well as takes into account the loss function with subsequent back-propagation for further optimization. Teng et al. discusses various CNNs and their iterations including attention-convolution and dilated convolution that can be applied to ICD coding. Ultimately, these advances in artificial intelligence applied to ICD coding may be merged with advances seen here in LLM technology to construct more accurate and efficient models.

2. As the paper focused on LLM, there is no baseline methods in comparison. However, the authors have highlighted that prompt and feedback fine-tuning LLM can potentially improve the performance. I think it would be useful to explore this to inform future development.

Response: Thank you for the comment. We agree with the reviewer that future research in this area with feedback fine-tuning would help to stratify and compare performance of LLMs. We have added a dedicated sentence to the manuscript to discuss this future research direction.

Lines 277-282:
Furthermore, providing the ICD codes with their definitions in the original prompt to ChatGPT may have greatly improved performance since then the model would not rely on its original training, which includes texts from across the internet, but instead focused on the definitions of ICD codes given to it in the prompt. Future research may be utilized to analyzing prompt engineering and feedback fine-tuning of LLMs in ICD coding against LLMs without feedback fine-tuning to stratify and compare the benefits of improving.

3. It might be clearer if a table or figure can be used to present the results. Also, I think the interesting error analysis in the discussion can be a separate subsection in the result section. This way, the error analysis can be richer.

Response: Thank you for the comment. We thank the reviewer for their perspective and suggestion. We have now added a table that concisely shows the results.

Lines 182-192 and Lines 420-426: A total 181 mockup retina encounters were evaluated. 84 eyes were right eyes, 97 eyes were left eyes. A total of 597 ICD codes were generated, with 305 consisting of retina codes (51% of total consisting of retina codes, 1.68 retina codes per eye). This total code count also included past medical history included in the note including hyperlipidemia, hypertension, and hyperthyroidism. 127/181 (70%) of responses resulted in a true positive result with at least one code provided matching a correct code. 54/181 (30%) responses did not generate a correct code from the text. An additional sub-analysis analyzed whether ChatGPT coded the retina encounter and diagnosis completely correct. If ChatGPT got any of it incorrect, it would be counted as incorrect in this analysis even if some were correct. In this analysis ChatGPT achieved a "Correct Only" in 106 of 181 encounters (59%) with the remaining 75 encounters (41%) having some form of incorrect diagnosis even if it included the correct diagnosis (Table 1.).

Subgroup	Results	Percentage
Correct	137/181	70%
Correct Only	106/181	59%
Incorrect	54/181	30%

Table 1. Results of ChatGPT's generation of ICD coding by "Correct", "Correct only", and "Incorrect" for mockup retina encounters. "Correct" was defined by producing at least one correct ICD code for a clinician to choose from. "Correct Only" was defined as generating only the correct ICD codes for the encounter. "Incorrect" was defined as not generating any correct ICD codes for the mockup retina encounter.

4. Is there any privacy concern on send sensitive patient data to ChatGPT?

Response: Thank you for the comment. We highly appreciate the reviewer's perspective on this matter. There are certainly privacy concerns with sensitive patient data when utilizing online LLMs. We have dedicated a paragraph that discusses future research including cybersecurity and validation of this cybersecurity research in order for this to be implemented clinically. We thank the reviewer again for the comment.

Lines 300-311: Research in other aspects of the LLM optimization have been explored including being able to visualize/analyze images and engaging with the technology with voice. In September 2023, OpenAI announced the ability of ChatGPT to analyze voice and image input (18). This optimization of LLM input may further optimize ICD coding. Along with further validation in cybersecurity and privacy, this technique may have the potential to take in voice-based clinical encounter discussions to generate ICD codes. This would further optimize clinic workflow compared to the methods of the aforementioned study as no text input would be required. Additionally, clinicians may not document their final encounter note immediately after the clinical encounter, thus,

ICD code generation with LLMs must be performed after this manual step. By employing voice-based LLMs, ICD code selections may be available immediately after the patient encounter has finished. Future research may be geared towards these optimizations in LLM technology in ICD coding as well as the cybersecurity research that must go behind these technologies to ensure safe clinical implementation.