Performance of ChatGPT on Corroborating North American Spine Society Clinical Guidelines for the Diagnosis and Treatment of Cervical Radiculopathy

Timothy Hoang, Lathan Liou, Justin Evan Tang, Ashley Rosenberg, Nancy Shrestha, Jun Kim, Samuel Kang-Wook Cho

INTRODUCTION:

Cervical radiculopathy (CR), a disorder characterized by dysfunction of the spinal nerve due to compression or inflammation, affects 83.2 per 100,000 persons. The nonoperative health-related costs associated with CR—stemming from time off work, decreased productivity, and medical expenses—are estimated to be $14.3 million annually. In 2010, the North American Spine Society (NASS) issued evidence-based clinical guidelines, featuring a list of educational questions and answers for practitioners addressing diagnosis and treatment of CR.

Chat Generative Pre-Trained Transformer (ChatGPT), a sophisticated large-language artificial intelligence (AI) tool, has shown potential in providing comprehensible answers to complex medical inquiries. Since its debut, it has sparked interest among clinicians because of its promise as a "curbside consultation" tool. Integrating AI solutions into the diagnostic process may expedite identification of CR, potentially mitigating the aforementioned health-related costs. Thus, a quantitative assessment of ChatGPT's potential to enhance diagnostic capacity is warranted. This study aimed to assess the utility and accuracy of ChatGPT by comparing its responses to NASS clinical questions with the evidence-based answers published in the NASS CR guidelines.

METHODS:

The NASS guidelines for CR included 18 questions, four of which were excluded because the guideline authors graded them as having insufficient evidence in the literature. The remaining questions were used as input to OpenAI's ChatGPT software. ChatGPT's answers were compared with NASS guidelines and evaluated for concordance. Selected key phrases within the NASS, CR guidelines were identified and each assigned a value of one accuracy point. A total of 100 accuracy points were identified across the 14 clinical guidelines. Accuracy was measured as the total number of key phrases identified between ChatGPT and NASS with one accuracy point given for each key phrase mentioned. Flesch reading ease scores were also measured to assess interpretability.

Jaccard Similarity Index, calculated as J(A,B)= ABAB, was used to assess agreement between ChatGPT responses and NASS guidelines. Several sensitivity analyses were performed to more robustly assess ChatGPT performance.

RESULTS:

When used with a complex prompt in a new session, ChatGPT-4 responses exhibited an accuracy of 45% compared with NASS guidelines. In contrast, ChatGPT-3.5, when used with a complex prompt in a new session, yielded an accuracy of 34%. Therefore, ChatGPT-4 outperformed ChatGPT-3.5 by a margin of 11%. The accuracy of ChatGPT responses improved by 1-4% when using complex versus simple prompts. Responses from a new or existing session enhanced the accuracy of ChatGPT responses by 3-6%. ChatGPT responses, when queried and collected on different days, varied by 3% (Table 1). With NASS guidelines as a reference, ChatGPT-3.5 registered a higher average Jaccard similarity index than ChatGPT-4 (Figure 1).

ChatGPT-4, with an average Flesch reading score of 15.24, was graded as very difficult to read, typically understood by those with a college graduate education level. On the other hand, ChatGPT-3.5 exhibited an average Flesch reading score of 8.73, requiring a professional education level and graded as extremely difficult to read. The NASS Guidelines also demonstrated a professional education level readability, with an average Flesch reading score of 4.58.

DISCUSSION AND CONCLUSION:

ChatGPT-4 demonstrated superior performance over ChatGPT-3.5 in delivering accurate responses to frequently asked questions about CR. However, the average similarity of ChatGPT-3.5 with NASS guidelines surpassed that of ChatGPT-4. A closer inspection of keyword sets showed that there was a large discrepancy in the respective Jaccard similarity indices for specific queries. Interestingly, ChatGPT-4 responses tended to be more succinct, while ChatGPT-3.5's outputs were typically more verbose. This observation may account for the divergence in Jaccard similarity indices between the two model versions.

The non-deterministic nature of ChatGPT was considered in this study and sensitivity analyses were used to quantify the effects of various input parameters. Interestingly, simple prompts consistently outperformed complex ones, potentially due to the AI model's proficiency in dealing with less complex language structures. Using a new session for each query improved the accuracy scores across scenarios, which was expected as ChatGPT's contextual understanding could introduce bias from prior prompts. The observed day-to-day variance in ChatGPT responses further substantiates existing literature on ChatGPT's non-deterministic behavior.

ChatGPT-4, while offering the most readable responses, produced outputs classified as "very difficult to read," requiring a high education level for comprehension. This poses challenges for its potential use as a "curbside consultation" AI tool and highlights the need for improvements in language accessibility in future AI models. Further research exploring whether ChatGPT can be prompted to provide a more interpretable response is needed. This study underscores the importance of ongoing evaluation of large language models like ChatGPT in the medical context, to provide healthcare professionals with evidence-based guidelines for their use and to ensure that they can effectively enhance patient understanding and engagement in their care.

Figure 1. Barplot of Jaccard Similarity Indices comparing ChatGPT-4 and ChatGPT-3.5 Responses to the NASS Clinical Guidelines. Horizontal lines represent average Jaccard Similarity Indices of ChatGPT 3.5 (red) and ChatGPT 4 (blue).

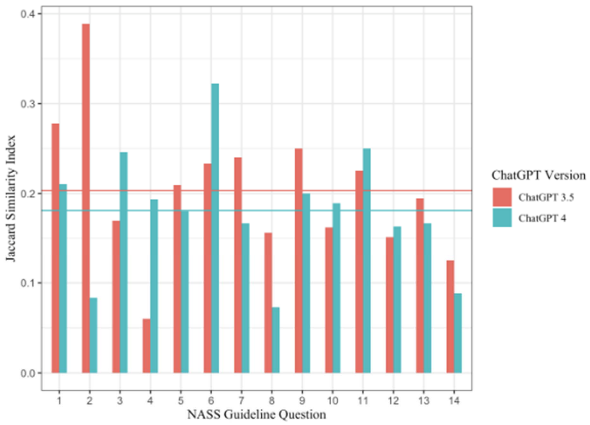

Table 1. ChatGPT-3.5 and ChatGPT-4 Accuracy Compared to the NASS Clinical Guidelines in Different Scenarios.

ChatGPT Version	Prompt Complexity	Session Type	Day	Points Earned (Out of 100)	Accuracy
4 (Reference)	Complex	New	1	45	45%
4	Complex	New	2	48	48%
4	Simple	New	1	49	49%
4	Complex	Existing	1	42	42%
4	Simple	Existing	1	46	46%
3.5	Complex	New	1	34	34%