Neurospine 2024;21(1):147-148.
https://doi.org/10.14245/ns.2448236.118

Neurospine
pISSN 2586-6583 eISSN 2586-6591

Commentary

Commentary on "Performance of a Large Language Model in the Generation of Clinical Guidelines for Antibiotic Prophylaxis in Spine Surgery"

Sun-Ho Lee

Department of Neurosurgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea

Corresponding Author

Sun-Ho Lee
https://orcid.org/0000-0003-3357-4329

Department of Neurosurgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, 81 Irwon-ro, Gangnam-gu, Seoul 06351, Korea
Email: sunho72.lee@samsung.com

See the article "Performance of a Large Language Model in the Generation of Clinical Guidelines for Antibiotic Prophylaxis in Spine Surgery" via https://doi.org/10.14245/ns.2347310.655.

The introduction of artificial intelligence (AI), particularly large language models (LLMs) such as the generative pre-trained transformer (GPT) series into the medical field has heralded a new era of data-driven medicine. AI's capacity for processing vast datasets has enabled the development of predictive models that can forecast patient outcomes with remarkable accuracy. LLMs like GPT and its successors have demonstrated an ability to understand and generate human-like text, facilitating their application in medical documentation, patient interaction, and even in generating diagnostic reports from patient data and imaging findings. Over the past 10 years, the development of AI, LLMs, and GPTs has significantly impacted the field of neurosurgery and spinal care as well.[1-5]

Zaidat et al.[6] studied performance of a LLM in the generation of clinical guidelines for antibiotic prophylaxis in spine surgery. This study delves into the capabilities of ChatGPT's models, GPT-3.5 and GPT-4.0, showcasing their potential to streamline medical processes. They suggest that GPT-3.5's ability to generate clinically relevant antibiotic use guidelines for spinal surgery is commendable; however, its limitations, such as the inability to discern the most crucial aspects of the guidelines, redundancy, fabrication of citations, and inconsistency, pose significant barriers to its practical application. GPT-4.0, on the other hand, demonstrates a marked improvement in response accuracy and the ability to cite authoritative guidelines, such as those from the North American Spine Society (NASS). This model's enhanced performance, including a 20% increase in response accuracy and the ability to cite the NASS guideline in over 60% of responses, suggests a more reliable tool for clinicians seeking to integrate AI-generated content into their practice.

However, the study's findings also highlight the inherent unpredictability of LLM responses and the potential for "artificial hallucination," where models generate spurious statements without a solid basis in their training data. This phenomenon raises concerns about the ethical implications of using LLMs in clinical settings, particularly regarding patient care and liability. The possibility of LLMs providing inaccurate responses, especially when prompted for medical advice, necessitates a cautious approach to their deployment. We also pay attention to the limitations of the study itself, including the outdated nature of the NASS guidelines, which have not been updated since 2013, and the potential biases and gaps in the medical knowledge contained within the LLMs' training data. These factors highlight the im-

portance of ongoing research and development to ensure that LLMs are trained on the most recent and relevant medical literature.

From the perspective of a spine surgeon, the advancements from GPT-3.5 to GPT-4.0 are noteworthy. Recent studies have showcased the diverse applications of these AI models, from enhancing clinical outcomes aiding in diagnostics and patient care.[7,8] While LLMs like GPT-3.5 and GPT-4.0 hold significant promise for enhancing clinical practice, their current application should be approached with caution. Clinicians must critically evaluate the information provided by these models and should not rely on them exclusively for clinical recommendations. The future direction of LLM development, including the anticipated release of GPT-4.0 Turbo and domain-specific models trained on medical literature, offers exciting possibilities for the field. However, the medical community must balance this enthusiasm with rigorous research to understand the models' limitations and to develop evidence-based guidelines for their safe and effective use in clinical settings. As we stand on the beginning of a new era in medical AI, it is imperative that we proceed with both caution and optimism, ensuring that patient care remains at the forefront of our priorities.

• **Conflict of Interest:** The author has nothing to disclose.

REFERENCES

1. Chang M, Canseco JA, Nicholson KJ, et al. The role of machine learning in spine surgery: the future is now. Front Surg 2020;7:54.
2. Kuang YR, Zou MX, Niu HQ, et al. ChatGPT encounters multiple opportunities and challenges in neurosurgery. Int J Surg 2023;109:2886-91.
3. Schwartz JT, Gao M, Geng EA, et al. Applications of machine learning using electronic medical records in spine surgery. Neurospine 2019;16:643-53.
4. Kim SH, Lee SH, Shin DA. Could machine learning better predict postoperative C5 palsy of cervical ossification of the posterior longitudinal ligament? Clin Spine Surg 2022;35:E419-25.
5. Noh SH, Lee HS, Park GE, et al. Predicting mechanical complications after adult spinal deformity operation using a machine learning model based on modified global alignment and proportion scoring with body mass index and bone mineral density. Neurospine 2023;20:265-74.
6. Zaidat B, Shrestha N, Rosenberg AM, et al. Performance of a large language model in the generation of clinical guidelines for antibiotic prophylaxis in spine surgery. Neurospine 2024;21:128-46.
7. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2023;93:1353-65.
8. Shrestha N, Shen Z, Zaidat B, et al. Performance of ChatGPT on NASS clinical guidelines for the diagnosis and treatment of low back pain: a comparison study. Spine (Phila Pa 1976) 2024 Jan 12. doi: 10.1097/BRS.0000000000004915. [Epub].

 https://doi.org/10.14245/ns.2448236.118