Artificial Intelligence in Orthopaedic Surgery: Can a Large Language Model Write a Believable Orthopaedic Journal Article?

Devon Tracey Brameier, Ahmad Alnasser, Jonathan Carnino, Abhiram Bhashyam, Arvind Gabriel Von Keudell, Michael John Weaver[1]
[1]Brigham and Women's Hospital

INTRODUCTION: The use of artificial intelligence (AI), particularly in the form of natural language processing (NLP) with large language models (LLM), in medicine and orthopaedic surgery is rapidly growing. Concerns have emerged regarding the potential for research misconduct and misinformation. This project sought to evaluate the use of LLMs to “write” academic research papers and comment on the peer review process of LLM-generated articles by accredited orthopaedic journals.
METHODS:
A qualitative study of LLM use in orthopaedic writing and the peer review of LLM-generated manuscripts was performed. ChatGPT was used to generate two manuscripts on topics in orthopaedic trauma surgery: 1) A randomized controlled trial of surgical management in minimally displaced femoral neck fractures (Article A), and 2) A randomized controlled trial of surgical management in distal humerus fractures (Article B). Researchers without formal orthopaedic training prompted ChatGPT to generate the title, abstract, body text, results, tables, and references. ChatGPT’s responses were compiled with limited editing by the researchers. The generated manuscripts were submitted to the Journal of Bone and Joint Surgery (JBJS) for formal review in collaboration with the editor-in-chief. The reviewers were blinded to the manuscripts’ fabricated nature.
RESULTS: Two manuscripts were generated using LLMs with minimal researcher and surgeon input. Approximately 20% of each manuscript was edited by one of our researchers to make them submission ready. ChatGPT “hallucinated” data and statistics, references, and conclusions within the discussion – providing confident responses which seem realistic but are not based on any real-world data. After blinded peer review, Article A was accepted for publication at JBJS and Article B was referred for additional review and consideration for publication by JBJS Open Access. Each article received 30-35 reviewer comments. The reviewers did not question the authorship of the manuscript or the validity of the study. All AI hallucinations, including falsified references, were missed by reviewers.
DISCUSSION AND CONCLUSION: LLM algorithms can generate text of publishable quality with minimal expert input. Like any tool, this could be used to improve the quality of papers written using real data, but it also raises significant risks. The use of AI is susceptible to hallucination. The current peer review process is not capable of managing this emerging threat. The responsibility for screening academic submissions for the malicious use of artificial intelligence assistance lies with the editorial offices. They must encourage safe AI use by establishing clear guidelines for the use of these tools across the orthopaedic literature and introducing additional steps in the screening process to identify AI-involvement in submitted manuscripts.