The Accuracy, Credibility and Relevance of the Artificial Intelligence Generated Information for the Use in General Surgery Practice

Danis Lester, DO MS, Audric Darian, MD, Onyinye Nnamdi, MD, Taylor Cardwell, MD, Leana Reagan, MD, Alexa Wacker, MD, Nabeel Hussain, MD, Niki Augustine, MD, Matthew Di Leo, MD, Darwin Ang, MD PhD

UCF College of Medicine
UNIVERSITY OF CENTRAL FLORIDA

HCA Florida Ocala Hospital

Background

The integration of large language models (LLMs) with artificial intelligence (AI) search engines is transforming information access and clinical decision-making across specialties, including general surgery. These hybrid systems combine advanced natural language generation with information retrieval capabilities. However, the accuracy, credibility, and relevance of information generated by these systems remain under scrutiny, especially in surgical contexts where precision and reliability are paramount.

Objective

This first-of-its-kind study aims to evaluate the accuracy, credibility, and relevance of information generated by six popular AI search engines that integrate LLMs as of 2024 for potential use in general surgery. By examining these facets, this research seeks to determine the suitability of AI-generated information in aiding clinical decisions and educational initiatives within the general surgery domain.

Methods

A selected AI search engine integrating LLMs was tasked with generating information on four specific general surgery cases: incarcerated/strangulated inguinal hernia, acute appendicitis, acute cholecystitis, and necrotizing soft tissue infection (NSTI). Nine physicians in general surgery evaluated the AI-generated responses using a 5-point Likert scale across five categories: relevance, authority, accuracy, completeness, and accessibility. The relevance scale assessed how directly the information addressed clinical scenario aspects; authority evaluated the credibility of sources; accuracy measured alignment with current clinical guidelines; completeness assessed the level of detail; and accessibility was based on paywall restrictions. Each category ranged from 1 (e.g., "Not relevant" or "Not accurate") to 5 (e.g., "Extremely relevant" or "Extremely accurate"). The AI-generated information was re-evaluated after three months, with no significant changes in content or quality observed.

This research was supported (in whole or in part) by HCA Healthcare and/or an HCA Healthcare affiliated entity. The views expressed in this publication represent those of the author(s) and do not necessarily represent the official views of HCA Healthcare or any of its affiliated entities.

Results

Among the six evaluated AI search engines integrating LLMs, Engine 5 demonstrated the highest relevance (4.76), while Engine 2 scored the lowest in authority (1.36). Accuracy was generally high, with Engine 3 leading at 4.66. Completeness varied, with Engine 3 scoring the highest (4.47) and Engine 2 the lowest (3.65). Accessibility scores ranged from 4.56 for Engine 2 to 3.12 for Engine 5, highlighting variability in user-friendliness and pay-to-use barriers. These results indicate mixed performance across engines, with strengths in relevance and accuracy but variability in authority/credibility and accessibility.

AI Search Engine	Engine 1	Engine 2	Engine 3	Engine 4	Engine 5	Engine 6
Overall Relevence	4.35	4.2	4.65	5	4.85	4.15
Overall Authority	3.35	1	4.5	3.55	2	2.27
Overall Accuracy	3.8	3.15	4.5	4.3	4.25	3.7
Overall Completeness	3.3	2.7	4.25	4.1	4.1	3.1
Overall Accessibility	5	5	3	5	3	5

Conclusion

While AI search engines integrating LLMs show promise as supportive tools for general surgery, outputs require careful expert review. This study highlights both the potential benefits and limitations of AI-generated information in surgical practice, underscoring the need for ongoing evaluation and model refinement to improve reliability for clinical and even potential public use.

References

1. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv. Preprint posted online May 28, 2020. doi:10.48550/arXiv.2005.14165
2. Goodman RS, Patrinely JR, Stone CA, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Netw Open. 2023;6(10):e2336483. doi:10.1001/jamanetworkopen.2023.36483
3. Shen Y, Heacock L, Elias J, et al. ChatGPT and Other Large Language Models Are Double-edged Swords. Radiology. 2023;307(2):e230163. doi:10.1148/radiol.230163
4. Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. arXiv. Preprint posted online February 17, 2023. https://arxiv.org/pdf/1706.03741.pdf
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. medRxiv. 2022:2022.12.19.22283643.
6. Gilson A, Safranek C, Huang T, Socrates V, Chi L, Taylor RA, et al. How Well Does ChatGPT Do When Taking the Medical Licensing Exams? The Implications of Large Language Models for Medical Education and Knowledge Assessment. medRxiv. 2022:2022.12. 23.22283901.
7. Johnson D, Goodman R, Patrinely J, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Preprint. Res Sq. 2023;rs.3.rs-2566942. Published 2023 Feb 28. doi:10.21203/rs.3.rs-2566942/v1
8. Hosseini M, Rasmussen LM, Resnik DB. Using AI to write scholarly publications. Account Res. Published onlne January 25, 2023. doi:10.1080/08989621.2023.2168535
9. Thorp HH. ChatGPT is fun, but not an author. Science. 2023;379(6630):313. doi:10.1126/science.adg7879
10. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. Nature. 2023;613 (7945):620-621. doi:10.1038/d41586-023-00107-z
11. Flanagin A, Bibbins-Domingo K, Berkwits M, Christiansen SL. Nonhuman "authors" and implications for the integrity of scientific publication and medical knowledge. JAMA. 2023;329(8):637-639. doi:10.1001/jama. 2023.1344
12. Goodman RS, Patrinely JR Jr, Osterman T, Wheless L, Johnson DB. On the cusp: Considering the impact of artificial intelligence language models in healthcare. Med. 2023;4(3):139-140. doi:10.1016/j.medj.2023.02.008

HCA Healthcare®