Beyond Clinical Accuracy: Considerations for the Use of Generative Artificial Intelligence Models in Gastrointestinal Care

See "ChatGPT answers common patient questions about colonoscopy," by Lee T-S, Staller K, Botoman V, et al, on page 509.

As the volume and complexity of health care data continue to grow, the field of gastroenterology has embraced the use of computational tools to identify, extract, and synthesize relevant information.[1] Rapidly expanding from the use of medical record data only, machine learning and artificial intelligence (AI) techniques now routinely integrate data from procedural images and free-text documents (eg, clinical notes, academic articles, and online resources).[2,3] Clinically, this has manifested in predictive models and risk stratification tools to improve prognosis, diagnosis, treatment, and patient management.[4] For patients, unparalleled access to digital resources has facilitated engagement in their care.[5] It is largely understood that these tools cannot, and should not, replace gastroenterologists. However, the ability to leverage this technology in an assistive capacity is fundamentally changing the way clinicians and patients interact with data.

One notable change already impacting health care is the ability to streamline navigation of health-centric resources. In this issue of *Gastroenterology,* Lee and colleagues[6] explored an application of such, assessing the potential of ChatGPT, an emerging natural language processing technology, to provide patients with accurate and understandable answers to common questions regarding colonoscopy. As patients turn increasingly to online resources,[7] this information is often available through health care system's online frequently asked questions. However, the static nature of these webpages limits utility when questions do not align with prespecified items. It has thus become necessary to expand both the way in which information can be requested and how, once identified, insights are provided for consumption.

Early natural language processing efforts used chatbots that allowed patients to dynamically enter free-text questions.[8] However, these models largely operated akin to voice assistants, capturing key phrases in questions and providing responses from a limited set of predefined data. Conversely, this emergent class of natural language processing (colloquially known as "generative AI"), leverages large-scale language models to capture the context of a question and produce meaningful responses based on information from a broader set of general health data.[9] The authors' primary result highlights a high-degree of clinical accuracy in ChatGPT's response to 8 questions. However, it is their secondary analyses, evaluating the ease of understanding and reliability of AI-generated responses, where we find a core theme surrounding the future use of these models in practice, that is, data and information quality.

The concept of quality is multifaceted, and to understand how these tools impact patients and clinicians, we must consider multiple factors jointly. We broadly characterize these factors into 3 domains—performance, appropriateness, and accessibility.

Performance

At its core, utility of generative AI is contingent on an ability to provide accurate, complete, and reliable information. Highlighted by Lee et al and others,[6,10,11] these tools can produce results with a reasonable degree of clinical accuracy. However, factual inaccuracies known to be contained in the web resources used to build responses[12] pose a risk for providing inaccurate information and must be monitored carefully. Moreover, response quality has been found to be dependent on how questions are asked.[13] Thus, appropriately framing clinical questions may present a barrier for lay patients to obtain precise answers.

Furthermore, the machine learning and AI models on which these tools are built are probabilistic and, as demonstrated by Lee et al,[6] can produce different answers to the same question. This lack of reliability creates a challenge for clinicians to guide patients to specific information and for counseling them based on expected results.

Recent studies have noted that even accurate and reliable results may be incomplete, failing to provide necessary information to fully contextualize health care scenarios.[14] As US Food and Drug Administration regulations on monitoring performance of dynamic models remain in the early stages,[15] a methodology to quantify uncertainty or safety of responses is needed.

Appropriateness

As medical knowledge continues to progress, the use of generative AI in practice will be dependent on its ability to provide up-to-date information. Although these tools can have access to even the most recent data, their ability to convey and account for data that changes over time remains unclear. Similarly, there exists a need to disambiguate data from varying sources, such as differentiating between established clinical standards and emerging research. This complexity is compounded in situations where multiple current resources may conflict (eg, different guidelines on colorectal cancer screening[16]). Addressing such variability will be a key factor for successful implementation of these tools in rapidly advancing gastroenterology practice.

Downloaded for Anonymous User (n/a) at Children's Mercy Kansas City from ClinicalKey.com by Elsevier on July 22, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.

In addition, the data used to build these tools present a challenge to their equitable use.[17] Historical data are known to contain biases perpetuated through society. They may not be representative of all individuals who intend to use this tool and may impact the ability to provide appropriate responses. As such, transparency around what data are used to build these models must be improved before widespread use.

Accessibility

At a fundamental level, these models are trained on a breadth of data beyond that accessible to most individuals. Synthesizing this information presents an opportunity to broaden access and may aid in reducing disparities in underserved communities.[18] However, information alone provides little utility if it cannot be understood. Health literacy remains a barrier in providing usable responses to complex health-related questions.[19] As reported by Lee at al,[6] ChatGPT's response readability exceeded the 8th-grade level, limiting utility for a subset of patients and potentially widening the gap to health care access. Moreover, as the underlying process to generate the responses results from the output of complex neural models, explaining why specific information was provided remains challenging, limiting clinician's ability to moderate or explain concerns that may arise based on the use of such tools.[20]

Although generative AI remains in its infancy, the ability to leverage the flexibility and broad knowledge base of these tools holds the potential to augment and assist multiple aspects of gastrointestinal care. However, work by Lee et al[6] and others lays a foundation for a range of quality metrics needed for its successful implementation. Although clinical accuracy of these tools is necessary, it is not sufficient, and addressing the full spectrum of such is a grand challenge for the coming years and will require the collaborative efforts of patients, clinicians, and computational scientists alike.

KEITH FELDMAN
Division of Health Services and Outcomes Research
Children's Mercy Kansas City
Kansas City, Missouri, *and*
Department of Pediatrics
University of Missouri-Kansas City School of Medicine
Kansas City, Missouri

FREDY NEHME
Division of Gastroenterology and Hepatology
Indiana University School of Medicine
Indianapolis, Indiana

References

1. Cheung K-S, Leung WK, Seto W-K. Application of Big Data analysis in gastrointestinal research. World J Gastroenterol 2019;25:2990.
2. Penrice DD, Rattan P, Simonetto DA. Artificial intelligence and the future of gastroenterology and hepatology. Gastro Hep Adv 2022;1:581–595.
3. Nehme F, Feldman K. Evolving role and future directions of natural language processing in gastroenterology. Dig Dis Sci 2021;66:29–40.
4. Adadi A, Adadi S, Berrada M. Gastroenterology meets machine learning: status quo and quo vadis. Adv Bioinform 2019;2019:1870975.
5. Catlow J, Bray B, Morris E, et al. Power of big data to improve patient care in gastroenterology. Frontline Gastroenterol 2022;13:237–244.
6. Lee T-C, Staller K, Botoman V, et al. ChatGPT answers common patient questions about colonoscopy. Gastroenterology 2023;165:509–511.
7. Tan SS-L, Goonawardene N. Internet health information seeking and the patient-physician relationship: a systematic review. J Med Internet Res 2017;19:e9.
8. Zand A, Sharma A, Stokes Z, et al. An exploration into the use of a chatbot for patients with inflammatory bowel diseases: retrospective cohort study. J Med Internet Res 2020;22:e15589.
9. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Adv Neural Inform Proc Syst 2022;35:27730–27744.
10. Levine DM, Tuwani R, Kompa B, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model. medRxiv. Preprint posted online February 1, 2023. https://doi.org/10.1101/2023.01.30.23285067.
11. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health 2023;2:e0000198.
12. Chiang AL, Jajoo K, Shivashankar R, et al. Scoping out misinformation: assessing factual inaccuracies among popular colonoscopy-related videos on social media. Gastro Hep Adv 2022;1:923–925.
13. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388:1233–1239.
14. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. arXiv. Preprint posted online December 26, 2022. https://doi.org/10.48550/arXiv.2212.13138.
15. US Food and Drug Administration. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD). Available at: https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf. Accessed June 11, 2023.
16. Bénard F, Barkun AN, Martel M, et al. Systematic review of colorectal cancer screening guidelines for average-risk adults: summarizing the current global recommendations. World J Gastroenterol 2018;24:124–138.
17. Uche-Anya E, Anyane-Yeboa A, Berzin TM, et al. Artificial intelligence in gastroenterology and hepatology: how to advance clinical practice while ensuring health equity. Gut 2022;71:1909–1915.

Downloaded for Anonymous User (n/a) at Children's Mercy Kansas City from ClinicalKey.com by Elsevier on July 22, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.

18. Li W, Zhang Y, Chen F. ChatGPT in colorectal surgery: a promising tool or a passing fad? [published online ahead of print May 10, 2023]. Ann Biomed Eng https://doi.org/10.1007/s10439-023-03232-y.
19. Kobayashi LC, Wardle J, von Wagner C. Limited health literacy is a barrier to colorectal cancer screening in England: evidence from the English Longitudinal Study of Ageing. Prev Med 2014;61:100–105.
20. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine 2023;90:104512.

Received May 31, 2023. Accepted June 9, 2023.

Correspondence
Address correspondence to: Fredy Nehme, MD, MS, Division of Gastroenterology and Hepatology, Indiana University School of Medicine, 240 W 10th Street, Indianapolis, Indiana, 46202. e-mail: nehme.fredy@gmail.com.

Conflicts of interest
The authors disclose no conflicts.

Most current article

© 2023 by the AGA Institute.
0016-5085/$36.00
https://doi.org/10.1053/j.gastro.2023.06.006

Downloaded for Anonymous User (n/a) at Children's Mercy Kansas City from ClinicalKey.com by Elsevier on July 22, 2023. For personal use only. No other uses without permission. Copyright ©2023. Elsevier Inc. All rights reserved.