Published in final edited form as:

Gastro Hep Adv. 2023 ; 2(8): 1040–1043. doi:10.1016/j.gastha.2023.07.008.

Assessing ChatGPT's Ability to Reply to Queries Regarding Colon Cancer Screening Based on Multisociety Guidelines

S. MUKHERJEE[1], **C. DURKIN**[1], **A. M. PEBENITO**[1], **N. D. FERRANTE**[1], **I. C. UMANA**[1], **M. L. KOCHMAN**[1,2]

[1]Gastroenterology Division, Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania

[2]Department of Medicine, Center for Endoscopic Innovation Research and Training, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania

ChatGPT™ is a chatbot defined Artificial Intelligence program launched by San Francisco–based OpenAI on November 30th, 2022, with the ability to hold human-like conversations.[1] Although literature commenting on ChatGPT ™'s abilities has grown over the past months, individual studies assessing its utility in clinical care, research and teaching in the field of Gastroenterology (GI) has been scarce with only 2 reported studies.[2,3] Our study assesses ChatGPT™'s ability to answer queries regarding appropriate colonoscopy intervals for colon cancer screening compared to currently applicable guidelines.

Utilizing the American Gastroenterological Association (AGA) 's recommendations for follow-up after colonoscopy and polypectomy,[4,5] 12 questions were developed to query ChatGPT ™ (Table). The queries were entered into ChatGPT ™ by the author (SM) with the responses being separately documented (Appendix 1). Each of the 12 query-response pairs underwent adjudication by 4 senior GI fellows (CD, AP, NF, IU) who graded the responses on a semi-qualitative scale over a set of 5 options ranging from "addresses the query and is factually entirely correct" to "does not address the query and is factually incorrect". A field to comment on the potential usefulness to patients was provided. Adjudicators were provided a copy of the AGA guideline as base truth to aid assessment of responses. All 4 adjudicators were blinded regarding the source of the responses to reduce potential bias.

Correspondence: Address correspondence to: Samiran Mukherjee, MD, Division of Gastroenterology, Perelman Center for Advanced Medicine South Pavilion, 4th Floor 3400 Civic Center Boulevard, Philadelphia, Pennsylvania 19104. samiranmukherjee93@gmail.com.

Conflicts of Interest:
The authors disclose no conflicts.

Ethical Statement:
The corresponding author, on behalf of all authors, jointly and severally, certifies that their institution has approved the protocol for any investigation involving humans or animals and that all experimentation was conducted in conformity with ethical and humane principles of research.

Reporting Guidelines:
Guidelines used were the AGA Multisociety Task force on Colon Cancer Recommendations for Follow-Up After Colonoscopy and Polypectomy.

Supplementary materials
Material associated with this article can be found in the online version at https://doi.org/10.1016/j.gastha.2023.07.008.

All adjudicators were informed that the responses were generated by ChatGPT ™ after conclusion of the study. The study did not meet criteria for institutional review board submission given the absence of human subjects.

Three of 4 (75%) adjudicators felt that ChatGPT™'s response to Q1 (*What is the risk developing a colon cancer leading to death after a clear colonoscopy?*) addressed the query and was factually correct. One of 4 stated it was inaccurate in its reporting of colon cancer incidence as a percentage (as opposed to a hazard ratio). Three of 4 felt the answers would be usable by patients.

Only 50% (2/4) of the adjudicators felt that ChatGPT™'s response to Q2 (*When should colon screening be repeated in a patient with a quality colonoscopy?*) addressed the query and was factually correct. 100% agreed that the answer would be usable by patients. ChatGPT™ had suggested starting colon cancer screening at 50, with repeat colonoscopies every 10 years. While it was accurate regarding the time interval for repeat colonoscopy, it was inaccurate regarding the age to initiate screening (45 for average risk).

Similarly, when assessing ChatGPT™ 's response to Q3 (*Repeat colonoscopy for patients who had 1–2 small tubular adenomas <10 mm in size that have been completely resected at a high-quality examination?*), 75% (3/4) felt that the queries would be usable by patients and 75% (3/4) agreed that while it did address the query, it contained both correct and incorrect responses. ChatGPT™'s response was that the interval was to be "5–10 years" (instead of 7–10 years).

Kappa for interrater reliability was 0.189 for all 12 questions, 0.248 for the first 3 questions and 0.704 when assessing patient usability. Analysis was performed using RStudio.[6]

A summary of all queries is presented in Table. Critical observations of all responses are presented in Appendix 1. None of the responses completely inaccurate as none were found by all adjudicators to be completely wrong. ChatGPT™ was also able to identify rare genetic syndromes in Q11–12.

ChatGPT™'s introduction has generated widespread interest the academic community. Its ability to draft entire essays and even pass the United States Medical Licensing Exam has led to debates about the ethics of its use.[1,7] One area which continues to generate discussion is the question of its authorship on publications.[8,9] Scholarly societies, such as World Association of Medical Editors state that chatbots cannot be authors as they do not create new knowledge.[10]

Its capabilities in GI education and research remains relatively unexplored, with only 2 studies describing early experience.[2,3] Lahat et al,[2] assessed its ability to identify questions related to GI research and concluded that while it was able to frame questions, they were not considered novel. Yeo et al[3] assessed its ability to answer questions on the management of liver cirrhosis and hepatocellular carcinoma where it performed favorably.

The purpose of our study was 2-fold: *First*, can ChatGPT™ accurately answer queries regarding colonoscopy intervals as held to the standard of currently active guidelines?

Second, could it be a tool in patient selfeducation? Regarding the former, its ability to respond to simple and direct questions (Questions 1–3) was greater in straightforward queries when compared to the more nuanced questions. Regarding the latter, while there was no patient data used in this project, adjudicator assessments suggest it may be a useful patient tool for background information to inform discussion with treating physicians. It is not felt to be useful for self-directed care due to potential imprecision.

The study has several strengths: we assessed the accuracy of ChatGPT™'s responses against a standard of care guideline and found that ChatGPT™'s ability to provide accurate responses diminishes with more complex medical queries. Additionally, our findings highlight a potential role for ChatGPT™ as an adjunct tool for patient education on the utility and timing of follow-up colonoscopy but should not replace information received from a licensed medical provider.

Regarding its limitations: First, human adjudication is prone to error and the small number of adjudicators and verbosity of ChatGPT™ responses have resulted in variability in adjudication, as reflected in the weak kappa statistic. Second, suitability of ChatGPT™'s responses for patient education was determined by the adjudicators as opposed to patients. Third, ChatGPT™'s training data is current through September 2021 which may have contributed to ChatGPT™'s inaccuracy.

In conclusion, we assessed ChatGPT™'s ability to answer queries regarding appropriate colonoscopy intervals for colon cancer screening and surveillance. Although in its current iteration it under-delivers, it does appear to be a potential source of background information for patient selfeducation. As global interest in ChatGPT™ continues to increase and the technology iterates, we expect that future renditions will be able address nuanced queries with increased precision, serving as a readily available resource for GI education.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding:

The following authors have been supported by the National Institute of Diabetes and Digestive and Kidney Diseases T32 training grants; Dr Samiran Mukherjee (DK007066-49), Dr Claire Durkin (DK007740), Dr Amanda M PeBenito (DK007066-49), Dr Nicole D Ferrante (DK007740) Dr Iboro C Umana (DK007066-49S1).This study has been supported by the The Willmott Family Fund for Endoscopic Innovation, Research and Training awarded to Dr Michael L Kochman.

Data Transparency Statement:

Data, analytic methods, and study materials will be made available to other researchers (Attached as Appendix 1).

References

1. Graham F. Nature 2022. 10.1038/d41586-022-04437-2.
2. Lahat A, et al. Sci Rep 2023;13:4164. [PubMed: 36914821]

3. Yeo YH, et al. Clin Mol Hepatol 2023;29:721–732. [PubMed: 36946005]
4. Gupta S, et al. Gastroenterology 2020;158:1131–1153.e5. [PubMed: 32044092]
5. Gupta S, et al. Gastrointest Endosc 2020;91:463–485.e5. [PubMed: 32044106]
6. RStudio Team. RStudio: integrated development for R. Boston: RStudio, PBC, 2020.
7. The Lancet Digital Health. Lancet Digit Health 2023;5:e102. [PubMed: 36754723]
8. Stokel-Walker C. Nature 2023; 613:620–621. [PubMed: 36653617]
9. Thorp HH. Science 2023;379:313. [PubMed: 36701446]
10. Zielinski C, et al. Chatbots, ChatGPT, and scholarly manuscripts: WAME recommendations on ChatGPT and chatbots in relation to scholarly publications. WAME. 2023. https://wame.org/page3.php?id=106. Accessed January 20, 2023.

Table.

Itemized List of 12 Clinical Queries With Overall Adjudicator Assessment of the Responses by ChatGPT™

Sno	Question	Does not address the query and is factually incorrect. Please state what was incorrect	Does not address the query but the information is factually correct though unrelated	Adresses the query and is factually incorrect to the query.	Adresses the query and has both factually correct and incorrect responses (whether or not related to the query).	Adressess the query and is factually entirely correct	Will the answers be usable by patients to address the question asked?
1	What is the risk developing a colon cancer leading to death after a clear colonoscopy?	25%				75%	Yes (75%) No (25%)
2	When should colon screening be repeated in a patient with a quality colonoscopy?		50%			50%	Yes (100%), No (0%)
3	When should repeat colonoscopy be scheduled for patients who had 1–2 small tubular adenomas <10 mm in size that have been completely resected at a high quality examination?				75%	25%	Yes (75%), No (25%)
4	When should colonoscopy be repeated for patients with 5–10 small tubular adenomas <10 mm in size that have been completely resected at a high quality examination?	25%		25%	50%		No (100%)
5	What is the recommended surveillance interval for colonoscopy after the resection of an adenoma ≥10 mm?				100%		No (100%)
6	When should colonoscopy be repeated for patients with any polyp with villous histology that was completely removed at a high quality examination?				100%		No (100%)
7	When should colonoscopy be repeated for patients with any polyp containing high grade dysplasia that has been completely removed at a high quality examination?			75%	25%		No (100%)
8	When should a colonoscopy be repeated for patients with ≤20 hyperplastic polyps <10 mm in size in the rectum or sigmoid colon removed at a high quality examination?			50%	50%		No (100%)
9	When should a colonoscopy be repeated for patients with piecemeal resection of adenoma or serrated polyp that was greater than 20 mm?			50%	25%%	25%	No (100%)
10	a) What is the recommended interval for the surveillance endoscopy if the baseline finding revealed an adenoma with tubulovillous/villous histology? (If answer correct, proceed to Qb) b) the repeat colonoscopy is normal, when should the next follow-up colonoscopy be?	25%		25%	50%		No (100%)
11	Clinical vignette 1: When should a colonoscopy be repeated for a 60 y old male who had ≥10 adenomas noted in a high quality examinations? Is this patient at risk for any syndromes? If so, what syndromes?			25%	75%		No (100%)
12	Clinical vigntette 2: What is the recommended interval for a surveillance colonoscopy after the removal of ≤20 hyperplastic polyps in the rectum or sigmoid colon <10 mm? Does this patient have a syndrome? If so what?			25%	75%		No (100%)