Health and Technology (2024) 14:463–469
https://doi.org/10.1007/s12553-024-00836-9

ORIGINAL PAPER

Dr. GPT will see you now: the ability of large language model-linked chatbots to provide colorectal cancer screening recommendations

Bright Huo[1] · Tyler McKechnie[1] · Monica Ortenzi[2] · Yung Lee[1] · Stavros Antoniou[3] · Julio Mayol[4] · Hassaan Ahmed[5] · Vanessa Boudreau[1] · Karim Ramji[1] · Cagla Eskicioglu[1]

Received: 23 November 2023 / Accepted: 27 February 2024 / Published online: 4 March 2024

Abstract

Purpose This study assessed the performance of LLM-linked chatbots in providing accurate advice for colorectal cancer screening to both clinicians and patients.

Methods We created standardized prompts for nine patient cases varying by age and family history to query ChatGPT, Bing Chat, Google Bard, and Claude 2 for screening recommendations to clinicians. Chatbots were asked to specify which screening test was indicated and the frequency of interval screening. Separately, the chatbots were queried with lay terminology for screening advice to patients. Clinician and patient advice was compared to guidelines from the United States Preventive Services Task Force (USPSTF), Canadian Cancer Society (CCS), and the U.S. Multi-Society Task Force (USMSTF) on Colorectal Cancer.

Results Based on USPSTF criteria, clinician advice aligned with 3/4 (75.0%), 2/4 (50.0%), 3/4 (75.0%), and 1/4 (25.0%) cases for ChatGPT, Bing Chat, Google Bard, and Claude 2, respectively. With CCS criteria, clinician advice corresponded to 2/4 (50.0%), 2/4 (50.0%), 2/4 (50.0%), and 1/4 (25.0%) cases for ChatGPT, Bing Chat, and Google Bard, respectively. For USMSTF guidelines, clinician advice aligned with 7/9 (77.8%), 5/9 (55.6%), 6/9 (66.7%), and 3/9 (33.3%) cases for ChatGPT, Bing Chat, Google Bard, and Claude 2, respectively. Discordant advice was given to clinicians and patients for 2/9 (22.2%), 3/9 (33.3%), 2/9 (22.2%), and 3/9 (33.3%) cases for ChatGPT, Bing Chat, Google Bard, and Claude 2, respectively. Clinical advice provided by the chatbots stemmed from a range of sources including the American Cancer Society (ACS), USPSTF, USMSTF, and the CCS.

Conclusion LLM-linked chatbots provide colorectal cancer screening recommendations with inconsistent accuracy for both patients and clinicians. Clinicians must educate patients on the pitfalls of using these platforms for health advice.

Keywords Colorectal cancer screening · Large language models · Artificial intelligence

1 Introduction

ChatGPT is a Large Language Model (LLM) that is trained using online datasets [1]. LLMs use Natural Language Processing (NLP) respond to user-generated text input with information-based, human-like responses [1]. LLM-linked chatbots like ChatGPT are capable of conversation simulation and writing tasks [2]. ChatGPT can correct mistakes, reject improper asks, and challenge inaccurate premises [2]. Since its release in 2022 [2], other chatbots such as Google Bard and Bing Chat have emerged [3]. Though these chatbots are not validated for healthcare application, there is mounting interest in their use in patient care [4].

The convenient user interface and wide accessibility of chatbots [3] make them well-positioned for patients seeking

✉ Bright Huo
brighthuo@dal.ca

1 Division of General Surgery, Department of Surgery, McMaster University, 112 King St E, Unit #713, Hamilton, ON L8N1A8, Canada

2 Department of General Surgery, Università Politecnica delle Marche, Ancona, Italy

3 Department of Surgery, Papageorgiou General Hospital, Thessaloniki, Greece

4 Hospital Clinico San Carlos, IdISSC, Universidad Complutense de Madrid, Madrid, Spain

5 Phelix AI, Hamilton, ON, Canada

health advice. Limited data suggests that ChatGPT can generate higher quality responses compared to standard search engines [5], and can produce quality and empathetic responses to patient questions [6, 7]. However, the accuracy of health advice generated by these LLM-linked chatbots is unclear. It is also not known whether chatbots give consistent clinical advice to both clinicians and patients. There is early interest in the use of chatbots to respond to questions regarding cancer screening [5, 8], but current studies apply heterogeneous methodology, clouding their interpretation [9].

There is a need for a deeper understanding of the ability of chatbots to generate accurate advice regarding cancer screening for both clinicians and patients. Studies attempting to address this deficiency must report their methods transparently. Thus, this study applied rigorous methodology to evaluate the performance of LLM-linked chatbots when providing clinical advice for colorectal cancer screening to both clinicians and patients. This study also assessed the quality of evidence cited by chatbots to support their recommendations.

The primary objective of this study was to identify whether chatbots could successfully provide accurate clinical advice regarding colorectal cancer screening for both patients and clinicians according to guideline recommendations from key societies. Secondary objectives were to identify discrepancies in clinical advice provided to clinicians and patients and to assess the quality of the evidence cited by LLM-linked chatbots in providing screening recommendations.

2 Materials and methods

2.1 Query strategy

Nine patient cases were constructed based on expert input and existing colorectal cancer screening guidelines with various permutations of age and family history of colorectal cancer [10–14]. Four patient cases were considered to be "high-risk" with a family history of colorectal cancer, while five patients were deemed "average-risk." Over July 27th to 28th 2023, ChatGPT (GPT-3.5), Bing Chat, Google Bard, and Claude 2 were queried simultaneously using various prompts from a computer server in Hamilton, Ontario, Canada. This was done to identify generic responses such as legal disclaimers which would be barriers to obtaining meaningful information. Follow-up prompts were trialled during this time to identify whether these could be overcome. This information was combined with input from expert general surgeons to generate standardized prompts to query each chatbot platform about whether colorectal cancer screening is indicated for each patient case for clinicians (Supplementary Appendix 1).

If a chatbot did not produce a clinically meaningful response, structured follow-up prompts were applied. Prompts were reviewed by two study members to ensure clinical accuracy. The same study member entered the prompts for all chatbot platforms, while a second team member reviewed all prompts and responses to ensure that all chatbots were queried in a consistent and reliable manner. For relevant cases, chatbots were asked to report which screening test was indicated, the frequency of interval screening if the test were negative, and the next steps in management for patients with positive screening test results. Separately, the chatbots were queried with lay terminology to ask whether patients should receive screening for colon cancer (Supplementary Appendix 1). To mitigate the likelihood of biasing any given chatbot platform toward a specific guideline, no prompt contained any reference to a society or organization. Responses were prompted in a fresh chat session to mitigate the likelihood of biasing responses from prior discourse for clinician and patient advice, separately. For clinicians, chat discussion was started with the prompt, "I am a doctor." For patients, chat discussion was started with the prompt "Should I be screened for colon cancer?" to ensure that patients would be prompted for information regarding age and family history of colorectal cancer. When chatbots provided legal disclaimers in place of responding to the prompt, they were instructed with standardized follow-up prompts including: "I acknowledge this", "Tell me based on what I've told you", and "Tell me based on what you know" (Supplementary Appendix 1).

2.2 Analysis

Chatbot-generated screening advice for clinicians was compared to guidelines from the American Cancer Society (ACS), Canadian Cancer Society (CCS), U.S. Multi-Society Task Force (USMSTF) on Colorectal Cancer, and the United States Preventive Services Task Force (USPSTF). Clinician advice was compared to patient advice to assess for the presence of discrepancies. A data collection form designed a priori on Microsoft Excel was used to amalgamate response data. We used descriptive statistics for outcome reporting including counts and percentages. All analyses were performed using Microsoft Excel. Count data were based on the applicability of the given guideline recommendations to the patient case. If patient cases were not applicable to the guideline, then the outcome event was not included in the count (i.e.: in the denominator). Additionally, the Appraisal of Guidelines for Research & Evaluation (AGREE)-II tool was applied to assess the quality of the guidelines which chatbots cited to support their recommendations.

Table 1 Alignment of chatbot-generated recommendations with ACS colorectal cancer screening guidelines

Chatbot	ChatGPT		Bing Chat		Google Bard		Claude 2	
Clinician (CI) or Patient Inquiries (PI)	*CI*	*PI*	*CI*	*PI*	*CI*	*PI*	*CI*	*PI*
Total Frequency Aligned with ACS	3/4	3/4	2/4	1/4	3/4	2/4	1/4	2/4
Cases								
49 y/o, average risk	Yes	Yes	X	X	Yes	Yes	X	Yes
44 y/o, average risk	X	X	Yes	X	Yes	X	X	X
53 y/o, average risk	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
77 y/o, average risk	Yes	Yes	X	X	X	X	X	No

‡X = Did not give clinically meaningful advice

Table 2 Alignment of chatbot-generated recommendations with CCS colorectal cancer screening guidelines

Chatbot	ChatGPT		Bing Chat		Google Bard		Claude 2	
Clinician (CI) or Patient Inquiries (PI)	*CI*	*PI*	*CI*	*PI*	*CI*	*PI*	*CI*	*PI*
Total Frequency Aligned with CCS	2/4	2/4	2/4	1/4	2/4	1/4	1/4	1/4
Cases								
49 y/o, average risk	X	X	X	X	No	No	X	No
44 y/o, average risk	X	X	Yes	X	Yes	X	X	X
53 y/o, average risk	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
77 y/o, average risk	Yes	Yes	X	X	X	X	X	No

‡X = Did not give clinically meaningful advice

2.3 Response classification

Successful chatbot performance in providing accurate clinical advice was defined as responses which were aligned with major colorectal cancer screening guideline recommendations. In the case of clinical practice guideline recommendations in favour of a clinical action, advice from chatbots to proceed with this action, or advice indicating that proceeding with this action may be "reasonable" or "appropriate" were considered to align with guideline recommendations. When evaluating discrepancies between patient and clinician advice, recommendations were classified as discordant when a recommendation for or against a clinical action was given, while the second recommendation was for either the opposite action, to consult a physician, or that the decision would be made based on patient preferences and values. When chatbots provided responses to seek consultation from a medical professional, these responses were considered to not adhere with guideline recommendations. Advice generated from chatbots which were inaccurate according to clinical practice guidelines were considered not to align with guideline recommendations. When evaluating discrepancies between patient and clinician advice, recommendations were not classified as discordant when both recommendations were to consult a physician, or that the decision would be made based on patient preferences and values. Response evaluation was performed independently by two researchers. If discrepancies were noted, a third researcher was consulted. All researchers were trained on response evaluation through exposure to the above criteria and three pilot cases.

3 Results

3.1 Average risk screening

Table 1 demonstrates the alignment of LLM-linked chatbot advice with ACS recommendations. Accurate advice was given to clinicians for 3/4, 2/4, 3/4, and 1/4 cases for ChatGPT, Bing Chat, Google Bard, and Claude 2, respectively. Accurate advice was given to patients for 3/4, 1/4, 2/4, and 2/4 cases for ChatGPT, Bing Chat, Google Bard, and Claude 2, respectively (Table 1). When assessed using CCS recommendations, accurate clinician advice was given for 2/4, 2/4, 2/4, and 1/4 patient cases for ChatGPT, Bing Chat, Google Bard, and Claude 2, respectively. Accurate patient advice was given for 2/4, 1/4, 1/4, and 1/4 cases for ChatGPT, Bing Chat, Google Bard, and Claude 2, respectively (Table 2). When evaluated using USPSTF recommendations, clinician advice was given accurately to 3/4, 2/4, 3/4, and 1/4 patient cases for ChatGPT, Bing Chat, Google Bard, and Claude 2, respectively. Patient advice was appropriately generated for 3/4, 1/4, 2/4, and 2/4 cases for ChatGPT, Bing Chat, Google Bard, and Claude 2, respectively (Table 3). When performance was evaluated according to USMSTF recommendations, accurate clinician advice was given for 7/9, 5/9, 6/9, and 3/9 patient cases for ChatGPT, Bing Chat, Google Bard, and Claude 2, respectively. Patient advice was given accurately for 5/9, 4/9, 6/9, and 4/9 patient cases for ChatGPT, Bing Chat, Google Bard, and Claude 2, respectively (Table 4).

Discrepancies were given by ChatGPT for two of nine patient cases involving screening for a 77-year-old patient

Table 3 Alignment of chatbot-generated recommendations with USPSTF colorectal cancer screening guidelines

Chatbot	ChatGPT		Bing Chat		Google Bard		Claude 2	
Clinician (CI) or Patient Inquiries (PI)	*CI*	*PI*	*CI*	*PI*	*CI*	*PI*	*CI*	*PI*
Total Frequency Aligned with USPSTF	3/4	3/4	2/4	1/4	3/4	2/4	1/4	2/4
Cases								
49 y/o, average risk	Yes	Yes	X	X	Yes	Yes	X	Yes
44 y/o, average risk	†X	X	Yes	X	Yes	X	X	X
53 y/o, average risk	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
77 y/o, average risk	Yes	Yes	X	X	X	X	X	No

‡X = Did not give clinically meaningful advice

Table 4 Alignment of chatbot-generated recommendations with USMSTF colorectal cancer screening guidelines

Chatbot	ChatGPT		Bing Chat		Google Bard		Claude 2	
Clinician (CI) or Patient Inquiries (PI)	*CI*	*PI*	*CI*	*PI*	*CI*	*PI*	*CI*	*PI*
Total Frequency Aligned with USMSTF	7/9	5/9	5/9	4/9	6/9	6/9	3/9	4/9
Cases								
49 y/o, average risk	No	No	X	X	No	No	X	No
44 y/o, average risk	†X	X	Yes	X	Yes	X	X	X
53 y/o, average risk	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
77 y/o, average risk	Yes	Yes	X	X	X	X	X	No
44 y/o. Sister CRC at 59 y/o	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
44 y/o. Sister CRC at 61 y/o	Yes	Yes	Yes	Yes	Yes	Yes	No	No
44 y/o. Mother CRC at 59 y/o	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
77 y/o. Son CRC at 59 y/o	Yes	X	X	X	X	Yes	No	Yes
77 y/o. Son CRC at 61 y/o	Yes	X	X	X	Yes	Yes	No	No

‡X = Did not give clinically meaningful advice

with a family member diagnosed with colorectal cancer in both cases, as patients were not given clinically useful advice. Bing Chat provided discordant advice for 1/9 cases for clinicians and patients. For a 44-year-old patient with no additional risk factors, Bing Chat endorsed no screening to clinicians. In contrast, patients were not given clinically meaningful advice. Discordant advice between clinicians and patients was given by Google Bard for 2/9 (22.2%) cases. For a 44-year-old patient at average risk for colorectal cancer, Bard did not endorse screening to clinicians, but patients were not given useful information. For a 77-year-old patient at higher risk for colorectal cancer secondary to having a son diagnosed with colorectal cancer at 59 years old, Bard would not make a decision when prompted at a clinician level. However, patients were told that they qualified for colorectal cancer screening (Table 1). Discordant advice between clinicians and patients was given by Claude 2 for 3/9 (33.3%) cases. For an average risk patient at 49 years of age, clinicians did not receive useful guidance, while patients were told that they qualified for colorectal cancer screening. For a 77-year-old patient at average risk for colorectal cancer, clinicians did not receive clinically useful advice, while patients were told that they qualified for colorectal cancer screening. For a 77-year-old patient at high-risk for colorectal cancer secondary to having a son diagnosed with colorectal cancer at 59 years old, clinicians were told that the patient did not qualify for screening, while patients were told that they did qualify for colorectal cancer screening.

Supplementary Appendix 2 demonstrates the results of follow-up questions for a 53-year-old patient with no additional risk factors. ChatGPT provided discordant advice for all follow-up questions. For clinicians, ChatGPT recommended annual fecal immunochemical test (FIT) testing which aligns with USPSTF guidance, while Bing Chat recommended FIT testing every 2 years, which supports CCS guidelines. Google Bard and Claude 2 both recommended colonoscopy every 10 years, supported by USPSTF guidance. For patients, Bing Chat recommended FIT testing every 2 years, while all other chatbots recommended colonoscopy screening every 10 years (Table 2).

3.2 High risk screening

Supplementary Appendix 3 demonstrates chatbot recommendations for a 44-year-old patient at high risk for colorectal cancer secondary to having a sister diagnosed with colorectal cancer at 59 years of age. According to USMSTF guidelines, this patient should receive colorectal cancer screening beginning at age 40 years old (i.e., colonoscopy occurring at this patient's earliest convenience). However, ChatGPT and Google Bard recommended that this patient be screened at 49 years old and at 39 years old, respectively. Bing Chat and Claude 2 recommended

that the patient seek consultation from their physician. ChatGPT recommended that the patient receive a colonoscopy, aligning with USMSTF recommendations. However, Google Bard advised the clinician to send the patient for FIT testing, while Bing Chat and Claude 2 stated that the patient should seek further consultation from their physician. If this patient received a negative screening colonoscopy, USMSTF guidelines would suggest screening every 5 years. Only Google Bard produced clinician advice which adhered to this guidance. ChatGPT, Bing Chat, and Claude 2 recommended interval screening every 10 years (Table 3). Notably, clinicians were instructed to begin screening this patient at 39 years of age, while patients were instructed to begin screening at 40 years of age. Only ChatGPT avoided illogical recommendations for screening patients at a previous age. Google Bard provided clinician and patient advice which was discrepant for 3/4 (75.0%) follow-up prompts (Table 3).

Supplementary Appendix 4 illustrates chatbot advice for a 77-year-old patient at higher-risk of colorectal cancer secondary to having a son diagnosed with colorectal cancer at 59 years of age. For clinicians, ChatGPT advised to begin screening at 77 years old with colonoscopy at 5-year intervals, aligning with USMSTF recommendations. In contrast, Google Bard recommended to begin screening before the age of 76 years old using colonoscopy every 10 years. Bing Chat and Claude 2 elected not to provide clinician guidance for this patient. Notably, Google Bard instructed patients that they should begin screening for colorectal cancer at age 55 years old, contrasting with clinician advice (Table 4).

3.3 Quality of evidence

ChatGPT cited clinical recommendations from the ACS, as well as the USMSTF. Bing chat cited advice stemming from the CCS and USPSTF. Google Bard cited advice based on recommendations from the ACS and USPSTF. Claude 2 produced recommendations based on the USPSTF guidance. Supplementary Appendix 5 lists the AGREE-II scores for each guideline cited by the chatbots. AGREE-II scores for the ACS, USMSTF, and USPSTF guidelines were 83.3%, 25.0%, and 33.3%, respectively.

4 Discussion

This study applied rigorous methodology to use LLM-linked chatbots to provide clinical advice for both clinicians and patients based on clinical guideline recommendations. We applied colorectal cancer screening guidelines from major societies to assess the accuracy of chatbot-generated clinical advice for patients and clinicians for various patient cases. ChatGPT yielded the most successful performance and provided accurate screening advice more often than Bing Chat, Google Bard, and Claude 2 when assessed using the CCS, USPSTF, and USMSTF screening guidelines. However, no chatbot was able to provide accurate screening advice for all cases. Claude 2 gave the highest rate of discordant advice to clinicians and patients, but all chatbots provided inconsistent advice to both clinicians and patients. Chatbots provided responses from various societies including organizations from the USA, despite all queries being conducted in Canada. The inconsistent clinical advice generated by chatbots in this study outline the areas of future work in this area for clinician-researchers.

Previous studies posing basic, patient-level questions regarding cancer screening to ChatGPT, Bing, and Google Bard found that responses yielded inconsistent reliability [5, 8]. Similarly, no chatbot provided accurate clinical advice for all patient cases in our study. Rahsepar and colleagues found that ChatGPT provided accurate responses to 70.8% of questions regarding lung cancer screening [5], which parallels the rate of initially accurate responses to 77.8% of cases reported here. Rahsepar and colleagues found that ChatGPT was able to answer all questions, while Google Bard failed to answer one-fifth of questions posed to it [5]. By applying structured follow-up prompting questions established a priori, the chatbots produced a meaningful response to all prompts included in this study. Both ChatGPT and Bing Chat were able to provide clinical advice in favour of or against screening for most patient cases but defaulted to consulting a physician for nuanced cases involving patients above the age of colorectal cancer screening with a higher risk of colorectal cancer. Google Bard and Claude 2 defaulted to consulting a physician even for straightforward patient cases with average risk patients that met age criteria for screening. Only Claude 2 provided clinical screening advice which directly contradicted USMSTF recommendations, advising both clinicians and patients not to screen in certain patient cases, despite screening being indicated according to USMSTF. This discrepancy has major implications on patient care, as these patients may have missed an opportunity to prevent the development of a malignancy. Additionally, both Google Bard and Claude 2 struggled to provide accurate recommendations for high-risk patients and advised to begin screening at an age several years prior to the age of the patient being presented. However, as both Google Bard and Claude 2 were recently released in 2023 as experimental models (applicable even more so to Claude 2 which launched the same month that this study was conducted), this higher rate of inconsistent responses is not surprising.

Prior studies have not addressed the consistency of chatbot health advice for clinicians and patients. Discrepant

recommendations between clinicians and patients was most prominent in follow-up prompts for specific information regarding the age to begin screening, the type of screening test indicated, and interval screening. Prior studies in this area have not used follow-up prompts [5, 8]. However, as chatbots may not answer all initial prompts [5], the use of structured prompts and follow-up prompts across chatbots in a consistent manner is an effective approach to obtain a response, particularly to bypass disclaimers for clinical advice. ChatGPT did not pull data from USPSTF guidelines, which were most recently published in May 2021 [12], whereas Bing Chat did. It is noteworthy that Bing Chat is grounded in web data, whereas ChatGPT was limited to data prior to September 2021 [3]. While ChatGPT should theoretically have been able to access USPSTF guidelines, from the authors' anecdotal experience, publications in 2021 seem to be relatively more difficult for ChatGPT to access. This remains a major advantage to chatbots such as Bing Chat, and though ChatGPT provided the most consistent responses in this study, researchers should turn their attention toward how this detail contributes toward the future development of these chatbots. Additionally, Bing Chat based primary recommendations on data from the CCS, which is logical considering the queries were conducted in Canada. The impact of location of search on results obtained from chatbot queries remains to be established. Additionally, chatbots including ChatGPT, Google Bard, and Claude 2 refer to sources of their information inconsistently, whereas Bing Chat provides footnotes with citations for each response [3]. However, ChatGPT and Google Bard supported some of their recommendations using high-quality guidelines, whereas Bing Chat and Claude 2 used moderate-quality guidelines to substantiate their advice.

With the wide popularity and availability of ChatGPT and other chatbots, both patients and healthcare providers must take note of the pitfalls which currently exist when using chatbots to answer health questions. It must be emphasized that advice given to clinicians and patients are not always aligned, even when chatbots are queried with consistent phrasing and terminology. Patients and clinicians may take further interest in the impact of location of search on chatbot results, particularly as chatbots are not yet consistent in citing the source of their responses. Developers of future chatbot platforms should note the importance of citing sources for each response as with Bing Chat, particularly as research turns toward the use of chatbots to address problems in medicine. Furthermore, researchers must note the ever-changing climate of chatbots when assessing the performance of chatbots, referred to here as Chatbot Assessment Studies. While chatbots will inevitably improve their accuracy and efficiency, all stakeholders must hold basic knowledge surrounding how these chatbots function, a prime example being in the case of ChatGPT being limited to data prior to September 2021. Chatbot Assessment Studies are unique in that prompt content and length will produce varying responses from chatbots, while responses also change over time. Additionally, the use of objective measures to evaluate chatbot output is needed to improve the internal validity of Chatbot Assessment Studies. With the anticipated exponential increase of future studies in this space, there is a need for rigorous reporting standards for studies assessing chatbot output, whether for clinical advice, or other purposes.

Strengths of this study include its rigorous and transparent methodology, as well as the use of objective measures of performance through the direct applicability of chatbot advice to clinical practice guideline recommendations. Limitations exist in this study. Firstly, all chatbots were queried at a single point in time (July 2023). As LLMs evolve rapidly, the responses generated may no longer be relevant for future versions and/or models. Moreover, not all chatbots received follow-up prompts for clarification which may bias results, though structured follow-up prompts were prepared in advance for specific situations to mitigate the inability of chatbots to produce an answer, including to acknowledge legal disclaimers. Additionally, the unpaid version of ChatGPT was used rather than the paid version which has GPT-4 similarly to Bing Chat, which may introduce inherent bias in the quality of responses produced. However, this was done for the purpose of maintaining external validity. Furthermore, the ability of LLMs to consistently provide the same clinical advice to the same repeated prompts has yet to be fully elucidated, and our results must be interpreted accordingly. Importantly, there was a lack of representation of Canadian practice guidelines supporting recommendations from these chatbots. Clinicians and patients should be aware of the applicability of the recommendations based on screening strategies tailored for their location.

5 Conclusion

LLM-linked chatbots hold potential in their ability to successfully provide clinicians and patients with accurate clinical advice regarding colorectal cancer screening. However, current chatbot technology provides screening recommendations with inconsistent accuracy, which could result in screen-eligible patients not seeking colorectal cancer screening. Even with standardized prompting, advice given to clinicians may differ from that given to patients for the same clinical scenarios. Clinicians must educate patients on the pitfalls of using these platforms for health advice. Standardized terminology must be identified for the purpose

of improving the consistency and quality of clinician and patient advice produced by chatbots.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12553-024-00836-9.

Acknowledgements Not applicable.

Author contributions All authors contributed to the study conception and design. SA, JM, HA, VB, KR, and CE provided expert guidance on the study methodology. Material preparation, data collection and analysis were performed by BH, TM, MO, and Yung Lee. The first draft of the manuscript was written by Bright Huo and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This study was unfunded.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Ethical approval Not applicable.

Consent to participate Not applicable.

Conflicts of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29:1930–40.
2. Bhattacharya K, Bhattacharya AS, Bhattacharya N, Yagnik VD, Garg P, Kumar S. ChatGPT in Surgical Practice—a New Kid on the Block. Indian J Surg. 2023. https://doi.org/10.1007/s12262-023-03727-x.
3. Rudolph J, Tan S, Tan S. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. JALT. 2023. https://doi.org/10.37074/jalt.2023.6.1.23.
4. Ferdush J, Begum M, Hossain ST. ChatGPT and clinical decision support: scope, application, and limitations. Ann Biomed Eng. 2023. https://doi.org/10.1007/s10439-023-03329-4.
5. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung Cancer questions: ChatGPT vs Google Bard. Radiology. 2023;307:e230922.
6. Ayers JW, Leas EC, Dredze M, Allem JP, Grabowski JG, Hill L. Clinicians' perceptions of barriers to avoiding Inappropriate Imaging for LowBack Pain— Knowing is not enough. JAMA Intern Med. 2016;176:1865–6.
7. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic surgery advice and counseling from Artificial Intelligence: a Rhinoplasty Consultation with ChatGPT. Aesthetic Plast Surg. 2023;47:1985–93.
8. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast Cancer Prevention and Screening recommendations provided by ChatGPT. Radiology. 2023;307:e230424.
9. Fu S, Wang L, Moon S, Zong N, He H, Pejaver V, et al. Recommended practices and ethical considerations for natural language processing-assisted observational research: a scoping review. Clin Transl Sci. 2023;16:398–411.
10. Kalyta A, De Vera MA, Peacock S, Telford JJ, Brown CJ, Donnellan F, et al. Canadian colorectal cancer screening guidelines: do they need an update given changing incidence and global practice patterns? Curr Oncol. 2021;28:1558–70.
11. Wolf AMD, Fontham ETH, Church TR, Flowers CR, Guerra CE, LaMonte SJ, et al. Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. CA Cancer J Clin. 2018;68:250–81.
12. Davidson KW, Barry MJ, Mangione CM, Cabana M, Caughey AB, Davis EM, et al. Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. JAMA. 2021;325:1965–77.
13. Rex DK, Boland CR, Dominitz JA, Giardiello FM, Johnson DA, Kaltenbach T, et al. Colorectal Cancer screening: recommendations for Physicians and patients from the U.S. Multi-society Task Force on Colorectal Cancer. Am J Gastroenterol. 2017;112:1016–30.
14. Bacchus C, Dunfield L, Gorber S, Holmes N, Birtwhistle R, Dickinson J, et al. Recommendations on screening for colorectal cancer in primary care. Can Med Assoc J. 2016;188:340–8.