

The Ability of Artificial Intelligence LLM-Linked Chatbots to Generate Recommendations for the Surgical Management of GERD Using SAGES Guidelines

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]

[1]McMaster University, Hamilton, Canada; [2]Università Politecnica delle Marche, Ancona, Italy; [3]Papageorgiou General Hospital, Thessaloniki, Greece; [4]Universidad Complutense de Madrid, Madrid, Spain; [5]Phelix AI, Hamilton, Canada.

McMaster University Surgery

Introduction

Large language model (LLM)-linked chatbots have tremendous potential in healthcare, but their clinical accuracy is poorly understood.

Objectives

To evaluate the accuracy of colorectal cancer screening advice given by Large Language Model (LLM)-linked chatbots for both clinicians and patients.

Methods

Nine patient cases based on eligibility criteria from conventional colorectal cancer screening guidelines.

LLMs queried Jul 27th–28th, 2023 with standardized prompts for colorectal cancer screening recommendations.

Responses generated for patients and separately, for clinician inquiries.

Evaluation based on alignment with guideline recommendations.

Results

Table 1. Alignment of Chatbot-Generated Recommendations with 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

Results

ChatGPT demonstrated the ability to exercise clinical reasoning.

Conflicting clinician and patient advice was given for 2/9 (22.2%), 1/9 (11.1%), 2/9 (22.2%), and 3/9 (33.3%) cases with ChatGPT, Bing Chat, Google Bard, and Claude 2, respectively.

Conclusions

LLM-based chatbots have promise in providing colorectal cancer screening recommendations, but still perform with inconsistent accuracy.

Clinicians and patients must be aware of the pitfalls of using these platforms for health advice, as their reliability is unknown.