

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

Bright Huo[1], Elisa Calabrese,[2] Sunjay Kumar,[3] Romeo C. Ignacio,[4] Rodolfo Oviedo,[5-7] Imran Hassan,[8] Bethany Slater,[9] Danielle S. Walsh,[10] Wesley Vosburg.[11]

[1]McMaster University, Canada; [2]University of California South California, USA; [3]Thomas Jefferson University Hospital, USA; [4]University of California, USA; [5]Nacogdoches Center for Metabolic and Weight Loss Surgery, USA; [6]University of Houston Tilman J. Fertitta Family College of Medicine, USA; [7]Sam Houston State University College of Osteopathic Medicine, USA; [8]University of Iowa, USA; [9]University of Chicago, USA; [10]University of Kentucky, USA; [11]Harvard University, USA

McMaster University Surgery

Introduction

Large language model (LLM)-linked chatbots may be an efficient source of clinical recommendations for healthcare providers and patients.

Objectives

To evaluate the performance of LLM-linked chatbots in providing recommendations for the surgical management of GERD.

Methods

Seven patient cases based on KQs in SAGES guidelines for the surgical treatment of GERD.

perplexity

BING CHAT

Bard

LLMs queried Nov 16th, 2023 with standardized prompts for advice on the surgical management of GERD.

Responses generated for patients and separately, for clinician inquiries

Evaluation based on alignment with guideline recommendations.

Results

Table 1. Alignment of LLM Responses with SAGES Guideline Recommendations for Adults.

	Chatbot	ChatGPT-4		Bing Chat		Google Bard		Perplexity		SAGES
	Surgeons (S) or Patient (P) Inquiries	*S*	*P*	*S*	*P*	*S*	*P*	*S*	*P*	
	Recommendations Aligned with SAGES	5/7	3/5	3/7	2/5	6/7	4/5	3/7	1/5	-
	Certainty Aligned with SAGES	4/7		2/7		3/7		0/7		-
	Case	49 y/o on pantoprazole 40mg po BID x 2 years for refractory GERD, with worsening reflux symptoms.								-
#	Clinical Questions									-
1	Medication vs surgery?	Y	Y	Y	Y	Y	Y	Y	Y	Surgery
	Certainty?	N		N		N		N		Very low
2	Robotics vs laparoscopy - PPI?	Y	X*	N	N	Y	Y	N	N	Robot
	Certainty?	Y		N		Y		N		Low
3	Robotics vs laparoscopy - complications?	N	N	N	N	N	N	N	N	Laparoscopy
	Certainty?	N		N		N		N		Low
4	Complete vs partial repair - symptoms?	N	Y	Y	N	Y	Y	N	N	Complete
	Certainty?	N		N		N		N		Low
5	Complete vs partial repair - dysphagia?	Y	Y	Y	Y	Y	Y	Y	N	Partial
	Certainty?	Y		N		N		N		Low
6	Dissection of gastric vessels - symptoms?	Y		N		Y		Y		Divide
	Certainty?	Y		Y		Y		N		Very low
7	Dissection of gastric vessels - gas bloat?	Y		N		Y		Y		Don't divide
	Certainty?	Y		Y		Y		N		Very low

Results

In a child, surgeons given accurate advice for 2/3 (ChatGPT-4.0) 3/3 (Bing Chat) 3/3)Google Bard), and 2/3 KQs by Perplexity

Patients given appropriate guidance for 2/2 (ChatGPT-4.0), 2/2 (Bing Chat), 1/2 (Google Bard,) and 1/2 KQs by Perplexity

Conclusions

Gastrointestinal surgeons, gastroenterologists, and patients should recognize both the promise and pitfalls of LLM's when utilized for advice on surgical management of GERD.

Additional training of LLM's using evidence-based health information is needed.

LLM-based research must move toward patient-centeredness.