Comparing Large Language Models Accuracy in Following Interval Surveillance Colonoscopy Guidelines

Olufemi Osikoya, MD[1] and Gregory Brennan, MD[1,2]
[1]Department of Internal Medicine, Medical City Arlington, Arlington, TX; [2]GI Alliance, Mansfield, TX

Medical City Arlington

Introduction

- Large language models (LLMs) such as ChatGPT and Google Bard, have shown promise in clinical workflows such as pathology results letters
- Aim was to test whether LLMs could provide appropriate surveillance recommendations based on current guidelines from the US multi-society task force for post polypectomy colonoscopy follow-up.
- Compare the accuracy of ChatGPT 3.5, ChatGPT 4, and Google Bard in providing appropriate interval surveillance recommendations

- Seventeen different post polypectomy surveillance queries and responses were analyzed (correct, partially correct, incorrect) compared to USMSTF guidelines
- Example prompt "Write a patient pathology result letter after a colonoscopy with one tubular adenoma polyp (< 10mm) resected. Include recommendations for when the next surveillance colonoscopy should be completed"

Google Bard was more accurate at following the USMSTF guidelines compared to ChatGPT 3.5 and ChatGPT4

Colonoscopy Finding	USMSTF Interval Surveillance Colonoscopy Recommendation	ChatGPT 3.5	ChatGPT 4	Google Bard
1 tubular adenoma < 10 mm	7- 10 years	3-5 years	5-10 years	7-10 years
3 tubular adenomas < 10 mm	3- 5 years	3 years	3 years	3-5 years
5 tubular adenomas < 10 mm	3 years	1-3 years	3 years	3 years
>10 tubular adenomas	1 year	1 year	1 year	3 years
One or more Adenoma > 10 mm	3 years	3 years	3 years	3 years
Adenoma with tubulovillous or villous histology	3 years	1-3 years	3 years	3 years
Adenoma with high-grade dysplasia	3 years	3-6 months	3 years	3 years
Piecemeal resection of adenoma > 20 mm	6 months	3-6 months	6 months	6 months
10 hyperplastic polyps < 10 mm (rectum or sigmoid)	10 years	5 years	3-5 years	10 years
1 SSP < 10 mm	5-10 years	5-10 years	5-10 years	7-10 years
3–4 SSPs < 10 mm	3-5 years	3 years	3 years	3-5 years
5–10 SSPs < 10 mm	3 years	1-3 years	3 years	3 years
SSP > 10 mm	3 years	3 years	3 years	3 years
SSP with dysplasia	3 years	6-12 months	3 years	3 years
Hyperplastic polyp > 10 mm	3-5 years	5 years	3-5 years	5 years
Traditional serrated adenoma	3 years	3-5 years	3 years	3-5 years
Piecemeal resection of SSP >20 mm	6 months	6-12 months	2-6 months	6 months

Table 1. Comparison of accuracy of large language models in generating appropriate interval surveillance colonoscopy recommendations. Green= correct recommendation. Orange = partially correct recommendation. Red= incorrect recommendation

Results

- Google Bard provided the most correct responses and the least incorrect responses.
- Bard provided correct recommendations in 76% of queries (13/17), partially correct recommendations in 18% of queries (3/17) and incorrect recommendations in 6% of queries (1/17).
- ChatGPT 4 provided correct recommendations in 70% of queries (12/17), partially correct recommendations in 24% of queries (4/17) and incorrect recommendations in 6% of queries (1/17).
- ChatGPT 3.5 provided the most incorrect recommendations at 24% (4/17).

Conclusion

- Bard provided the most correct rec; connectivity to the internet.
- Bard and ChatGPT4 referenced the USMSTF guidelines. ChatGPT 3.5 had no refs.
- ChatGPT 4 also occasionally referenced the British Society of Gastroenterology (BSG) and European Society of Gastrointestinal Endoscopy (ESGE).
- Overall, partially correct recommendations were common in all LLMs.
- Using LLMs shows promise but, their current accuracy limits real world adoption.

References

1. Ayers, John W., et al. "Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum." JAMA Internal Medicine, vol. 183, no. 6, 1 June 2023, p. 589. https://doi.org/10.1001/jamainternmed.2023.1838.
2. Ge, Jin, and Jennifer C. Lai. "Artificial intelligence-based text generators in Hepatology: ChatGPT is just the beginning." Hepatology Communications, vol. 7, no. 4, 24 Mar. 2023, https://doi.org/10.1097/hc9.0000000000000097.
3. Gupta, Samir, et al. "Recommendations for follow-up after colonoscopy and polypectomy: A consensus update by the US Multi-Society task force on colorectal cancer." Gastrointestinal Endoscopy, vol. 91, no. 3, Mar. 2020, https://doi.org/10.1016/j.gie.2020.01.014.

This research was supported (in whole or in part) by HCA Healthcare and/or an HCA Healthcare affiliated entity. The views expressed in this publication represent those of the author(s) and do not necessarily represent the official views of HCA Healthcare or any of its affiliated entities.

HCA Healthcare℠