DR GOOGLE TO DR CHATGPT: ASSESSING THE CONTENT AND QUALITY OF ARTIFICIAL INTELLIGENCE-GENERATED MEDICAL INFORMATION ON APPENDICITIS

Yazid K Ghanem, MD MCh [1, 2]; Armuan Rouhi, BA [3]; Ammr Al-Houssan MD [4]; Zena Saleh, MD [1, 2], Matthew Moccia, MD [1, 2], Hansa Joshi, MD [1, 2], Kristoffel Dumon, MD [4]; Young Hong, MD [1, 2] ; Francis Spitz, MD [1, 2]; Amit R Joshi, MD [1, 2]; Michael Kwiatt, MD [1, 2]

[1] Department of Surgery, Cooper University Hospital, Camden, NJ

[2] Cooper Medical School of Rowan University, Camden, NJ

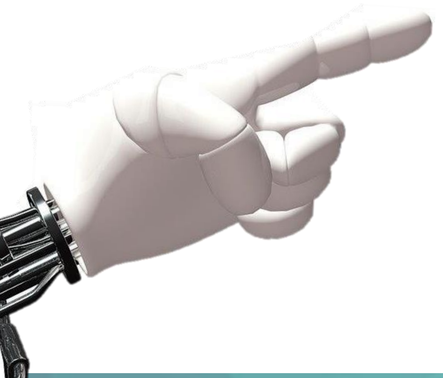
[3] Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

[4] Department of Surgery, University of Connecticut, Hartford, CT



INTRODUCTION

- >70% of adults use the internet for their health-related questions. [1]
- Existing online information is of *variable quality and difficult readability*. [2]
- **Poor health literacy** associated with worse surgical patient outcomes. [3]
- **Study Objective:** To assess the quality and readability of Artificial Intelligence (AI)-generated medical information about appendicitis.



METHODOLOGY

- **Four widely available AI chatbots were chosen:**
- *ChatGPT 3.5* (OpenAI, free), *ChatGPT 4* (OpenAI, subscription), *Bard* (Google, free), *and Claude-2* (Anthropic, free).
- **Four Standardized Prompts used:**
  - *"Please tell me about appendicitis"*
  - *"Please tell me more"*
  - *"What are the side effects and complications of the treatments?"*
  - *"Please provide a list of sources"*
- A **Modified DISCERN instrument was** used to assess quality, scored by four physicians independently, blinded to the identity of the chatbots.
- Readability was assessed using the **Flesch Kincaid Grade Level** and **Flesch Reading Ease**.

RESULTS

1. **Quality Analysis:** Bard, ChatGPT-3.5, ChatGPT-4 higher quality than Claude-2 ($p<0.001$)
2. All chatbots recommended **consulting a physician**, except Claude-2
3. **Readability:** Overall difficult readability, at the high-school student to a college graduate level
4. **Bard** provided verifiable sources **Claude-2 provided** fabricated sources **ChatGPT-3.5 and ChatGPT-4 did not provide any sources**

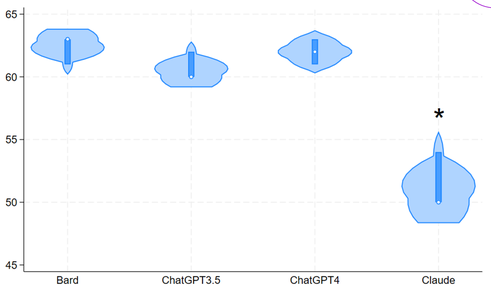


Figure 1: Violin plot of quality scores by artificial intelligence platform. *Indicates significance ($p < 0.001$)

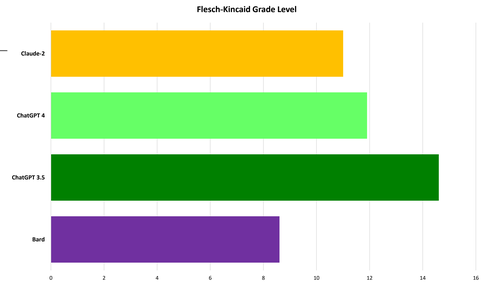


Figure 2: Readability scores (Flesch Kincaid Grade Level), showing Bard with the easiest readability, and ChatGPT-3.5 with the most difficult readability.

CONCLUSIONS

- AI produces high quality content about appendicitis
- Questionable credibility (except Bard)
- Challenging readability
- AI can be hugely powerful and can revolutionize patient care

REFERENCES

1. Link E, Baumann E. Nutzung von Gesundheitsinformationen im Internet: personenbezogene und motivationale Einflussfaktoren. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz.* " [Use of health information on the internet: personal and motivational influencing factors] 2020;63(6):681-689. doi:10.1007/s00103-020-03144-5
2. Rouhi AD, Ghanem YK, Hoeltzel GD, et al. Online Resources for Patients Considering Hiatal Hernia Repair: a Quality and Readability Analysis. *Journal of Gastrointestinal Surgery*. 2022;27(3):598-600. doi:10.1007/s11605-022-05460-4
3. Rouhi AD, Ghanem YK, Hoeltzel GD, et al. Quality and readability of online patient information on adolescent bariatric surgery. *Obesity Surgery*. 2022;33(1):397-399. doi:10.1007/s11695-022-06385-2
4. Trutner Z, Furlough K, Martinez AB, et al. Is health literacy associated with surgical outcomes? A Systematic review. *Journal of Surgical Research*. 2023;291:720-733. doi:10.1016/j.jss.2023.06.044

Cooper Medical School of Rowan University

Cleveland