Disclosures: I (and my co-authors) have nothing to disclose.

Using text network analysis to assess if Chat GPT generates comparable responses to Google Sources when asked questions related to Gastrointestinal Pathologies

Shankar S Thiru, BS[1]; Matthew C Chiang, BS[2]; Margaret E Greer, BS[1]; Thomas P Stirrat, BS[1], Karim S Trad, MD[3]
[1]Georgetown University School of Medicine, [2] DataChat Inc., [3]MedStar Georgetown University Hospital, Department of Surgery

MedStar Georgetown University Hospital

INTRODUCTION

- Role of patient has evolved from more passive to active participant
- Chat GPT (OpenAI) is a language learning model (LLM) that takes user-inputted text and generates a narrative response
- ChatGPT increasingly relied upon as alternative to Google for health searches
- **Purpose:** To compare ChatGPT 4o mini to Google web sources for GI pathology queries
- **Hypothesis:** ChatGPT 4o mini would be similar to Google web sources for GI pathology information

METHODS

Three queried questions of varying medical consensus:
(1) What causes appendicitis? (high)
(2) What causes heartburn? (medium)
(3) Fecal microbiota transplant for IBS? (low)

Study Procedure:
- ChatGPT asked each question 20x
- Text network analysis was performed, calculating cosine comparison of text frequency inverse-document frequency (TF-IDF)
- Statistical reference (p=0.001) was generated using random text samples with same length distribution as Google searches
- Analyzed similarities of ChatGPT and Google web sources by source type

Outcome Measures:
- TF-IDF for individual sources, TF-IDF for source type

$$ChatGPT\ node: v_i^*$$

$$Website\ node: v_j$$

$$Element-wise\ mean\ of\ weights: w_j = \frac{1}{N*M}\sum_{i=1}^{N}\sum_{j=1}^{M} e_{ij}$$

Figure 1. Mathematical schema for generating the text network of vectors. Each text sample is represented as a node. Connections between nodes are weighted based on strength of similarity, which is determined by the cosine similarity score.

** ChatGPT Responses were dominated by top Google search results**

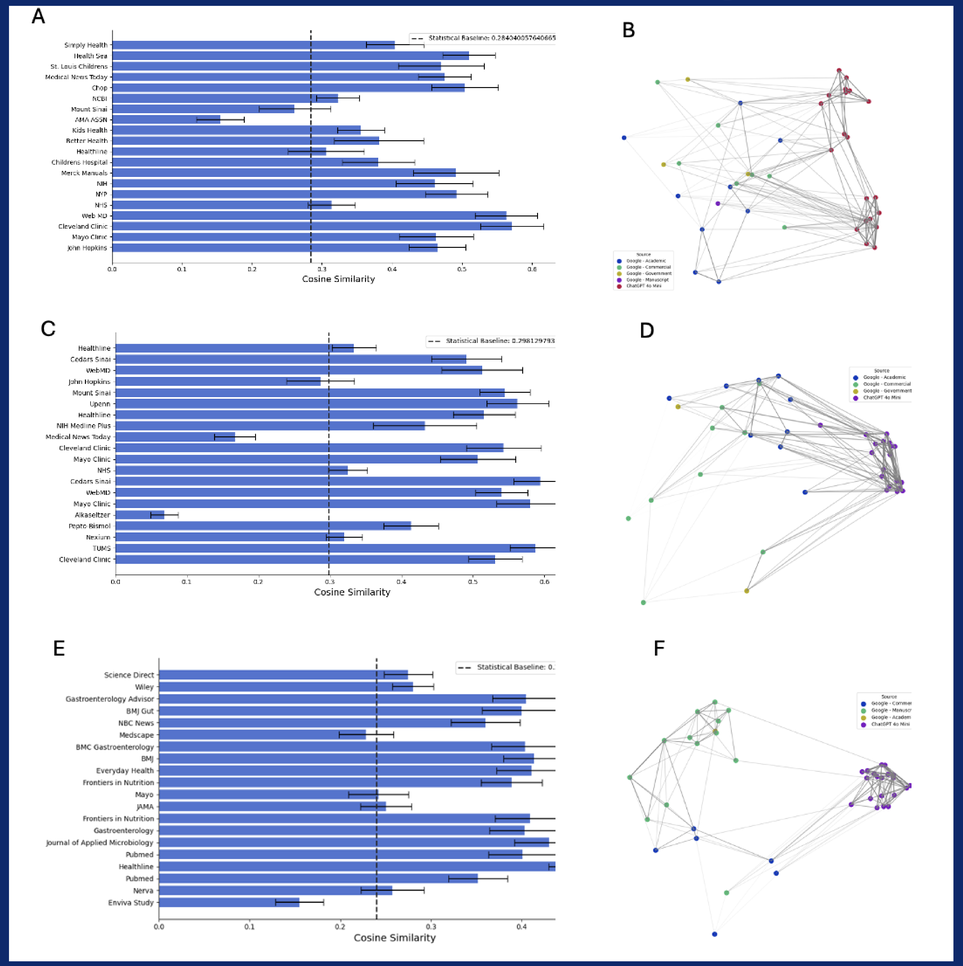


Figure 2. The TF-IDF similarity and text networks for individualized websites and aggregate ChatGPT responses for (A-B) **appendicitis**, (C-D) **heartburn**, (E-F) **fecal biota transplant** are shown. The statistical baseline was determined from textual similarity to a random generated text.

RESULTS

Cohort Characteristics
- n = 3
- Reference TF-IDF score of randomly generated text:
  - 1st question: 0.284
  - 2nd question: 0.298
  - 3rd question: 0.240

TF-IDF Similarities by Source Type
- Highest Medical Consensus
  - **Commercial: TF-IDF 0.445±0.096**
  - Academic: TF-IDF 0.419±0.133
  - Government: TF-IDF 0.386±0.079
  - Manuscript: TF-IDF 0.322±0.030
- Medium Consensus
  - Commercial: TF-IDF 0.384±0.172
  - **Academic: TF-IDF 0.516±0.097**
  - Government: TF-IDF 0.379±0.077
- Lowest Medical Consensus
  - Commercial: TF-IDF 0.313±0.073
  - **Academic: TF-IDF 0.430±0.113**
  - Manuscript: TF-IDF 0.356±0.038

DISCUSSION

- **ChatGPT responses are similar to individual Google search results for gastrointestinal pathologies, but the distribution of source information can vary based on the topic's level of consensus.**
- Only one similarity analysis was used
  - Future research could use other analysis (eg word2vec, bert vectorization)
- ChatGPT provides easily understandable information that patients have reportedly appreciated.
- **Conclusion: ChatGPT seems to have a role in patient online health information acquisition with the caveat that its utility may vary based on existing evidence of the topic**