Responsiveness of Online Generative AI with Large Language Models in Asian American-Specific Cancer Screening Inquiries: NLP Analysis

Ryan Suk PhD[1], Milkie Vu PhD[2], Alec Zhang[3], Alexander C. Tsai MD PhD[4]

1 Emory University (SON & Dept of Health Policy and Management SPH), Winship Cancer Institute; 2 Northwestern University (Dept of Preventive Medicine SOM); 3 Northwestern University (CAS); 4 Harvard University (SOM & Dept of Epidemiology SPH); Massachusetts General Hospital

Introduction

- ❑ The integration of Generative AI through large language models (LLMs) has influenced recent trends in information-seeking behaviors, particularly in health-related inquiries.
- ❑ While research has started to evaluate the quality of training data used in these online AI platforms, there is still a significant gap in understanding how these data translate to their responsiveness across diverse populations.
- ❑ This study explores the accuracy, bias, and cultural appropriateness of LLM-generated cancer screening information offered to Asian Americans compared with the general population.

Methods

Data source: Publicly available Large Language Model (LLM) platforms (ChatGPT, CoPilot, Gemini)

Prompts: Identical prompts were used to elicit cancer screening recommendations. Prompts were designed to self-identify the user, by gender, as either Asian American, of a specific Asian origin group (Chinese, Filipino, Indian, Japanese, Korean, or Vietnamese), or in race-neutral terms (e.g., 'I am a woman').

Iterations: Conducted testing at three timepoints (April-May 2024) to improve reliability, with all prompts run across platforms on the same day.

Natural Language Processing (NLP) analyses: We used OpenAI's large NLP model (ChatGPT, versions 4 and 3.5 Turbo) through its application programming interface (**API**). All data were analyzed using Python (ver. 3.13) and Pandas. **1) Bias Detection:** Identified and explained potential biases related to ethnicity and gender. **2) Sentiment Analysis:** Categorized responses as positive, neutral, or negative. **3) Compliance Assessment:** Evaluated adherence to established medical guidelines**. 4) Parameter settings:** Set temperature to 0.3 for bias detection and compliance assessment to promote deterministic and focused responses, minimizing randomness. For sentiment analysis, it was set to 0.0 to ensure consistency. **5) Qualitative Summarization:** Aggregated analysis results by AI platform and prompt type (ethnicity-specific vs. race-neutral) and generated narrative summaries.

What is OpenAI Application Programming Interface (API)?

The OpenAI API provides access to advanced language models developed by OpenAI, such as GPT-4, enabling developers to integrate natural language processing (NLP) capabilities into their applications through HTTP requests.

Results

***Quote**:"..For example, the mention of liver cancer screening in the context of chronic hepatitis B virus (HBV) infection could be seen as a stereotype, as HBV is more prevalent in certain Asian populations…."*

	ChatGPT 3.5	CoPilot	Gemini
Prompts specific to Asian American & Asian origin groups	Overall alignment with clinical guidelines, with minor deviations. Liver, stomach, lung cancer (in women), and chronic HBV infection mentioned additionally for elevated risk in Asian groups. Neutral sentiment. Emphasized individualized risk assessment and discussions with healthcare providers.	Mostly compliant with guidelines, with some deviations (old guidelines). Additionally, gastric and liver cancer mentioned. High incidence of non-smoking-related lung cancer in Asian women. Positive sentiment. Occasionally mislabels Indian American as American Indian.	Mostly aligns with guidelines. Emphasis on individualized decision-making, based on discussions with healthcare providers. Additionally, stomach and liver cancers mentioned. Neutral sentiment. Emphasis on regular screening.
Race-neutral prompts	Largely compliant with guidelines, with minor deviations. Neutral sentiment. Mainly colorectal, skin, prostate, testicular, breast, cervical, and smoking-related lung cancers.	Emphasized the importance of regular cancer screenings. Positive sentiment. Largely compliant with guidelines. Mainly colorectal, prostate, lung, breast, cervical, skin, endometrial cancers.	Emphasis on early detection and tailored screening for at-risk individuals. Neutral sentiment. Mainly colorectal, testicular, prostate, breast, and cervical cancers.

Discussion

- ❑ Online LLMs can offer tailored cancer screening advice, potentially enhancing the effectiveness of health interventions for specific populations.
- ❑ However, there is a potential risk of perpetuating biases and/or stigma if these variations are not presented and explained in a **culturally sensitive and scientifically accurate manner, with appropriate context**.
- ❑ **Developers' efforts:** Developers should focus on curating diverse and representative datasets to minimize bias in AI models and implement robust fairness metrics. AI platforms should provide users with contextual prompts or warnings when engaging with AI-generated healthcare content.
- ❑ **Policymakers' ethical guidelines:** Regulations mandating transparency and accountability in AI development, particularly for publicly available models that influence health and social issues, may be essential to ensure ethical and responsible use.
- ❑ **User education:** Users should be educated on how to critically evaluate AI-generated information, and public campaigns should promote media literacy, teaching users to verify AI outputs against trusted sources. Users should be encouraged to treat AI-generated health advice as preliminary information and always consult healthcare providers for complex or personalized medical decisions.