Public Use of Large Language Models to Obtain Information About Cardiovascular Disease

VISHALA MISHRA and JOSEPH P. DEXTER

ABSTRACT

Background: Despite the availability of multiple effective therapies and strong practice guidelines, concerning gaps persist in the implementation of cardiovascular disease (CVD) prevention strategies. These gaps are due in part to ineffective communication, patient hesitancy, and misinformation about important topics. Current information sources, such as online patient educational materials from government or academic sources and social media, often have limitations related to accessibility and information quality. In addition to their increasing prominence in clinical workflows, large language models (LLMs) are also rapidly becoming a key informal source of health information for the general public; as such, they offer a promising opportunity to improve patient education and public health communication regarding CVD prevention.

Research Question: Most research about health-related applications of LLMs has involved experts developing careful prompts for well-defined tasks. These data are not generalizable to complex real-world situations, in which users with limited prior exposure to generative AI may compose vague or contradictory prompts, and there is an urgent need for more evaluation studies that involve human ratings of realistic tasks. To address this gap, we conducted a two-part user interaction study with ChatGPT in July 2024.

Methods: For this study, we recruited participants through Prolific, with quota sampling to achieve samples representative of the U.S. population for age, race, ethnicity, and education (N = 388). Participants were instructed to use ChatGPT to complete two tasks: i) finding health information about any topic of interest, and ii) structured exploration of two CVD prevention topics (what to do if someone is having a heart attack and clinical associations between diabetes and hypertension). After completing each task, participants copied a complete transcript of their ChatGPT session into the survey form and answered impression questions about the LLM-generated information they received.

Results: From the first set of transcripts, we tabulated which topics related to CVD were of greatest interest to our participants and the kinds of prompts they used to explore them (e.g., explicit requests for tailored or plain language information). For the structured transcripts, we developed a four-part rubric for clinician evaluation encompassing i) information quality (e.g., disagreements with medical consensus and off-topic information), ii) fidelity of user prompts, iii) impressions of clarity and understandability, and iv) completeness. To quantify completeness, we used two custom eight-point rubrics, one for the heart attack scenario (topic 1) and one for the diabetes/hypertension scenario (topic 2). Members of the study team coded 250 transcripts chosen at random. The median completeness score was 4 (IQR 2) for topic 1 and 7 (IQR 1) for topic 2; transcripts for topic 2 were rated as significantly more complete ($p < 0.001$ by a two-way Wilcoxon rank-sum test). Automated completeness scores generated by GPT-5 agreed substantially with the expert ratings (weighted $\kappa = 0.861$ for topic 1, 0.813 for topic 2).

Conclusions: The gaps in completeness identified from both manual and automated coding suggest the importance of rigorous evaluation of LLM usage by the public and of physician-in-the-loop workflows for precision health communication.

Implications: Future research should evaluate medical AI technologies under both ideal and average use conditions.

STUDY OVERVIEW

Participants

N = 388 online participants
Quota sampled to be representative of U.S. population
(age, race, ethnicity, education)

Prolific

Tasks

Use ChatGPT:

- For unstructured exploration of any health topic
- To learn about bystander response to heart attack
- To learn about association between hypertension and diabetes

Evaluation

Expert coding of information quality and completeness
Random selection of 250 transcripts coded
"LLM as a judge" evaluation using GPT-5

TASK INSTRUCTIONS

Tasks adapted from Zhang 2013 *JMIR* (study of online search behavior)

Task 1
For this task, use ChatGPT to learn about any topic(s) related to medicine and health.
It's up to you to decide what you want to explore and how to interact with ChatGPT.
Remember that ChatGPT is interactive - you can and should write more after receiving your first response.

Task 2
For this task, you will use ChatGPT to learn about two specific topics related to heart health.
As before, it's up to you to decide how to interact with ChatGPT to research each topic.
Your goal should be to gather information and learn more about each topic -
you won't have to answer any specific questions based on the responses you get from ChatGPT.

Heart Attack
A heart attack is a medical emergency and prompt treatment increases the chance for survival.
According to the American Heart Association, heart attacks cause 1 out of every 5 deaths.
According to the National Institutes of Health (NIH) more than 1.2 million heart attacks
occur each year in the United States and about 460,000 of these are fatal. Approximately 300,000
people die annually from heart attacks before they can receive medical treatment. To be prepared
for possible emergencies, you decide to find out what to do when a person around you has a heart attack.

Diabetes/Hypertension
Imagine that one of your close family members has lived with diabetes for years. Recently, he was
also diagnosed with hypertension. You decided to do some research on the clinical associations
between the two conditions so that you are able to effectively discuss with him about various
implications of this diagnosis.

COMPLETENESS RUBRIC (HEART ATTACK)

1) Check for three common symptoms of a heart attack:
chest pain or discomfort; shortness of breath or difficulty breathing;
and pain or discomfort in arms or shoulders.

2) Check for other symptoms of a heart attack: nausea; extreme fatigue,
feeling lightheaded or faint or having cold sweats; and jaw, neck, or back pain.

3) Remember that not all heart attacks are the same,
and symptoms may vary substantially between people.

4) Call 911 or another number for emergency services immediately.

5) Check if the person has been previously prescribed sublingual nitroglycerin for such a situation.

6) Check if the person has been previously prescribed aspirin for such a situation.

7) Check for a pulse if the person is unconscious.

8) Start CPR, including defibrillation, if the person has no pulse.

COMPLETENESS RUBRIC (DIABETES/HYPERTENSION)

1) Diabetes and hypertension share risk factors, such as suboptimal diet and sedentary lifestyles.

2) Both diabetes and hypertension are risk factors for cardiovascular disease.

3) Both diabetes and hypertension are risk factors for kidney disease.

4) Dietary changes, for instance involving heart-healthy diets like the DASH diet, are essential to control both.

5) Other healthy lifestyle measures, including exercise and weight management, are important to control both.

6) It is important to visit the doctor to make sure appropriate testing and surveillance is done,
such as blood pressure checks and laboratory tests for glucose levels, renal function,
and hemoglobin A1c levels. Specialist referral may be required as well.

7) Medications may be needed to control these conditions and require good adherence.

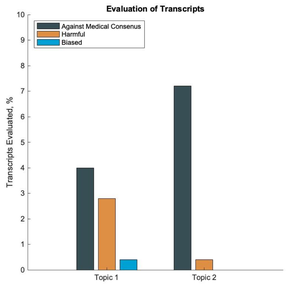
8) These conditions are chronic and lifelong, not acute or short-term.

TRANSCRIPT SUMMARY

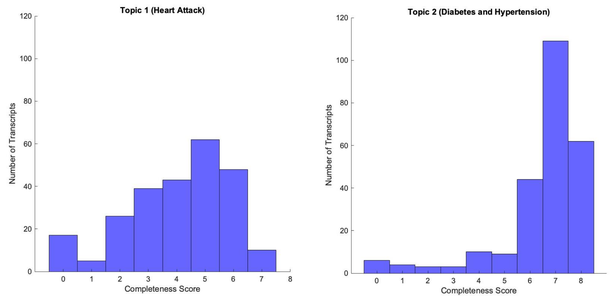
Table. Summary of Transcripts of Participant Conversations with ChatGPT

	Task 1	Task 2	
	Open Topic	Topic 1 (Heart Attack)	Topic 2 (Diabetes and Hypertension)
Number of Transcripts	389	389	389
Number of Turns	8 (4) [2 - 26]	6 (4) [2 - 22]	6 (2) [2 - 16]
Length of Prompts (Words)	37 (35) [0, 1475]	34 (29) [0, 1050]	25 (27) [0, 542]
Length of AI-Generated Responses (Words)	1170 (907) [49, 2750]	948 (819) [45, 2853]	1048 (770) [32, 2820]

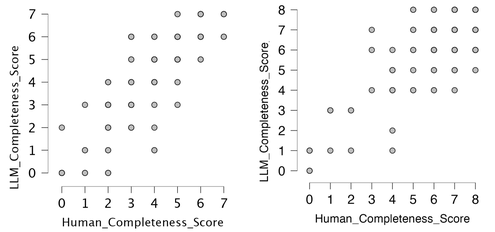
INFORMATION QUALITY



COMPLETENESS SCORE DISTRIBUTIONS



AGREEMENT OF HUMAN/LLM SCORING



Full Project Team

Cleveland Clinic: Ashish Sarraju, Astefanos Al-Dalakta, Bliss Chang, Anmol Multani, Roshan Patel

University of Macau: Anqi Gu, Xiaoyu Lu, Yihao Zhang

Harvard Medical School: Chirag Patel

The Roux Institute

HDSI

协同创新研究院
Institute de Innovação Colaborativa
Institute of Collaborative Innovation