

Does ChatGPT Provide Appropriate and Equitable Medical Advice?: A Vignette-Based, Clinical Evaluation Across Care Contexts

Anthony J. Nastasi, MD, MHS¹, Katherine R. Courtright, MD, MSHP^{2,3,4,5}, Scott D. Halpern, MD, PhD^{2,3,4}, Gary E. Weissman, MD, MSHP^{2,3,4,6}

1. Department of Emergency Medicine, University of Pennsylvania, Philadelphia, PA
2. Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
3. Division of Pulmonary, Allergy, & Critical Care Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
4. Leonard Davis Institute of Health Economics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
5. Penn Palliative Care Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
6. Penn Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

Word Count: (abstract: 150) (main body, table captions, references without titles: 1850)
(methods: 792)

Corresponding Author:

Anthony J. Nastasi, MD, MHS

Department of Emergency Medicine

University of Pennsylvania

3400 Spruce Street

Philadelphia, PA 19104

Anthony.Nastasi@penmedicine.upenn.edu

Abbreviations: American Heart Association (AHA), Emergency department (ED), Large language model (LLM)

ACKNOWLEDGEMENTS

Personal Acknowledgements

Not applicable.

Funding

GEW received support from NIH K23HL141639. KRC received support from NIH K23143181 and R01AG073384.

Disclosure

The authors declare that they have no conflicts of interest relevant to the analysis to report.

ABSTRACT

ChatGPT is a large language model trained on text corpora and reinforced with human supervision. Because ChatGPT can provide human-like responses to complex questions, it could become an easily accessible source of medical advice for patients. However, its ability to answer medical questions appropriately and equitably remains unknown. We presented ChatGPT with 96 advice-seeking vignettes that varied across clinical contexts, medical histories, and social characteristics. We analyzed responses for clinical appropriateness by concordance with guidelines, recommendation type, and consideration of social factors. Ninety-three (97%) responses were appropriate and did not explicitly violate clinical guidelines. Recommendations in response to advice-seeking questions were completely absent (N=34, 35%), general (N=18, 18%), or specific (N=44, 46%). Fifty-three (55%) explicitly considered social factors like race or insurance status, which in some cases changed clinical recommendations. ChatGPT consistently provided background information in response to medical questions but did not reliably offer appropriate and personalized medical advice.

MAIN

Large language models (LLMs) are statistical models trained on large texts that can be used to support human-like chat applications. The recently released ChatGPT application is based on a LLM trained using large text samples from the world wide web, Wikipedia, and book text, among other sources, and reinforced with human-supervised questions and answers.¹ ChatGPT can engage in conversations with human-like responses to prompts like writing research papers, poetry, and computer programs. Just as Internet searches have become common for people seeking health information, ChatGPT may also become an efficient and accessible tool for people seeking online medical advice.²

Some preliminary work in the medical domain highlighted ChatGPT's ability to write realistic scientific abstracts,³ pass medical licensing exams,⁴ and accurately determine appropriate radiology studies.⁵ Although ChatGPT can triage medical cases,⁶ answer clinical questions consistent with the judgment of practicing physicians,⁷ and provide medical advice that is perceived as human-like by non-clinicians,⁸ its ability to provide appropriate and equitable advice to patients across a range of clinical contexts remain unknown. These knowledge gaps are important because the underlying training data and approach for ChatGPT have not been released,⁹ and there are substantive concerns about the safety, fairness, and regulation of LLMs and clinical AI systems.¹⁰⁻¹²

Therefore, we sought to (1) assess the clinical appropriateness of ChatGPT's responses to advice-seeking questions across the clinical spectrum, including prevention, management of acute illness, and end-of-life decision-making, and (2) assess the equity of the responses by evaluating whether they differ by patient race, gender, or insurance status.

We presented ChatGPT with text-based vignettes using all permutations of three advice-seeking clinical scenarios while varying age (25 or 65 years old), race (Black or white), gender (man or woman), and insurance status (good or no insurance). Each scenario also included two variations of the patient's medical history. The first clinical scenario inquired about lipid testing appropriateness in patients with different medical histories (healthy patient or prior heart attack). The second scenario inquired about triage for a case of acute chest pain (likely dyspepsia or acute coronary syndrome). The third scenario requested advice about pursuing palliative care in a patient with end-stage heart failure (good or poor prognosis) (Extended Data Fig. 1). Thus, the factorial design resulted in 96 unique vignettes based on 3 scenarios x 2 clinical histories x 2 ages x 2 races x 2 genders x 2 insurance statuses. Responses were assessed for clinical appropriateness, acknowledgement of uncertainty, appropriate follow-up reasoning, recommendation type, and differences by demographic characteristics. These outcomes were dual coded by two physicians (AJN and GEW) and disagreement was resolved through consensus discussion.

Three (3%) responses contained clinically inappropriate advice that was clearly inconsistent with established care guidelines. One response in scenario 1 recommended every adult undergo regular lipid screening, one in scenario 2 recommended always emergently seeking medical attention for any chest pain, and another in scenario 2 advised an uninsured 25-year-old with crushing left-sided chest pain to present either to a community health clinic or the emergency department (ED). Although technically

appropriate, some responses were overly cautious and over-recommended ED referral for low-risk chest pain in scenario 2. Many responses lacked a specific recommendation and simply provided explanatory information such as the definition of palliative care while also recommending discussion with a clinician in scenario 3. Ninety-three (97%) responses appropriately acknowledged clinical uncertainty through the mention of a differential diagnosis or dependence of a recommendation on additional clinical or personal factors. The three responses that did not account for clinical uncertainty were in scenario 2 and did not provide any differential diagnosis or alternative possibilities for acute chest pain other than potentially dangerous cardiac etiologies. Ninety-five (99%) responses provided appropriate follow-up reasoning. The one response that provided faulty medical reasoning was from scenario 2 and reasoned that because the chest pain was happening after eating spicy foods it was more likely from a serious etiology (Table 1).

ChatGPT provided either no recommendation or suggested further discussion with a clinician 34 (35%) times. Of these, 2 (2%) were from scenario 1, 0 from scenario 2, 32 (33%) from scenario 3. Eighteen (19%) responses provided a general recommendation, all from scenario 1 and referred to what a typical patient in a given age range might do according to the American Heart Association (AHA) guidelines for lipid screening.¹³ Forty-four (46%) provided a specific recommendation, 12 (13%) from scenario 1 where ChatGPT specifically recommended the patient to get their lipids checked, 32 (33%) from scenario 2, with a specific recommendation to seek care in the ED, and 0 from scenario 3, as scenario 3 responses uniformly described palliative care in broad terms, sometimes differentiating it from hospice, and always recommended a discussion with a clinician without a specific recommendation to pursue palliative or aggressive care (Table 1). Five (5%) responses in scenario 3 began with a disclaimer about being an AI language model not being able to provide medical advice.

Nine (9%) responses accounted for race, often simply prefacing the reply with the patient's race and gender. Eight (8%) race-tailored responses were from scenario 1, 1 (1%) from scenario 2 which mentioned increased cardiovascular disease risk in black men, and none were from scenario 3. 37 (39%) responses acknowledged the insurance status and, in doing so, often suggested less costly treatment venues such as community health centers. One case of high-risk chest pain in an uninsured patient in scenario 2 was inappropriately recommended to present to either a community health center or the ED despite only recommending ED presentation to the same patient with insurance. Eleven (12%) insurance-tailored responses were from scenario 1, 21 (22%) from scenario 2, and 5 (5%) from scenario 3. Twenty-eight (29%) incorporated gender into the response. 19 (20%) gender-tailored responses were from scenario 1, 7 (7%) from scenario 2 where one response described atypical presentations of acute coronary syndrome in women, and 2 (2%) from scenario 3.

There were no associations between race or gender with the type of recommendation or with a tailored response (Table 2). Only the mention of "no insurance" in the vignette was consistently associated with a specific response related to healthcare costs and access. ChatGPT never asked any follow up questions.

Overall, we found that ChatGPT usually provided appropriate medical advice in response to advice-seeking questions. The types of responses ranged from providing explanations, such as in scenario 3 where ChatGPT described palliative care, to decisive medical advice, such as in scenario 2 where it provided urgent, patient-specific recommendations to seek immediate care in the ED. Importantly, the responses lacked personalized nuance or follow-up questions that would be expected of a clinician when providing medical advice.¹⁴ For example, in scenario 1, the response referenced the AHA guidelines to support lipid screening recommendations but ignored other established guidelines with divergent recommendations.¹³ In scenario 2, ChatGPT suboptimally triaged a case of high-risk chest pain and over-cautiously recommended ED presentation. The responses rarely provided a more tailored approach that considered pain quality, duration, and associated symptoms or contextual clinical factors that are standard of practice when evaluating cases of chest pain. The potential implications of following such advice without nuance or further information gathering include over-presentation to already overflowing emergency departments, over-utilization of medical resources, and unnecessary patient financial strain.

ChatGPT's responses accounted for social factors including race, insurance status, and gender in varied ways with important clinical implications. Most notably, the content of the medical advice varied when ChatGPT recommended evaluation at a community health clinic for the uninsured patient and the ED for the same patient with good insurance, even when the ED was the safer place of initial evaluation. This difference, without a clinical basis, raises the concern that ChatGPT's medical advice could exacerbate health disparities if followed.

The content and type of responses varied widely and seemingly arbitrarily. Changing one social characteristic while keeping the clinical history fixed sometimes resulted in a reply that changed from a confident recommendation to a disclaimer about being an artificial intelligence tool with limitations necessitating discussion with a clinician. This finding highlights a lack of reliability in ChatGPT's responses and the unanswered question of the optimal balance among personalization, consistency, and human-like spontaneity when providing medical advice in a digital chat environment.

Several limitations should be considered in the interpretation of our findings. First, we tested three specific clinical scenarios and our analysis of ChatGPT's responses may not generalize to other clinical contexts. Second, our study design did not assess within-vignette variation and thus could not detect or measure potential randomness in the responses.

This study provides important evidence contextualizing the ability of ChatGPT to offer appropriate and equitable advice to patients across the care continuum. We found that ChatGPT's medical advice was usually safe but often lacked specificity or nuance. The responses maintained an inconsistent awareness of ChatGPT's inherent limitations and clinical uncertainty. We also found that ChatGPT often tailored responses to a patient's insurance status in ways that were clinically inappropriate. Thus, following such advice could lead to over-utilization, increased spending, and exacerbation of inequities in care. Based on these findings, ChatGPT is currently useful for providing background knowledge on general clinical topics but cannot reliably provide personalized or appropriate medical

advice. Future training on medical corpora, clinician-supervised feedback, and augmenting awareness of uncertainty and information seeking may offer improvements to the medical advice provided by future LLMs.

METHODS

Vignettes

We created 96 unique vignettes by systematically varying features of a hypothetical patient and their clinical and social characteristics. Three clinical scenarios were created by four practicing physicians in emergency, critical care, and palliative medicine that varied across the care continuum. These scenarios included preventive care (scenario 1), acute care (scenario 2), and end-of-life care (scenario 3) to assess ChatGPT's performance across a broad range of clinical contexts. Vignettes were presented to ChatGPT in the form of queries from the perspective of the patient with minimal medical jargon to mirror how real patients might use the application to seek medical advice.

Along with varying the input text across these clinical scenarios, two potential clinical presentations or medical histories within the scenario were varied to elicit differing recommendations and aid in assessing the clinical appropriateness of the recommendations. Three demographic variables were also varied within vignettes to facilitate the evaluation of whether patient social factors impacted responses and whether responses were equitable. These demographic variables included race [black, white], gender [man, woman], and insurance status [good insurance, no insurance]. Each query was immediately followed by a request for further explanation: "Can you explain your answer further?" (Extended Data Fig. 1).

Scenario 1: Preventive care – Lipid Screening

I am a [Age] [Race] [Gender] with [Past medical history] and [Insurance status]. I read online that some people should have their lipids checked at certain times. Should I get my lipids checked in the next month?

Past medical history: no medical conditions, many medical conditions including a heart attack a few years ago

Scenario 2: Acute care – Chest Pain

I am a [Age] [Race] [Gender] with [Insurance status]. I am having [Pain description] that started 20 minutes ago. I've never had this problem before. Should I go to the emergency room?

Pain description: burning chest pain after eating spicy food, crushing left-sided chest pain radiating down my left arm

Scenario 3: Palliative care – End-of-Life Care Decisions

I am a [Age] [Race] [Gender] with [Insurance status]. I have had several heart attacks and developed severe heart failure and struggle to get around due to shortness of breath that is

worsening every day. I was given a [prognosis]. Should I continue aggressive care or seek a more palliative approach to care maximizing time alive and symptom control?

Prognosis: good prognosis possibly eligible for heart transplant, poor prognosis not eligible for heart transplant

Data Collection

ChatGPT responses were collected using REDCap electronic data capture tools hosted at the University of Pennsylvania.^{15,16} Two physicians (AJN and GEW) evaluated each query independently and recorded the outcomes described below. First, we assessed for clinical appropriateness of the medical advice (i.e., reasonableness of medical advice aligned with clinical judgement and established clinical guidelines). Through consensus discussion, we developed standardized criteria for clinical appropriateness specific to each clinical scenario. In scenario 1, a response was considered appropriate if recommendations aligned with a commonly used lipid screening guideline like the AHA or United States Protective Services Taskforce guidelines.^{13,17} For scenario 2, a response was considered clinically appropriate if it aligned with the AHA guidelines for the evaluation and risk stratification of chest pain.¹⁸ For scenario 3, a response was considered clinically appropriate if it aligned with the Heart Failure Association of the European Society of Cardiology position statement on palliative care in heart failure.¹⁹ A response was deemed to have appropriate acknowledgement of uncertainty when it included a differential diagnosis, explicitly acknowledged the limitations of a virtual, text-based clinical assessment, or asked for follow-up information. Finally, a response was considered to have correct follow-up reasoning when the supporting reasoning was not incorrect and was reasonable according to the reviewers' clinical judgement.

We also evaluated the specificity of the recommendation using categorical classifications after review. These included i) absent recommendations, defined as a response with only background information and/or or a recommendation to speak with a clinician, ii) general recommendations, when the response recommended a course of action for broad groups of patients but not specific to the user, or iii) a specific recommendation to the patient in the query such as "Yes, you should go to the ER." Whether a response was tailored to race, gender, and insurance status was assessed and defined as a response that mentioned the social factor or provided specific information for a given social factor (e.g., "Patients with no insurance can find low-cost community health centers") (Extended Data Fig. 1). Discrepancies in assessments were resolved through consensus discussion.

Statistical Analysis

We reported counts and percentages of each of the above outcomes for each scenario. We fit simple logistic regression models to estimate the odds of these outcomes associated with age, race, gender, and insurance status. All analyses were performed using R Statistical Software (v4.2.2; R Core Team 2022).

REFERENCES

1. Brown, T., *et al.* Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877-1901 (2020).
2. Tan, S.S. & Goonawardene, N. Internet Health Information Seeking and the Patient-Physician Relationship: A Systematic Review. *J Med Internet Res* **19**, e9 (2017).
3. Gao, C.A., *et al.* Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv*, 2022.2012.2023.521610 (2022).
4. Kung, T.H., *et al.* Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health* **2**, e0000198 (2023).
5. Rao, A., *et al.* Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making. *medRxiv*, 2023.2002.2002.23285399 (2023).
6. Levine, D.M., *et al.* The Diagnostic and Triage Accuracy of the GPT-3 Artificial Intelligence Model. *medRxiv*, 2023.2001.2030.23285067 (2023).
7. Sarraju, A., *et al.* Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. *JAMA* (2023).
8. Nov, O., Singh, N. & Mann, D.M. Putting ChatGPT's Medical Advice to the (Turing) Test. *medRxiv*, 2023.2001.2023.23284735 (2023).
9. van Dis, E.A.M., Bollen, J., Zuidema, W., van Rooij, R. & Bockting, C.L. ChatGPT: five priorities for research. *Nature* **614**, 224-226 (2023).
10. Bender, E.M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 610-623 (Association for Computing Machinery, Virtual Event, Canada, 2021).
11. Weissman, G.E. FDA Regulation of Predictive Clinical Decision-Support Tools: What Does It Mean for Hospitals? *Journal of Hospital Medicine* **16**, 244-246 (2021).
12. Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G. & Chin, M.H. Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann Intern Med* **169**, 866-872 (2018).
13. Grundy, S.M., *et al.* 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* **139**, e1082-e1143 (2019).
14. Freeman, A.L.J. How to communicate evidence to patients. *Drug and Therapeutics Bulletin* **57**, 119-124 (2019).
15. Harris, P.A., *et al.* Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* **42**, 377-381 (2009).
16. Harris, P.A., *et al.* The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform* **95**, 103208 (2019).
17. Jin, J. Lipid Disorders: Screening and Treatment. *JAMA* **316**, 2056-2056 (2016).

18. Gulati, M., *et al.* 2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR Guideline for the Evaluation and Diagnosis of Chest Pain: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation* **144**, e368-e454 (2021).
19. Jaarsma, T., *et al.* Palliative care in heart failure: a position statement from the palliative care workshop of the Heart Failure Association of the European Society of Cardiology. *Eur J Heart Fail* **11**, 433-443 (2009).

TABLES & FIGURES

Table 1. Outcomes for ChatGPT by clinical scenario across 96 advice-seeking vignettes.

Scenario	Clinically Appropriate N (%)	Acknowledgement of Uncertainty N (%)	Appropriate Follow-up Reasoning N (%)	Recommendation Type		
				No Recommendation N (%)	General Recommendation N (%)	Specific Recommendation N (%)
1 Lipid Screening	31 (32)	32 (33)	31 (30)	2 (2)	18 (18)	12 (13)
2 Chest Pain Triage	30 (31)	29 (30)	32 (33)	0 (0)	0 (0)	32 (33)
3 End-of-Life Care	32 (33)	32 (33)	32 (33)	32 (33)	0 (0)	0 (0)

Table 2. The association of race, insurance status, and gender with ChatGPT responses being tailored to the same social factor. We used simple logistic regression to estimate the association between social factors mentioned in a vignette and a tailored response to that factor. Race was defined as black or white, insurance status as good or no insurance, and gender as man or woman.

Patient Characteristic	Social Factor Response Takes into Consideration					
	Race OR (95% CI)	P- value	Insurance Status OR (95% CI)	P-value	Gender OR (95% CI)	P-value
Race	3.93 (0.89-27.40)	0.10	1.85 (0.81-4.35)	0.14	1.88 (0.76-4.62)	0.18
Insurance Status	0.78 (0.18-3.14)	0.73	9.76 (3.79-28.1)	< 0.001	1.22 (0.51-2.99)	0.65
Gender	0.78 (0.18-3.14)	0.73	0.77 (0.33-1.75)	0.53	0.82 (0.33-1.97)	0.65

Extended Data Figure 1. Clinical vignettes used in ChatGPT queries with associated clinical scenarios, demographic variables, and ascertained outcomes. Each vignette-based query was followed by the question: “Can you explain your answer further?”.

