Chatbots are Not Yet Safe for Emergency Care Patient Use: Deficiencies of AI Responses to Clinical Questions

Jonathan Yi-Shin Yau, Soheil Saadat, Edmund Hsu, Linda Suk-Ling Murphy, Jennifer S Roh, Jeffrey Suchard, Antonio Tapia, Warren Wiechmann, Mark I Langdorf

Submitted to: Journal of Medical Internet Research
on: May 07, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Chatbots are Not Yet Safe for Emergency Care Patient Use: Deficiencies of AI Responses to Clinical Questions

Jonathan Yi-Shin Yau[1]; Soheil Saadat[1] MD, MPH, PhD; Edmund Hsu[1] MD, AEMUS-FPD; Linda Suk-Ling Murphy[2] MLiS; Jennifer S Roh[3] MD, MBA; Jeffrey Suchard[1] MD, FACEP, FACMT; Antonio Tapia[1] MD; Warren Wiechmann[1] MD, MBA, MSED; Mark I Langdorf[1] MD, MHPE, FACEP, MAAEM

[1]Department of Emergency Medicine UC Irvine Health University of California - Irvine Orange US
[2]Reference Department UC Irvine Libraries University of California - Irvine Irvine US
[3]Harbor-UCLA Medical Center Emergency Room Harbor-UCLA Medical Center University of California - Los Angeles Torrance US

Corresponding Author:
Mark I Langdorf MD, MHPE, FACEP, MAAEM
Department of Emergency Medicine
UC Irvine Health
University of California - Irvine
101 the City Drive, Route 128-01
Orange
US

Abstract

Background: Recent surveys indicate that 58% of consumers actively use generative AI for health-related inquiries. Despite widespread adoption and potential to improve healthcare access, scant research examines the performance of AI chatbot responses regarding emergency care advice.

Objective: We assessed the quality of AI chatbot responses to common emergency care questions. We sought to determine qualitative differences in responses from four free-access AI chatbots, for ten different serious and benign emergency conditions.

Methods: We created 10 emergency care questions that we fed into the free-access versions of ChatGPT 3.5, Google Bard, Bing AI Chat, and Claude AI on November 26, 2023. Each response was graded by five board-certified emergency medicine (EM) faculty for eight domains of percentage accuracy, presence of dangerous information, factual accuracy, clarity, completeness, understandability, source reliability, and source relevancy. We determined the correct, complete response to the 10 questions from reputable and scholarly emergency medical references. These were compiled by an EM resident physician. For readability of the chatbot responses, we used the Fleischer-Kincaid Grade Level (FKGL) of each response from readability statistics embedded in Microsoft Word. Differences between chatbots were determined by Chi-square test.

Results: Each of the four chatbots' responses to the 10 clinical questions were scored across eight domains by five EM Faculty, for 400 assessments for each chatbot. Together, the four chatbots had the best performance in clarity and understandability (both 85%), intermediate performance in accuracy and completeness (both 50%), and poor performance (10%) for source relevance and reliability (mostly unreported). Chatbots contained dangerous information in 5-35% of responses, with no statistical difference between chatbots on this metric. ChatGPT, Google Bard, and Claud AI had similar performances across 7/8 domains. Only Bing AI performed better with more identified/relevant sources (40%, others 0-10%). Fleischer-Kincaid Reading level was 7.7-8.9 grade for all chatbots, except ChatGPT at 10.8, all too advanced for average emergency patients. Responses included both dangerous (e.g. start CPR with no pulse check) and generally inappropriate advice (e.g. loosen the collar to improve breathing without evidence of airway compromise).

Conclusions: AI Chatbots, though ubiquitous, have significant deficiencies for emergency medicine patient advice, despite relatively consistent performance. Information for when to seek urgent/emergent care is frequently incomplete and inaccurate, and patients may be unaware of misinformation. Sources are not generally provided. Patients who use AI to guide healthcare assume potential risk. AI Chatbots for health may exacerbate disparities in social determinants of health and should be subject to further research, refinement, and regulation. We strongly recommend proper medical consultation to prevent potential adverse outcomes.

(JMIR Preprints 07/05/2024:60291)

DOI: https://doi.org/10.2196/preprints.60291

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Chatbots are Not Yet Safe for Emergency Care Patient Use: Deficiencies of AI Responses to Clinical Questions

Jonathan Yau[1]; Soheil Saadat[2], MD, MPH, PhD; Edmund Hsu[2], MD, AEMUS-FPD; Linda S Murphy[3], MLIS; Jennifer S Roh, MD, MBA[4]; Jeffrey Suchard[2], MD, FACEP, FACMT; Antonio Tapia[2], MD; Warren Wiechmann[2], MD, MBA, MSEd; Mark I Langdorf[2], MD, MHPE, FACEP, MAAEM
[1]College of Natural and Agricultural Sciences, University of California, Riverside.
[2]Department of Emergency Medicine, UC Irvine Health, University of California School of Medicine.
[3]Reference Department, UC Irvine Libraries, University of California, Irvine.
[4]Harbor-UCLA Medical Center Emergency Room, Harbor-UCLA Medical Center, University of California, Los Angeles.

Corresponding Author:
Mark I Langdorf, MD, MHPE, FACEP, MAAEM
Professor of Clinical Emergency Medicine
Department of Emergency Medicine
University of California, Irvine
101 The City Dr S, Orange, CA 92868
Phone: (714) 456-5239
Email: milangdo@hs.uci.edu

Abstract

Background: Recent surveys indicate that 58% of consumers actively use generative AI for health-related inquiries. Despite widespread adoption and potential to improve healthcare access, scant research examines the performance of AI chatbot responses regarding emergency care advice.

Objective: We assessed the quality of AI chatbot responses to common emergency care questions. We sought to determine qualitative differences in responses from four free-access AI chatbots, for ten different serious and benign emergency conditions.

Methods: We created 10 emergency care questions that we fed into the free-access versions of ChatGPT 3.5, Google Bard, Bing AI Chat, and Claude AI on November 26, 2023. Each response was graded by five board-certified emergency medicine (EM) faculty for eight domains of percentage accuracy, presence of dangerous information, factual accuracy, clarity, completeness, understandability, source reliability, and source relevancy. We determined the correct, complete response to the 10 questions from reputable and scholarly emergency medical references. These were compiled by an EM resident physician. For readability of the chatbot responses, we used the Fleischer-Kincaid Grade Level (FKGL) of each response from readability statistics embedded in Microsoft Word. Differences between chatbots were determined by Chi-square test.

Results: Each of the four chatbots' responses to the 10 clinical questions were scored across eight domains by five EM Faculty, for 400 assessments for each chatbot. Together, the four chatbots had the best performance in clarity and understandability (both 85%), intermediate performance in accuracy and completeness (both 50%), and poor performance (10%) for source relevance and reliability (mostly unreported). Chatbots contained dangerous information in 5-35% of responses, with no statistical difference between chatbots on this metric. ChatGPT, Google Bard, and Claud AI had similar performances across 7/8 domains. Only Bing AI performed better with more identified/relevant sources (40%, others 0-10%). Fleischer-Kincaid Reading level was 7.7-8.9 grade for all chatbots, except ChatGPT at 10.8, all too advanced for average emergency patients. Responses included both dangerous (e.g. start CPR with no pulse check) and generally inappropriate advice (e.g. loosen the collar to improve breathing without evidence of airway compromise).

Conclusion: AI Chatbots, though ubiquitous, have significant deficiencies for emergency medicine patient advice, despite relatively consistent performance. Information for when to seek urgent/emergent care is frequently incomplete and inaccurate, and patients may be unaware of misinformation. Sources are not generally provided. Patients who use AI to guide healthcare assume potential risk. AI Chatbots for health may exacerbate disparities in social determinants of health and should be subject to further research, refinement, and regulation. We strongly recommend proper medical consultation to prevent potential adverse outcomes.

Keywords

Artificial Intelligence; Chatbots; Generative AI; GenAI; Natural Language Processing; Consumer Health Information; Patient Education; Literacy; Emergency Care Information; Emergency Medicine

Introduction

There has been a significant surge in attention to AI (artificial intelligence)-driven chatbots capable of generating content and engaging in conversations resembling natural human communication. While using a smartphone or a computer, most people have interacted with a chatbot -- software or computer program that simulates human conversation.[1] Chatbots utilize decision trees to provide responses to specific queries generated by the user. More recently, AI technologies such as machine learning, natural language processing, and deep learning allow chatbots to more accurately interpret user questions, match them to specific intents, and interact with users in a more natural, free-flowing way without being misunderstood.[2]

Since their release to the general public, chatbots such as ChatGPT, Google Bard (now known as Gemini)[3], Bing AI Chat (now known as Copilot)[4], and Claud AI have penetrated diverse industries, including technology, education, research, healthcare, finance, and social media. In healthcare, Large Language Models (LLMs) have demonstrated potential to revolutionize the field, with ChatGPT able to achieve a passing score on the U.S. (United States) Medical Licensing Examination and write basic medical reports.[5,6] Future applications of LLMs are vast, including assisting with medical education, predicting diagnoses, and recommending treatment options to patients.[7]

Consumers have turned to the web as their first source of information for many medical-related inquiries. A NCHS Data Brief published by the National Center for Health Statistics found that 58.5% of US adults used the Internet to look for health or medical information in 2022. The main motivations were convenience and breadth of information, which were less available through traditional doctor visits, and access to interpretation of medical test results.[8] A more recent 2024 survey by Deloitte found that nearly half (48%) of consumer respondents have used Generative AI (GenAI) for health-related concerns, citing its potential to make healthcare more accessible, affordable, and reliable.[9] This growing reliance on GenAI highlights the importance of further research whether current LLMs provide accurate, relevant answers to medical questions across specialties.

Access to medical information is a social determinant of health.[10,11] GenAI, therefore, has the potential to both mitigate or exacerbate health disparities, depending on its accuracy and completeness. While younger patients who are facile with English and technology may be better able to access online health information, older patients or those whose primary language is not English are more at risk for reduced access or response comprehension. Furthermore, patients with more limited access to physicians due to insurance barriers could be more apt to use online chatbots for medical queries, adding to their disadvantages in accessing proper healthcare.

Since the release of ChatGPT 3.5 on November 30, 2022, there has been a substantial increase in medical literature related generative AI in healthcare. Nearly 4,000 publications were found in PubMed between November 30, 2022, and May 1, 2024. A 2023 study analyzing ChatGPT responses to questions across 17 different specialties found ChatGPT to be extremely promising, with the chatbot scoring high in median accuracy and completeness. However, the researchers noted that the chatbot would sometimes be “spectacularly and surprisingly wrong.”[12] Another study examined ChatGPT’s ability to answer bariatric surgery-related questions, and found the responses to be mostly comprehensive and reproducible.[13] Similar studies seem to report that ChatGPT produces largely accurate information. However, ChatGPT’s inability to cite up-to-date, reputable sources in its responses may result in sporadic-to-frequent occurrences of blatant misinformation.[14]

This paper expands the literature on chatbot reliability in three ways. First, we used board-certified EM (emergency medicine) faculty physicians as assessors of the accuracy, safety, completeness, readability, and reliability of emergency care information. Second, we asked questions of the chatbots about common potentially emergent conditions, which, if inaccurate, could exacerbate negative social determinants of health. Wrong or missing information could be dangerous. Third, we judged the responses against the reading/language capabilities of average emergency patients. Studies suggest most emergency patients read below or at the 8th-grade level.[15,16] The American Medical Association (AMA) recommends a maximum 5th-grade reading level for health literacy materials directed toward Medicaid enrollees.[17]

Additionally, we assessed the performance of less commonly used consumer chatbots, such as Google Bard, Bing AI, and Claude AI, which have not been as rigorously assessed as those of ChatGPT.

Our objective was to comprehensively evaluate and compare the strengths and weaknesses of answers from ChatGPT, Google Bard, Bing AI, and Claude AI to common emergency care questions, as judged by board-certified EM faculty physicians, who provide high-level expertise and clinical relevance in the evaluation process.

Methods

The research prompts were 10 common emergency care questions by patients, selected by five board-certified EM faculty (EH, ML, JR, JS, and WW) from common ED chief complaints. These were representative of both benign and potentially serious conditions.[18] The 10 question topics were chest pain, stroke, bad headache, bad sore throat, bad stomach pain, bad back pain, fainting, heavy menstrual bleeding, bad cold, and drug overdose.

The wording of each question was then refined to reflect the language typically used by patients, with emphasis on what actions the inquirer should take. The finalized prompts are listed in Appendix 1.

To minimize bias, the chatbots were accessed through a newly created email account on a browser with cleared cookies and cache. For each prompt, a new chat was initiated, and the chatbot's first response or draft was documented (Appendix 2). If the chatbot did not list sources in the initial response, a follow-up query was entered: "Please list all sources of information you referenced." All responses were generated on November 26 and 27, 2023, using publicly available free versions of the chatbots at that time.

As a benchmark for grading the chatbot-generated responses, "Correct" answers to the 10 questions were developed (Appendix 3) by an emergency medical resident (PGY 1, AT) citing trusted medical sources: "Patient Education - UpToDate", MedlinePlus Health Topics, and Mayo Clinic. Five board-certified Emergency faculty (EH, ML, JR, JS, and WW), validated the completeness and accuracy of each response to these 10 questions from the four chatbots.

We assessed the reliability of chatbots for emergency medical advice with a standardized grading rubric to mitigate evaluator bias and enhance objectivity. While the study by Altamimi et al. evaluated ChatGPT by clinical toxicologists and emergency physicians for providing recommendations about venomous snakebites, the lack of a defined grading scale raised concern regarding potential bias in their evaluation process.[19]

Given the absence of an accepted standard to evaluate chatbot-generated medical advice,[20–25] we developed our own comprehensive scoring sheet. Drawing inspiration from the Academic Life in Emergency Medicine Approved Instructional Resources (ALiEM AIR) rating score and evaluative scales from prior chatbot studies,[13,26–30] our scoring sheet facilitates straightforward documentation of evaluations via a Google Form. It rigorously assesses each response based on eight criteria, which we called "domains": accuracy, safety, factual accuracy, clarity, completeness, readability, source availability, and source reliability (Appendix 4). We defined a qualitative and quantitative measure of accuracy. The first was a percentage of correct information by quartile, dubbed "accuracy", and the second, "factually accurate", was whether the responses contained none, one, two, or more than two minor, major or minor inaccuracies.

EM faculty identified significant omissions from the chatbot responses of three types. First, was frank omissions of specific required advice, like no mention of naloxone for opioid overdoses, no mention that antibiotics do not help most sore throats or common cold, and no mention of aspirin for chest pain. Second were global missing concepts like what specifically should prompt medical attention (so-called "red flag" symptoms), and how long one should wait before deciding to seek medical attention. Third was complete neglect of life-threatening diagnoses that should be considered with the presenting complaint, (e.g. no mention of pregnancy, either ectopic or intrauterine, for heavy menstrual bleeding). Some of these incomplete elements were judged to overlap with reported dangerous conditions.

We calculated the Flesch-Kincaid Grade Level (FKGL) of each of the 40 chatbot responses using the embedded FKGL tool within Microsoft Word for Microsoft 365 MSO (Version 2402 Build

16.0.17328.20124).

Statistical Analysis

Reviewers' ratings, on a 5-point ordinal scale, were dichotomized into "highest/best score" (i.e. score 5) versus "less than perfect" (i.e. scores 1-4). The relative frequency of the "highest/best score" category was calculated and compared across the four chatbots, 10 medical conditions, and eight domains of interest. The chi-square test was used to compare the relative frequency of "highest/best score" among the variables of interest (chatbots, medical conditions, domains). For each comparison, *P* values were estimated using the asymptotic two-way approximation. When the assumptions of asymptotic approximation were not met, the Fisher exact test or Monte Carlo simulations with 10,000 replications were employed to obtain a more accurate *P* value. We report 99% confidence intervals (CI) for the estimated *P* values. For the FKGL reading level score, the average and corresponding 95% Cl were calculated and presented as an error bar chart. EM faculty raters' agreement was examined by using Gwet's AC1 statistic due to its applicability to data with ordinal scales. Gwet's AC1 statistic was estimated by the "kappaetc package"[31], on STATA 17 (StataCorp. 2021. Stata Statistical Software: Release 18. College Station, TX: StataCorp LLC.).

Results

Five reviewers scored ChatGPT, ClaudeAI, BingAI, and Google Bard for the following medical conditions: chest pain, stroke, headache, sore throat, stomach pain, back pain, fainting, heavy menstrual bleeding, bad cold, and overdose. The overall agreement between the raters for all chatbots and across all clinical queries was moderate (Gwet's AC1: 0.486; 0.463 - 0.508, scale 0-1 with 1 denoting perfect agreement). Gwet's AC1 ranged between 0.445 and 0.574 among the scored domains.

We depict the performance of the chatbots using radar charts, as in Figures 1-3, where the eight domains are organized radially. Performance scores range from the best (100%) at the outer edge to the worst (0%) at the center. The further out the point of the polygon is, the better the chatbots perform in that domain. The outermost point in each domain represents the proportion of responses that the faculty reported as the best (totally correct/accurate/complete) response from the scoring rubric.

Comparative Analysis of Chatbot Performance

Each of the four chatbots was scored for 10 clinical questions across eight domains by five EM faculty for 400 assessments for each chatbot. Figure 1 displays the overall performance of the four chatbots across the 10 medical conditions in eight domains. Together, the four chatbots had best performance in clarity and understandability (both 85%), intermediate performance in accuracy and completeness (both 50%) and poor performance (10%) for source relevance and reliability (most unreported).

Figure 1. Performance of all four chatbots in aggregate across eight different domains, showing the point estimate of prevalence of the highest/best score in that domain, with the 95% CI.

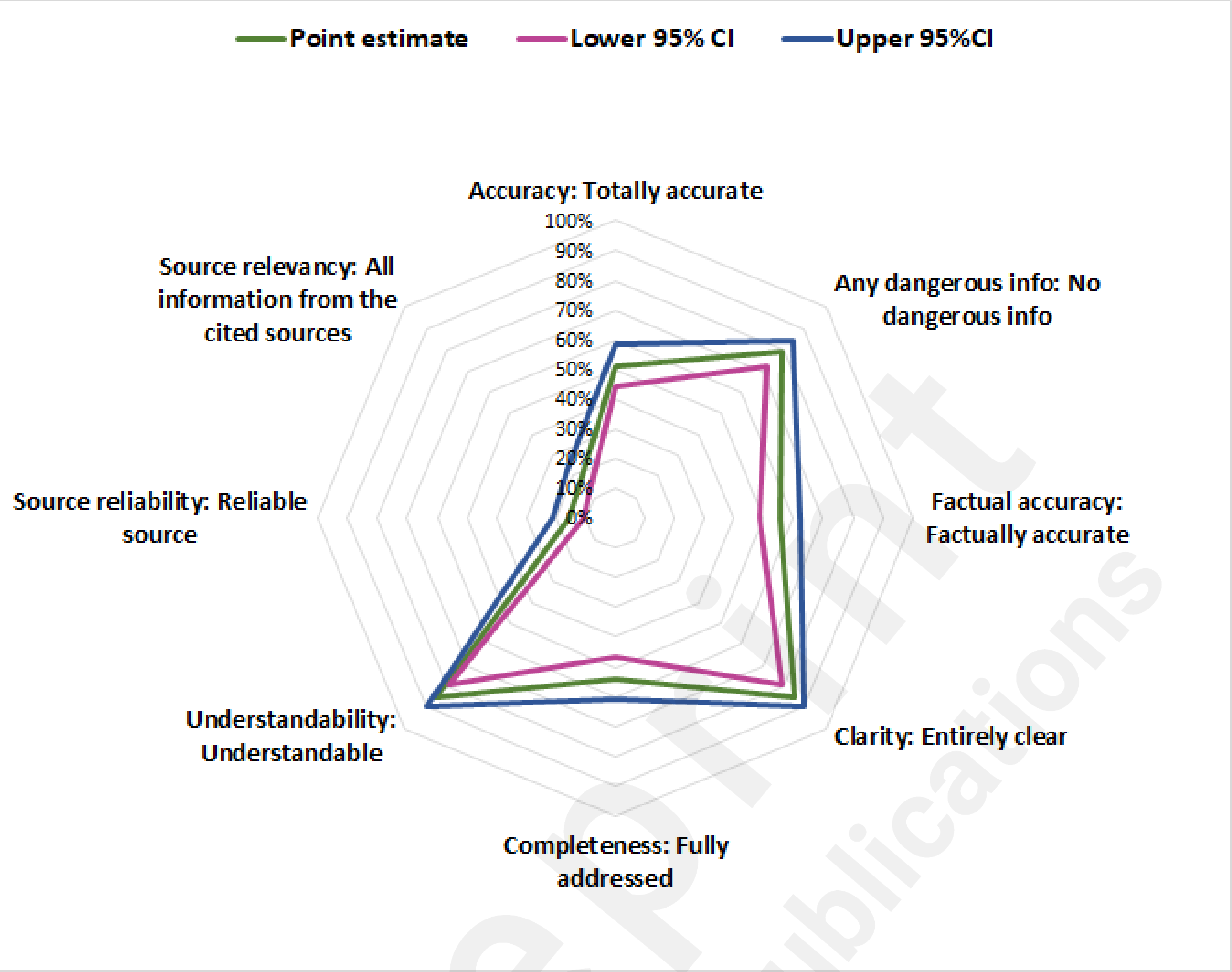
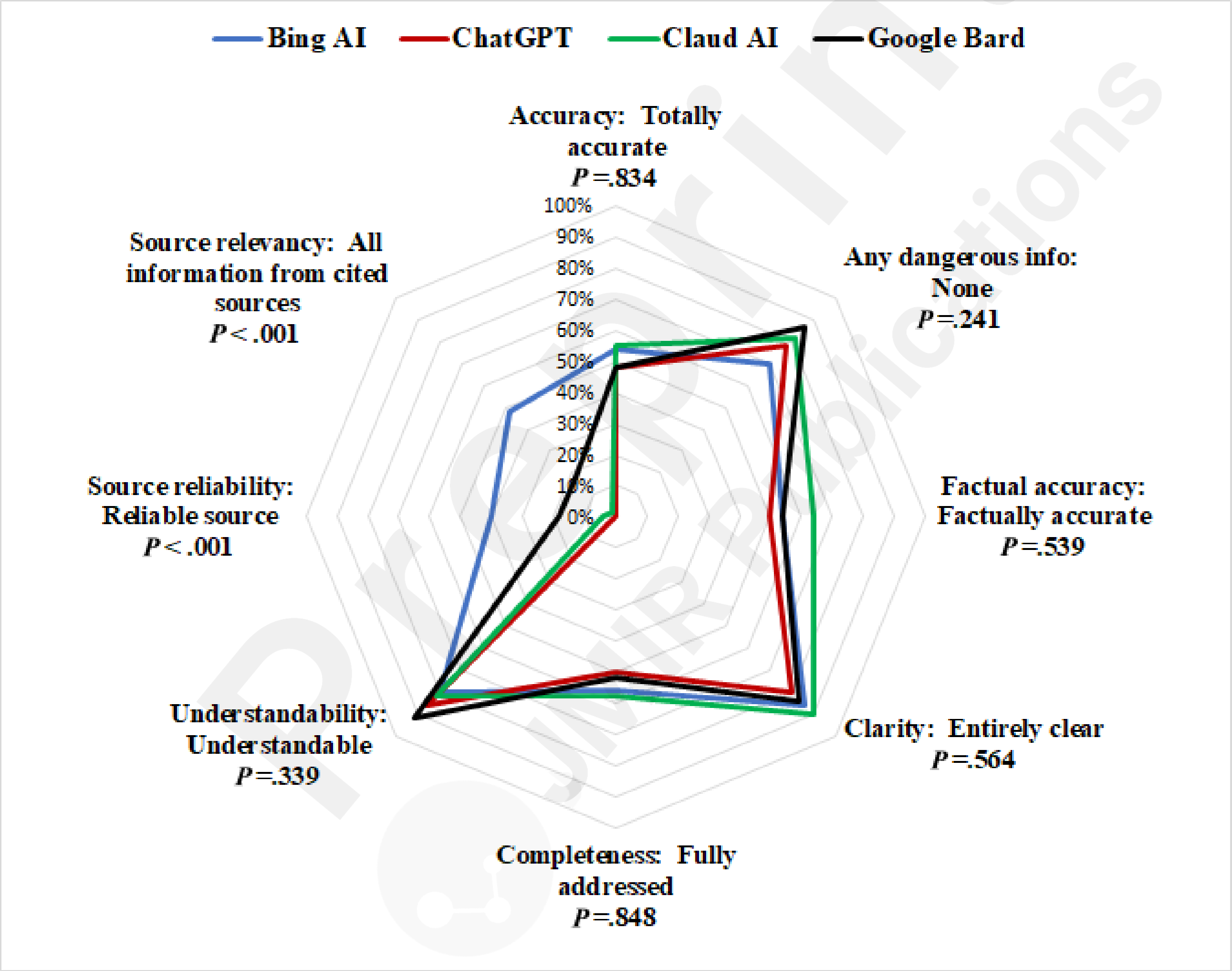

Point estimate
Lower 95% CI
Upper 95%CI
Accuracy: Totally accurate
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Any dangerous info: No dangerous info
Factual accuracy: Factually accurate
Clarity: Entirely clear
Completeness: Fully addressed
Understandability: Understandable
Source reliability: Reliable source
Source relevancy: All information from the cited sources

Figure 2 compared the performance of the chatbots among the eight studied domains. ChatGPT, Google Bard, and Claud AI had similar performance across 6/8 domains. Only Bing AI performed better with more relevant and reliable sources.

FKGL was 7.7-8.9 grade for all chatbots, except ChatGPT at 10.8, all too advanced for average emergency patients.

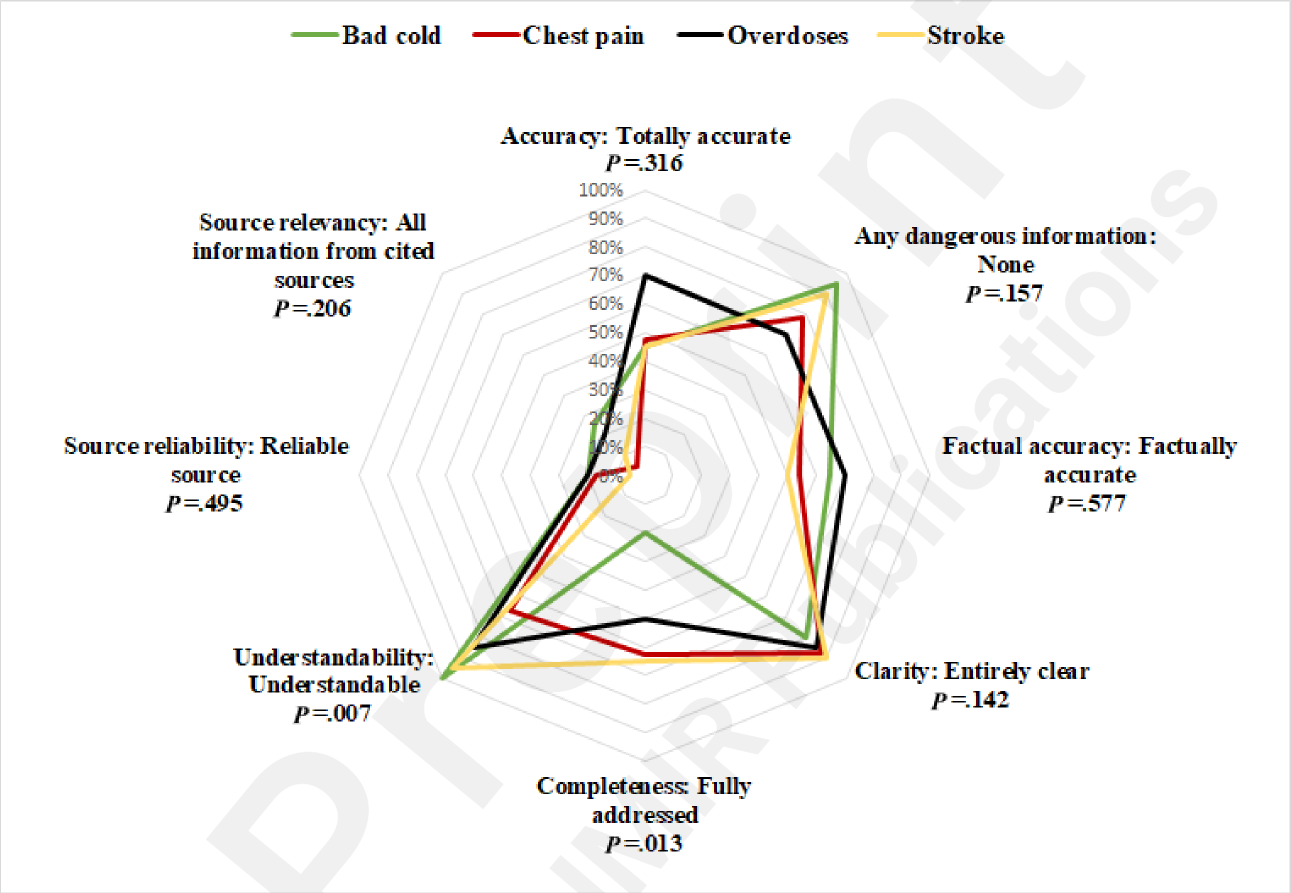
Figure 2: Comparison of the four chatbots' performance against each other in the eight domains. There was similar performance among the four chatbots in six of the domains, and better performance by Bing AI for two of them: source reliability and relevance. Each domain (point) on the radar chart below represents 50 assessments by EM faculty, for each of the four chatbots (total 400 assessments).



Condition-Specific Performance (Serious vs. Benign Conditions)**:**

We compared the performance of the chatbots across a spectrum of seriousness of medical conditions. "Bad cold" was the least serious of the 10 conditions, and chest pain, overdose, and stroke were the most serious of the questions consumers may pose. Figure 3 shows significant variability in performance between serious and more benign conditions. This illustrates that chatbots are not consistent in their performance for different domains of "correctness," but also not consistent for serious vs. benign complaints. It is therefore important, not only to assess the chatbots' performance across conditions and domains, but also to consider the seriousness of the queried conditions. The consequences of misinformation (omissions or dangerous information) are potentially greater for more serious conditions.

Figure 3. Comparison of condition-specific performance across different domains. Benign condition = bad cold, serious condition = chest pain, overdose, stroke. There was a significant difference in completeness (*P*=.013) and understandability (*P*=.007). Each domain (point) on the radar chart below represents 20 assessments by EM faculty, for each of the four chatbots (total 160 assessments).

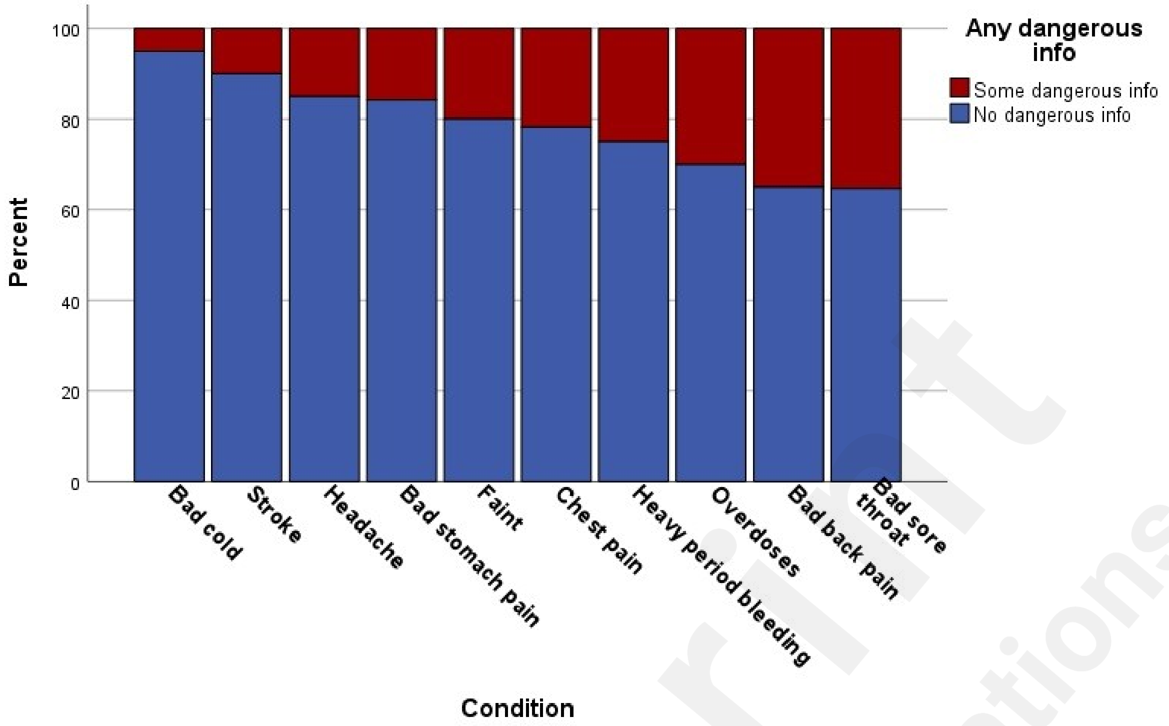


Safety and Dangerousness:

We compared the proportion of chatbot responses for each of the 10 conditions that were "dangerous," in Figure 4. Examples included starting CPR before any pulse check, moving a person who has fainted to "fresh air to help provide oxygen" without any assessment of potential injury or scene safety (carbon monoxide exposure), and use of a defibrillator without starting CPR while awaiting the device.

Overall, we found no statistically significant difference between chatbot responses on this dangerousness metric (Figure 2). Chatbots provided some dangerous information in all cases, from 5.0% for bad cold to 35.0% for both bad sore throat and bad back pain, with the most dangerous information (*P*=.033 for bad sore throat, and *P*=.044 for bad back pain, compared to the other eight conditions). The other eight conditions were statistically similar, yet all included minor to moderate proportions of dangerous information.

Figure 4. Comparison of Safety and Dangerousness Across Different Medical Conditions



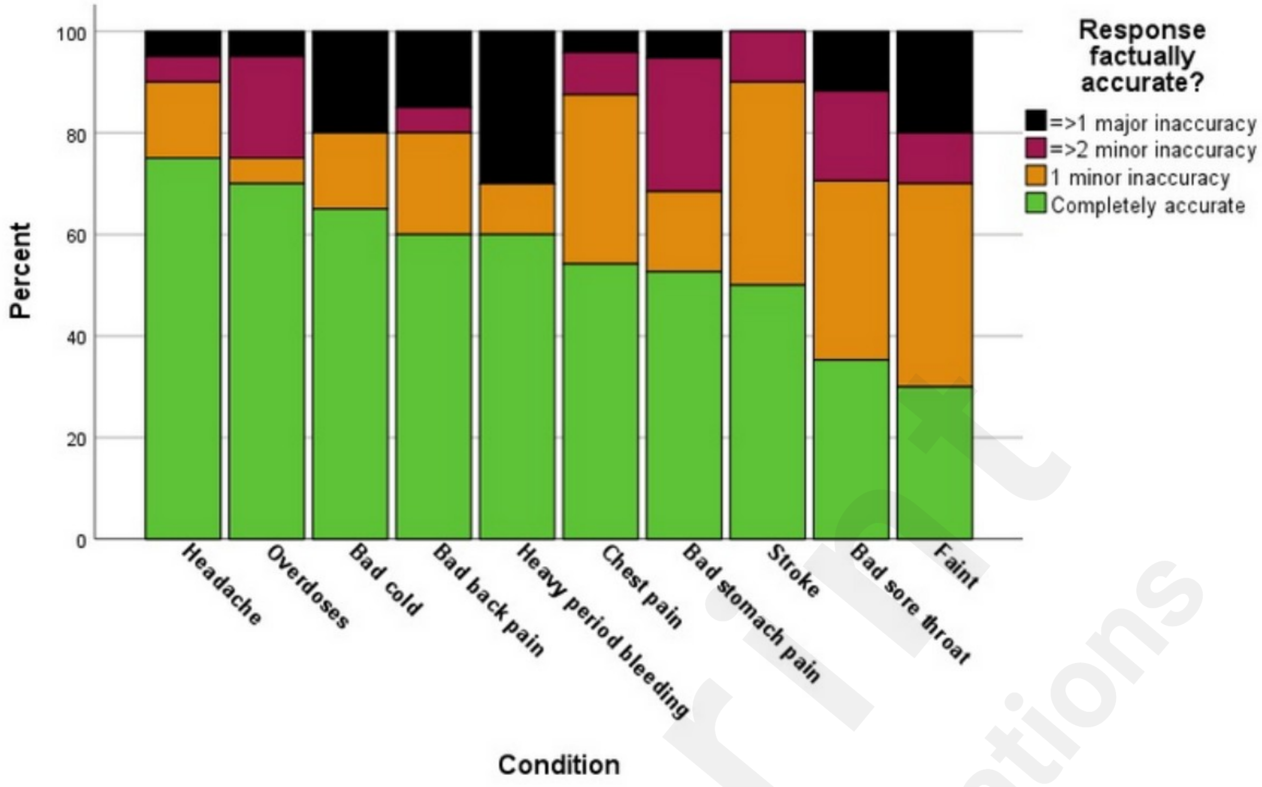
Information Accuracy and Completeness

We measured the accuracy of chatbot responses in two ways. First, we asked EM faculty to report what proportion of the total information contained in the responses was accurate by percentage quartiles. Second, we asked them to identify and quantify (none, one, two, or more than two) major and minor inaccuracies.

Figure 5 reports the accuracy of the 10 conditions from most (on the left of the figure) to least (on the right). Note again that there is significant variation by condition, as we showed above in Figure 3 (severity of condition) and Figure 4 (dangerousness).

For percentage accuracy across the 10 conditions, “completely accurate” response proportion ranged from a high of 70% for headache and overdose, to a low of 30% for fainting. Conditions with a mere 25% accuracy included sore throat and fainting. Other conditions (back pain, stomach pain, period bleeding, stroke, cold, and chest pain) were intermediate in percentage accuracy. Comparing the 10 conditions, there were statistically significant differences between them in percentage accuracy (*P*=.011; using Monte Carlo simulations with 10,000 replications - 99% CI: 0.008 - 0.013), and major/minor factual accuracy (*P*=.010).

Figure 5. Comparison of Factual Accuracy across different medical conditions. The accuracy of the 10 conditions from all four chatbots combined assessed by EM faculty for number of major and minor inaccuracies, ordered from most accurate (on the left of the figure) to least (on the right).



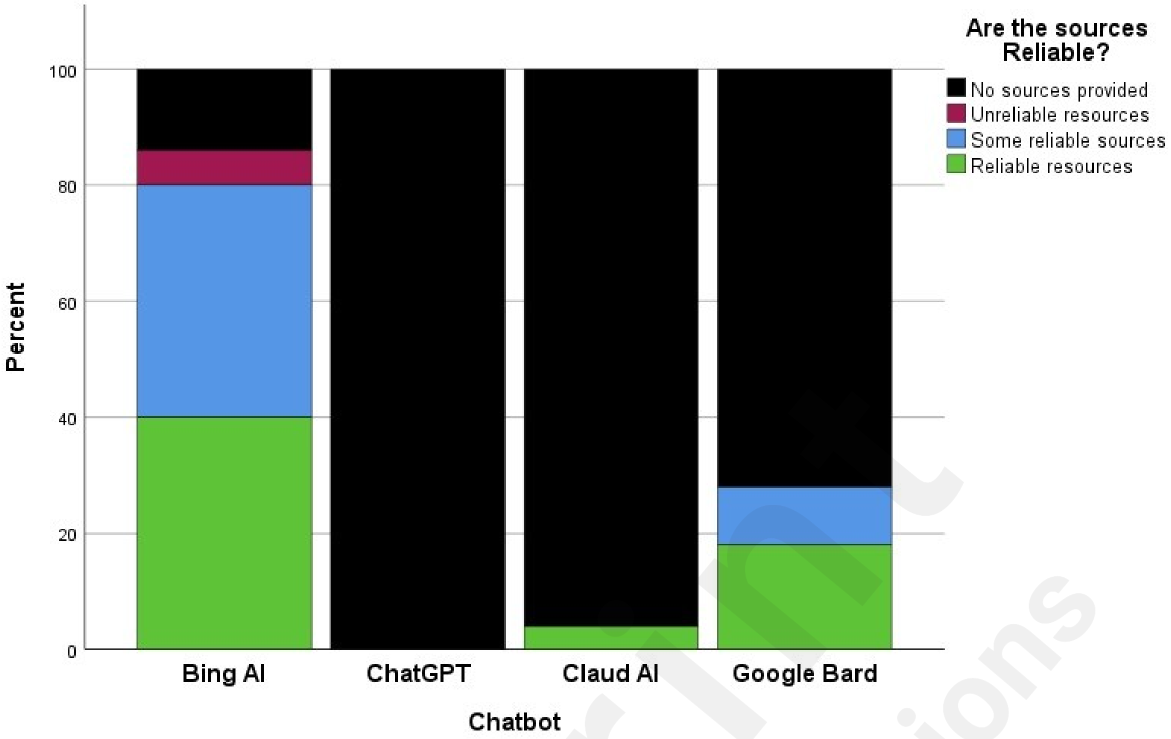
Because the order of conditions from most to least accurate, by the alternate percentage by quartile measure, closely matched the order of Figure 5, we elected not to present this percentile data.

As depicted in Figure 2, the aggregate measure of completeness across all 10 questions was similar, and ranged from 50-60% for each of the four chatbots (*P*=.848). However, completeness of response did vary depending on the clinical condition (*P*=.040), from a high of 79% for stomach pain to a low of 20% for bad cold. The responses that were most deficient, as defined by missing more than two important pieces of information, varied from a low of 6% (bad sore throat) to a high of 40% (bad cold). The responses from all 10 conditions were judged by some faculty to be missing 1-2 pieces of important information in up to 20% of scores, while 5-20% of scores identified more than two missing pieces of information. None of the 10 medical condition responses were judged by the faculty to contain all necessary information.

Source Reliability and Relevance

As seen graphically in Figure 2, source reliability and source relevance were outliers from the other six domains. As shown in Figure 6, ChatGPT was a further outlier from the other chatbots (*P*<.001), providing essentially no sources for its responses, while Google Bard was slightly better than Claud AI (*P*=.003). Only Bing AI provided what the EM faculty judged as generally reliable sources for a majority of responses to the clinical queries. We found similar results for both domains of source reliability and source relevance, so present only the former here.

Figure 6: Comparison between four chatbots for the domain of source reliability, as judged by five EM experts.



Readability

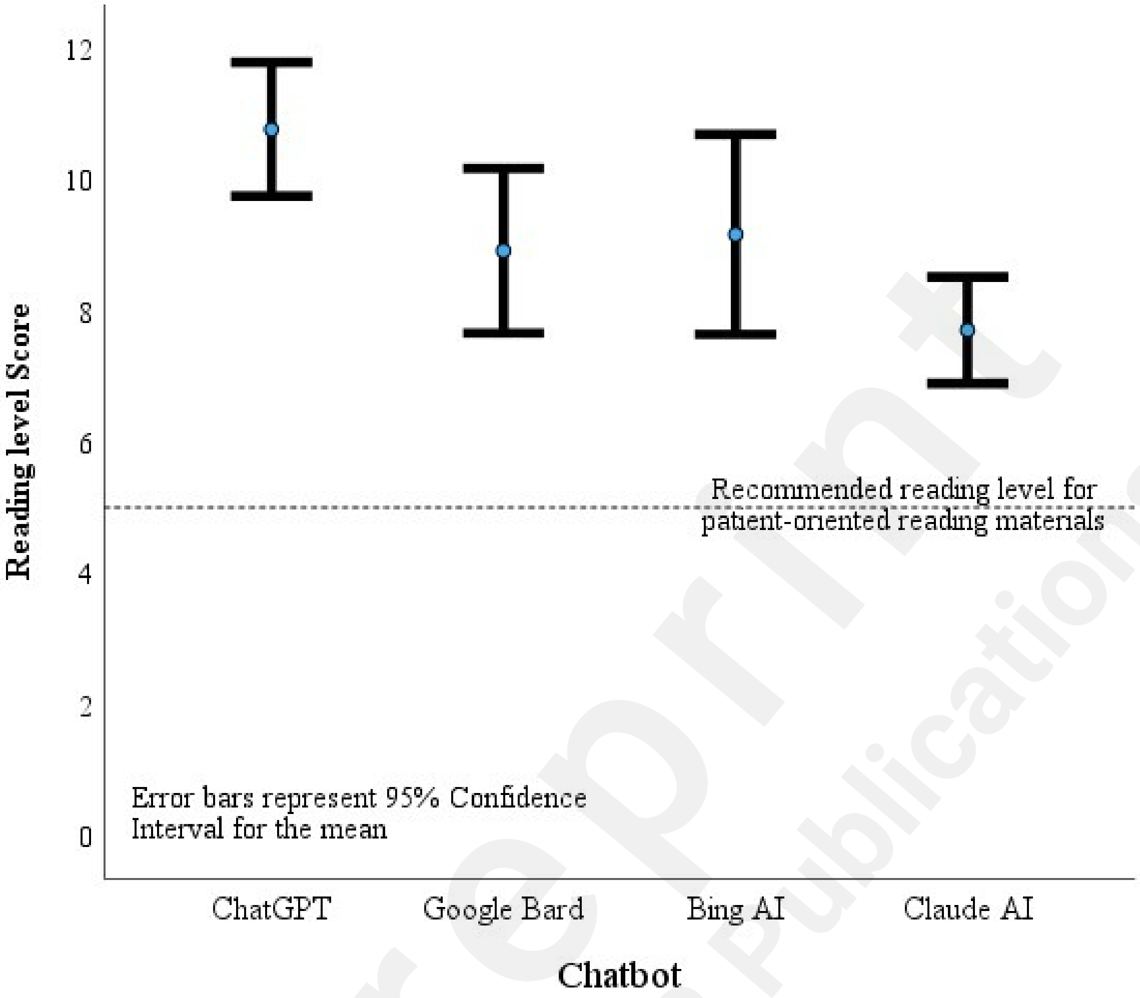
We processed all 10 clinical query responses from each chatbot through the embedded Readability Statistics Tool in Microsoft Word (FKGL) to calculate a global grade reading level for each chatbot. Table 1 presents the distribution of grade level and chatbot average. The mean grade level for Google Bard, Bing AI, and Claude AI was 7.7-9.16, with ChatGPT scoring 10.76 grade level. As shown in Figure 7, ChatGPT scored higher than Claude AI (*P*=.003). The grade level across all four chatbots was statistically similar as the error bars overlap (Figure 7). Within any of the three chatbots with lower overall grade-level reading scores, the level varied from a low of 5.9 to a high of 12.4. Of the 40 scores, the proportion of chatbot responses for all questions/conditions at or below each grade level was 0% (5th or below), 5% (6th), 15% (7th), 32.5% (8th), 52.5% (9th), 65% (10th), 80% (11th) and 90% at or below 12th grade.

Table 1. Emergency Care responses to 10 clinical queries from four chatbots: Flescher-Kincaid Grade Level (FKGL) scores from Microsoft Word

FKGL Reading Levels among all Four Chatbots (Microsoft Word)					
Medical Condition	ChatGPT	Google Bard	Bing AI	Claude AI	Mean
Chest pain	9.8	8.5	10	8.2	9.13
Stroke	8.6	7.5	6.9	7	7.50
Headache	10.9	9.8	9.8	7.1	9.40
Sore throat	10.7	8.2	6.3	7.6	8.20
Stomach pain	12.1	12.3	11.7	8	11.03
Back pain	12.4	7.6	11.3	7.2	9.63
Faints	8.5	6.3	6	5.9	6.68

Period bleeding	11.4	10.3	9.8	10.3	10.45
Bad cold	12.4	10.3	11.3	7.6	10.40
Overdose	10.8	8.3	8.5	8.1	8.93
Mean (chatbot)	10.76	8.91	9.16	7.70	9.13

Figure 7: Comparison of mean and 95% confidence interval (error bars) for reading level per Microsoft Word Fleisher-Kincaid Grade Level (FKGL) score for four chatbots as judged by EM faculty.



Therefore, for the 5th grade level recommended by the American Medical Association for patient-oriented reading materials, all of the chatbot responses failed this standard. For the average reading level of 6th grade for Medicaid-insured patients, the chatbot responses failed 95% of the time. Finally, the chatbots failed the US National Institutes of Health recommendation of 8th grade for reading materials 67.5% of the time.[17,32]

Discussion

We found that four chatbots had significant deficiencies in their performance compared to the opinions of five board-certified EM faculty and compared to optimal information drawn from respected mainstream EM texts and reliable websites. Problem areas included inadequate medical/scientific accuracy, incomplete information, dangerous information, and lack of source information.

The four chatbots were good and statistically similar in performance in clarity and understandability of presentation (to the educated EM faculty), intermediate in accuracy and completeness, but performed poorly for source disclosure and relevance. All four provided what the EM faculty considered dangerous advice with similar frequency. Grade level of English language presentation was above recommended norms for emergency patients (5th to 8th grade level) for all of them, with ChatGPT having the highest, and therefore potentially least user-friendly, grade-level assessment (10th grade).

Dangerous Information/Patient Safety

Because of the high stakes inherent in chatbot queries for emergency conditions, the authors were concerned about the potential for dangerous errors of omission and commission within the chatbot answers. These concerns were validated. There was wide variation on the accuracy of information depending on the clinical query and by chatbot. As all the chatbots had similar rates of dangerous advice and omissions of critical actions for some questions, we did not find any to be superior or inferior to others. We conclude, then, that chatbots may, if at all, be useful for minor conditions that are low risk (perhaps common cold), but should not be used for many EM complaints with potential for serious consequences and complications (chest pain, stroke, overdose).

Non-scientific information

Another aspect of veracity of chatbot responses is the degree to which they contain nonsensical/irrelevant information, without a scientific evidence base. Although our study design did not specifically ask the EM specialists to quantify these, we noted that responses contained such examples: gargling to relieve sore throat, using a heating pad for abdominal pain, eating vitamin-C-rich foods to “boost your immune system” (multiple mentions of this), and “stay warm, as cold air can worsen sore throat, and avoid strong odors.”

The EM faculty also noticed many instances of vague advice that may not be understood by laypersons. Examples include: “support the person’s breathing”; may need “triage” to a stroke center; and references to “neurologic symptoms” or “other concerning symptoms” relating to headache.

Reading Level and Understandability

For the specific use of chatbots in EM, readability/understandability is critical, perhaps beyond its importance for general medicine. Regardless of accuracy and completeness, if the consumer cannot readily comprehend the information from the chatbot query, this system of real-time health education fails on a fundamental level. Three of the chatbots had average responses written at 7th to 9th grade level, with ChatGPT statistically higher at almost 11th grade average. Even the “better” performing chatbots had some answers written at 10th, 11th, and 12th grade levels. The lowest grade was grade level 5.9, and only 6 of the 40 answers were written at or below the 7th grade level (see Table 1, Emergency Care Information: FKGL Reading Level). The authors conclude that these answers exceed the reading capability recommended in previous literature and US government recommendations and, therefore, may be unsuitable for potential emergency medical conditions.[15–17,32]

AI Chatbots and their Impact on Social Determinants of Health Disparities

As the move to mainstream society for AI progresses, the use of chatbots will increase. A 2022 survey revealed that 58% of Americans have already used AI to obtain health information.[8] Recently, AI chatbot capability has been installed in widely used internet browsers, making their availability ubiquitous, assuming internet access. Microsoft has embedded Bing Copilot within their browser, Edge, calling it your, “everyday AI companion” in promotional materials.[33] This has the potential to exacerbate disparities relating to social determinants of health, as internet access varies depending on socioeconomic status. Socially disadvantaged patients already have reduced access to in-person medical consultation or telemedicine, difficulties with office visit transportation, and reduced ability to navigate a complex medical system with insurance requirements and constraints.[34]

On one hand, lower socioeconomic/disadvantaged patients may have limited access to AI chatbot information, based on lower personal computer use, lack of internet access, language barriers, more limited general and health reading literacy, and simple unawareness of the chatbot tools themselves.[35,36] This may be, in some ways, protective from the misinformation we found in the AI responses. Conversely, these same patients may have limited resources, insurance, and a lower level of primary care access, and therefore may turn to AI Chatbots for medical advice more frequently.[9]

Another variable that may influence chatbot use is age and the resultant technological facility. Older patients with more complex medical histories may have greater need for information potentially provided by AI, yet, paradoxically, may have more difficulty with computer access and use.[10,37] Furthermore, this disparate access may be more pronounced in low-resource countries, where regulations requiring prescription or pharmacist intervention to obtain medications are generally lacking.[38,39] In these settings, the authors conjecture that such chatbot advice would be less likely to be checked with, or balanced by, information from a qualified human physician. As AI becomes ubiquitous, the direction of these possibilities of AI as an additional social determinant of health disparities demands further study.[34]

If AI chatbots were to reach their potential as correct, comprehensive, understandable and ubiquitous, it is possible that they could favorably impact health care costs, perhaps avoiding more costly in-person healthcare visits. While the authors recognize this potential, at this point, the current iteration of chatbots is far from capable of this potential healthcare cost reduction.[40,41]

Spectrum of queries from not likely dangerous to life-threatening, purposeful spectrum

We chose 10 questions for the chatbots and EM faculty to reflect some of the most common presenting complaints to American emergency departments (EDs).[18,42] We purposely chose conditions with both benign and serious possibilities, and hoped the responses would cover both ends of the acuity spectrum, and provide sound information for when, and in what time course, to seek further care. We found these features only inconsistently in each of the chatbots, with some queries containing such information, and some devoid of it. None of the chatbots performed consistently well on this measure.

Call for Regulation of Chatbots as “Medical Devices.”

We would no sooner accept approval of a new pacemaker or orthopedic implant without FDA clearance than we would allow wide access to incomplete, dangerous, nonscientific information to guide patients’ health care decisions. The authors recognize the enormous challenge of regulating words produced by these systems. Nevertheless, we would be remiss in not calling for progress in this area.

Both the European Parliament and the US federal government, as well as China’s Ministry of Industry and Information Technology have recognized the need for regulation of AI and put forth draft regulations. As these efforts are just beginning and very fluid, we direct the reader to these references for further information.[43–46]

Standardized Study Reporting Framework and Assessment Tools are Lacking

We searched for, and then recognized, that there are no standard tools to evaluate health advice generated by chatbots. Hence, we derived our tool from recent papers which studied AI medical advice in other settings,[5,6,13,26–30] and the expert opinions by EM faculty. We used both qualitative (minor and major missing information) and quantitative methods (percent of response that was correct) to judge the chatbots.

With the increasing reliance of both physicians and patients on the internet for health advice, previous authors have stressed the importance of a standardized format for reporting AI-related health advice. Huo and colleagues wrote of their concern that these chatbots are a risk for patient safety. Further, they wrote, and we agree, that chatbot information needs accuracy, and must avoid false and misleading information/sources. These authors have convened an international group of stakeholders to address the importance of developing reporting standards for studies of chatbots used for health advice.[25]

Comparisons to Previous Work

Our paper's methodology drafts from previous work, and substantially expands and improves on previous assessments of AI healthcare information.

A recent paper from *JAMA Oncology* shared some of our methodology; the current paper extended those methods and tailored our evaluation to EM. The *JAMA Oncology* paper used a DISCERN tool (published in 1999) to evaluate information on four common cancers, and graded accuracy, quality, uncertainty, and reliability, on a 1-5 Likert scale.[47] We refined and expanded this by identifying eight logical domains of information to assess the chatbot responses. We did not use a Likert scale purposely to limit subjectivity. Instead, we used specific percentages and numbers of inaccurate pieces of information, as judged by content experts in EM.

The previous paper's scoring was not apparently done by content experts. We had our EM experts compare the chatbot responses to "correct" answers drawn from reputable and widely-used sources. The *JAMA Oncology* authors similarly assessed the readability of the chatbot responses, and found overall 11-12 grade reading level, certainly too high for the average emergency patient.[47]

A more recent grading tool for online medical education information was published in 2016 by ALIEM and shared some of our domains for evaluation: accuracy, utility, evidence base, and reference quality.[26] They used a 1-7 Likert scale and similarly had multiple (8) experts evaluate each of the on-line resources. As these were geared toward physician learners, they did not evaluate readability.

Categories and Examples of Response Deficiencies:

The chatbot responses had both categorical and individual deficiency. The major categories were important omissions of critical information, advice to act with insufficient situational information, absence of information regarding when and how to access healthcare, vague or technical language not expected to be understandable by lay people, and extraneous information without scientific support.

There were components of some responses that were considered “omissions,” but were so important that the omissions were also considered “dangerous actions.” Examples include omitting a pulse check with chest pain, and yet advising to start CPR, and omitting any mention of lack of need or efficacy of antibiotics for the common cold. Other overtly dangerous actions were advising the use of a defibrillator without directing anyone to start CPR, and advice to move a patient who fainted to fresh air without checking for any injury. Some responses also made no mention of pregnancy, either ectopic or intrauterine, among the advice for vaginal bleeding. Finally, the advice for sore throat omitted any mention of symptoms of airway obstruction, like stridor.

Advice to consumers about when and how to seek medical care was inconsistent and vague. We recognize that EM faculty would have a bias toward expecting inclusion of “red-flag” symptoms. For example, for severe or persistent headache, advice to go to the emergency room for “concerning neurological symptoms” is too vague to be useful. Converseley, some responses mentioned serious conditions to watch for, but omitted any advice for common conditions with the same complaint. For example, Bing AI provided specific signs to watch for stroke or meningitis but did not give advice about migraine or tension headache. Finally, one chatbot advised in the possible stroke information, for the reader to start CPR, without assessing if the patient was unconscious or pulseless.

Sources of information for three of the four chatbots were not included in the responses. We asked each chatbot, “Please list all sources of information you referenced.”, and they generally responded, “I don’t have direct access to my training data or know where it came from. I was trained on a mixture of licensed data, data created by human trainers, and publicly available data.” Only one chatbot, Bing AI, provided references to “learn more,” but no specific sources from which the responses were drawn. While some “learn more” links were to reliable sources like the National Health Service in the UK, the USA CDC, Mayo and Cleveland Clinics, others were clearly not authoritative, for example medical news sources like MSNBC, proprietary information from various urgent cares, computer resources like Microsoft Start, and layperson sites like Verywell Health, and WikiHow.

Limitations

We conducted a pilot test of the scoring sheet/methodology on three of the included clinical conditions (bad cold, fainting, chest pain) and refined the scoring sheet and methodology with the content experts/respondents. We did not pilot all 10 queries. Despite including definitions for each question in the survey, there is still a risk that the EM faculty might interpret the questions differently. We did not examine the causes of discrepancies among the faculty in their assessments of source relevance and reliability. We did not do a cluster analysis for the five raters, leaving the potential that some were inherently more severe graders of AI than others. Some raters interpreted, “learn more,” at the end of the Bing AI chatbot response to be a reporting of sources from which the answers were derived, and some did not. We did not further study this variability in interpretation. We did not distinguish the clinical importance of the ten commonly asked emergency care questions.

CONCLUSION

AI Chatbots have important deficiencies for emergency medicine patient advice, despite their consistent performance. Advice for when to seek urgent/emergent care is frequently incomplete and inaccurate, and patients may be unaware of misinformation. Sources of information are not generally disclosed. Patients who use AI to guide their healthcare assume potential risk. Use of AI Chatbots for health information may worsen social determinants of health disparities, and should be subject to further research, refinement, and regulation. We strongly recommend proper medical consultation to prevent potential adverse outcomes. Finally, we call for further validation of scoring tools, and development of literature reporting guidelines.

Multimedia Appendix 1-4

Word doc

Acknowledgments

None.

Conflict Of Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FUNDING

The author(s) received no financial support for the research and authorship. Funding of open access publication charges provided by the University of California/JMIR Transformative Publishing Agreement.

Data Availability

Upon request to authors.

Authors' Contributions

JY designed and carried out the research method, collected the assessment data, and drafted, reviewed, and approved the final manuscript.
SS assisted with the research design, performed statistical analysis, drafted the results section, and reviewed and approved the final manuscript.
LSM served as the research project coordinator, assisted with the research design, and drafted, reviewed, and approved the final manuscript.
JSR assisted with the research design, served as one of the five assessors, and reviewed and approved the final manuscript.
JS assisted with the research design, served as one of the five assessors, and reviewed and approved the final manuscript.
AT prepared the “Correct” answers to the 10 questions from credible sources.
WW assisted with the research design, served as one of the five assessors, and reviewed and approved the final manuscript.
MIL served as the senior advisor of the project and one of the five assessors, designed the research method, and assisted with drafting and final approval of the manuscript.

References

1. Hoffman M. What is a chatbot + how does it work? The ultimate guide. Zendesk. 2020. Available from: https://www.zendesk.com/blog/what-is-a-chatbot/ [accessed Apr 15, 2024]
2. What Is a Chatbot? | IBM. Available from: https://www.ibm.com/topics/chatbots [accessed Apr 15, 2024]
3. Metz C, Grant N. Google Updates Bard Chatbot With 'Gemini' A.I. as It Chases ChatGPT. N Y Times 2023 Dec 6; Available from: https://www.nytimes.com/2023/12/06/technology/google-ai-bard-chatbot-gemini.html [accessed Mar 21, 2024]
4. Mehdi Y. Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. Off Microsoft Blog. 2023. Available from: https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/ [accessed Mar 19, 2024]
5. Kung TH, Cheatham M, Medenilla A, Sillos C, Leon LD, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health Public Library of Science; 2023 Feb 9;2(2):e0000198. doi: 10.1371/journal.pdig.0000198
6. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, Ayoub W, Yang JD, Liran O, Spiegel B, Kuo A. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023 Jul;29(3):721–732. PMID:36946005
7. Reddy S. Evaluating large language models for use in healthcare: A framework for translational value assessment. Inform Med Unlocked 2023 Jan 1;41:101304. doi: 10.1016/j.imu.2023.101304
8. Wang X, Cohen RA. Health Information Technology Use Among Adults: United States, July-December 2022. Hyattsville, MD: National Center for Health Statistics (U.S.); 2023 Oct. doi: 10.15620/cdc:133700
9. Can GenAI Help Make Health Care Affordable? Consumers Think So. Deloitte U S. Available from: https://www2.deloitte.com/us/en/blog/health-care-blog/2023/can-gen-ai-help-make-health-care-affordable-consumers-think-so.html [accessed Jan 5, 2024]
10. Hanebutt R, Mohyuddin H. The Digital Domain: A "Super" Social Determinant of Health. Prim Care 2023 Dec;50(4):657–670. PMID:37866838
11. Digital Access: A Super Determinant of Health. 2023. Available from: https://www.samhsa.gov/blog/digital-access-super-determinant-health [accessed Mar 18, 2024]
12. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, Chang S, Berkowitz S, Finn A, Jahangir E, Scoville E, Reese T, Friedman D, Bastarache J, van der Heijden Y, Wright J, Carter N, Alexander M, Choe J, Chastain C, Zic J, Horst S, Turker I, Agarwal R, Osmundson E, Idrees K, Kiernan C, Padmanabhan C, Bailey C, Schlegel C, Chambless L, Gibson M, Osterman T, Wheless L. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Res Sq 2023 Feb 28; doi: 10.21203/rs.3.rs-2566942/v1
13. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, Srinivasan N, Park J, Burch M, Watson R, Liran O, Samakar K. Assessing the Accuracy of Responses by the Language Model ChatGPT to Questions Regarding Bariatric Surgery. Obes Surg 2023 Jun;33(6):1790–1796. PMID:37106269
14. Jin Q, Leaman R, Lu Z. Retrieve, Summarize, and Verify: How Will ChatGPT Affect Information Seeking from the Medical Literature? J Am Soc Nephrol 2023 Aug 1;34(8):1302–1304. PMID:37254254
15. Powers RD. Emergency department patient literacy and the readability of patient-directed materials. Ann Emerg Med 1988 Feb;17(2):124–126. PMID:3337429

16. Rooney MK, Santiago G, Perni S, Horowitz DP, McCall AR, Einstein AJ, Jagsi R, Golden DW. Readability of Patient Education Materials From High-Impact Medical Journals: A 20-Year Analysis. J Patient Exp 2021 Mar 3;8:2374373521998847. PMID:34179407
17. Weiss BD, Blanchard JS, McGee DL, Hart G, Warren B, Burgoon M, Smith KJ. Illiteracy among Medicaid recipients and its relationship to health care costs. J Health Care Poor Underserved 1994;5(2):99–111. PMID:8043732
18. Arvig MD, Mogensen CB, Skjøt-Arkil H, Johansen IS, Rosenvinge FS, Lassen AT. Chief Complaints, Underlying Diagnoses, and Mortality in Adult, Non-trauma Emergency Department Visits: A Population-based, Multicenter Cohort Study. West J Emerg Med 2022 Oct 31;23(6):855–863. PMID:36409936
19. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah MH. Snakebite Advice and Counseling From Artificial Intelligence: An Acute Venomous Snakebite Consultation With ChatGPT. Cureus 2023 Jun;15(6):e40351. PMID:37456381
20. Fahy E. Quality of patient health information on the internet: reviewing a complex and evolving landscape. Australas Med J 2014 Feb 1;7(1):24–28. PMID:24567763
21. Liu J, Zhang Y, Kim Y. Consumer Health Information Quality, Credibility, and Trust: An Analysis of Definitions, Measures, and Conceptual Dimensions. Proc 2023 Conf Hum Inf Interact Retr New York, NY, USA: Association for Computing Machinery; 2023. p. 197–210. doi: 10.1145/3576840.3578331
22. Eysenbach G, Powell J, Kuss O, Sa E-R. Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide Web: A Systematic Review. JAMA 2002 May 22;287(20):2691. PMID:12020305
23. Robillard JM, Jun JH, Lai J-A, Feng TL. The QUEST for quality online health information: validation of a short quantitative tool. BMC Med Inform Decis Mak 2018 Dec;18(1):87. PMID:30340488
24. Breckons M, Jones R, Morris J, Richardson J. What Do Evaluation Instruments Tell Us About the Quality of Complementary Medicine Information on the Internet? J Med Internet Res 2008 Jan 22;10(1):e3. PMID:18244894
25. Huo B, Cacciamani GE, Collins GS, McKechnie T, Lee Y, Guyatt G. Reporting standards for the use of large language model-linked chatbots for health advice. Nat Med 2023 Dec;29(12):2988. PMID:37957381
26. Chan TM-Y, Grock A, Paddock M, Kulasegaram K, Yarris LM, Lin M. Examining Reliability and Validity of an Online Score (ALiEM AIR) for Rating Free Open Access Medical Education Resources. Ann Emerg Med 2016 Dec 1;68(6):729–735. PMID:27033141
27. Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O. Can ChatGPT, an Artificial Intelligence Language Model, Provide Accurate and High-quality Patient Information on Prostate Cancer? Urology 2023 Oct 1;180:35–58. PMID:37406864
28. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. Radiology 2023 Jun;307(5):e230922. PMID:37310252
29. Lahat A, Shachar E, Avidan B, Glicksberg B, Klang E. Evaluating the Utility of a Large Language Model in Answering Common Patients' Gastrointestinal Health-Related Questions: Are We There Yet? Diagn Basel 2023 Jun 2;13(11). PMID:37296802
30. Morath B, Chiriac U, Jaszkowski E, Deiss C, Nurnberg H, Horth K, Hoppe-Tichy T, Green K. Performance and risks of ChatGPT used in drug information: an exploratory real-world analysis. Eur J Hosp Pharm 2023 Jun 1; PMID:37263772
31. Klein D. Implementing a General Framework for Assessing Interrater Agreement in Stata. Stata J Promot Commun Stat Stata 2018 Dec;18(4):871–901. doi: 10.1177/1536867X1801800408
32. Weiss BD. Health literacy: A manual for clinicians. Am Med Assoc 2003; Available from: http://lib.ncfh.org/pdfs/6617.pdf

33. Mehdi Y. Announcing Microsoft Copilot, your everyday AI companion. Off Microsoft Blog. 2023. Available from: https://blogs.microsoft.com/blog/2023/09/21/announcing-microsoft-copilot-your-everyday-ai-companion/ [accessed Apr 16, 2024]
34. Ong JCL, Seng BJJ, Law JZF, Low LL, Kwa ALH, Giacomini KM, Ting DSW. Artificial intelligence, ChatGPT, and other large language models for social determinants of health: Current state and future directions. Cell Rep Med Elsevier; 2024 Jan 16;5(1). PMID:38232690
35. Adepoju OE, Dang P, Jacobs W, Baiden P. Long Way to Go: Attitudes, Knowledge, and Perception of Artificial Intelligence in Health Care, Among a Racially Diverse, Lower Income Population in Houston, New York, and Los Angeles. Popul Health Manag 2024 Feb;27(1):90–93. PMID:38100075
36. Swenson K, Ghertner R. People in Low-Income Households Have Less Access to Internet Services.
37. Wilson J, Heinsch M, Betts D, Booth D, Kay-Lambkin F. Barriers and facilitators to the use of e-health by older adults: a scoping review. BMC Public Health 2021 Aug 17;21(1):1556. PMID:34399716
38. Clement V, Lemus S, Newman E. In search of health: medical tourism at the US-Mexico border/lands. J Tour Cult Change Routledge; 2024 Jan 2;22(1):118–136. doi: 10.1080/14766825.2023.2248088
39. LaPelusa M, Verduzco-Aguirre H, Diaz F, Aldaco F, Soto-Perez-de-Celis E. Cross-border utilization of cancer care by patients in the US and Mexico – a survey of Mexican oncologists. Glob Health 2023 Oct 27;19(1):78. PMID:37891675
40. Hegeman P, Vader D, Kamke K, El-Toukhy S. Patterns of digital health access and use among US adults: A latent class analysis. Res Sq 2024 Jan 29;rs.3.rs-3895228. PMID:38352382
41. Sutton J, Gu L, Diercks DB. Impact of Social Determinants of Health, Health Literacy, Self-perceived Risk, and Trust in the Emergency Physician on Compliance with Follow-up. West J Emerg Med 2021 May 5;22(3):667–671. PMID:34125044
42. Estimates of Emergency Department Visits in the United States, 2016-2021. 2023. Available from: https://www.cdc.gov/nchs/dhcs/ed-visits/index.htm [accessed Apr 15, 2024]
43. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. White House. 2023. Available from: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ [accessed Apr 15, 2024]
44. Artificial Intelligence 2023 Legislation. Natl Conf State Legis. Available from: https://www.ncsl.org/technology-and-communication/artificial-intelligence-2023-legislation [accessed Apr 15, 2024]
45. Ye J. China issues draft guidelines for standardising AI industry. Reuters 2024 Jan 18; Available from: https://www.reuters.com/technology/china-issues-draft-guidelines-standardising-ai-industry-2024-01-17/ [accessed Apr 16, 2024]
46. EU AI Act: first regulation on artificial intelligence. Top Eur Parliam. 2023. Available from: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [accessed Apr 15, 2024]
47. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. JAMA Oncol 2023 Oct 1;9(10):1437–1440. PMID:37615960

Abbreviations

AI: Artificial Intelligence
EM: Emergency Medicine
FKGL: Flesch-Kincaid Grade Level
GenAI: Generative AI
Info: Information
LLMs: Large Language Models

Supplementary Files

Multimedia Appendixes

Contains 10 clinical questions, chatbot prompts, reference answers, and the scoring form.
URL: http://asset.jmir.pub/assets/07e4c5d65f56e16234ea06669053ab9f.docx