Information Research, Special Issue: Proceedings of the 15th ISIC - The Information Behaviour Conference, Aalborg, Denmark, August 26-29, 2024

Can ChatGPT provide health information as physicians do? Preliminary findings from a cross-sectional study of online medical consultation

Siqi Luo, Hongyi Qin, Hanlin Li, and Cui Huang

DOI: https://doi.org/10.47989/ir292837

Abstract

Introduction. ChatGPT has shown promise in medical consultation. This paper presents the preliminary findings of evaluating the capability of ChatGPT in responding to real-world patient questions from a patient point of view, using physician responses as a benchmark.

Method. 24 patient questions and physician responses were collected from a Chinese professional medical consultation platform. The corresponding ChatGPT responses were also collected. Five evaluators without medical background were given the patient questions and responses from both sources in random order. Evaluations were made in terms of the quality and presented empathy of collected responses.

Analysis. Evaluation scores were analysed using descriptive statistical method.

Results. Preliminary findings demonstrated that ChatGPT could be considered as a dependable source for acquiring useful health information. However, it was not able to present the feeling of empathy to patients compared with human physicians.

Conclusion. We recommend that physicians consider utilizing ChatGPT as a supplementary information source when addressing general medical consultations to improve the experience of seeking medical information for patients.

Introduction

In today's digital age, individuals seek health information regarding their personal concerns on the Internet to fulfil their information needs, especially when immediate support from professional physicians is unavailable (Zhao and Zhang, 2017). The advent of recent technologies, such as large language models, has spurred some individuals to turn to platforms like ChatGPT for medical advice (Choudhury and Shamszare, 2023). Therefore, this paper aims to evaluate ChatGPT's capability of providing health information.

Information communication in the context of medical consultation is demanding and sensitive (Seitz et al., 2022). Not only do the patients expect to acquire accurate and useful health information through the interaction with physicians, but they also need the physicians to express empathy and show human warmth (Fitzpatrick et al., 2017; B. Liu & Sundar, 2018). For example, a medical student who performed extraordinarily in all exams may not be qualified to be a good doctor. According to a study using online reviews to research the drivers of patient satisfaction, the level of information quality and empathy were found to be the main reasons (Shah et al., 2021).

ChatGPT represents a new generation of AI technologies driven by advances in large language models and is widely recognized for its ability to generate near-human-quality texts across a wide range of contexts (Dwivedi et al., 2023). Although the system was not developed to provide health care, it has shown promise in addressing patient questions (Ayers et al., 2023; Beets et al., 2023; Budler et al., 2023; Eriksen et al., 2023; Lee et al., 2023; J. Liu et al., 2023; Singhal et al., 2023; Walker et al., 2023). A survey investigating 607 adults in the United States revealed that ChatGPT users have used the system for health-related queries (n=44, 7.2%) (Choudhury and Shamszare, 2023). ChatGPT applied in medical consultation may save patients with minor health concerns from a visit to the doctor and allow clinicians to spend more time to treat patients who need a consultation at the most (Bibault et al., 2019; Budler et al., 2023; Li et al., 2023).

However, it is yet questionable in the capability of ChatGPT to provide health information, since previous research have mostly focused on rigid criteria such as accuracy using structured professional medical board exams instead of real-world patient questions (Jin et al., 2021). And although valuable, most of the assessments on the ability of AI technologies in medical consultation were conducted from the perspective of specialized physicians (Ayers et al., 2023; Bickmore et al., 2018), while the actual needs of the patients remained unexplored. Less is known about patient satisfaction of health information provided by ChatGPT. If patient questions could be responded by ChatGPT with high quality information as well as empathy, it might reduce the workload of physicians, freeing them up for those patients who have more urgent needs (Li et al., 2023; Rasu et al., 2015).

Therefore, this work-in-progress paper presents the preliminary findings into evaluating the capability of ChatGPT in responding to real-world patient questions from a patient point of view. For comparing purposes, the physician responses were applied as a benchmark. This cross-sectional study addressed the following research questions:

1. Generally, from the perspective of patient satisfaction, can ChatGPT provide health information as physicians do?

2. In terms of quality, can ChatGPT provide health information as physicians do?

3. In terms of empathy, can ChatGPT provide health information as physicians do?

Methods

The study took a quantitative approach by using a cross-sectional design. We collected patient questions and physician responses from Dingxiangyisheng (https://dxy.com/questions/), a Chinese professional medical consultation platform with over 5.5 million users and over 2 million licensed doctors, where each patient's question was answered by a certificated physician (Zhou et al., 2023). During November 2023, we selected general patient questions related to daily life including consultations about insomnia, diet, having a cold, HPV prevention, headache and so on. The physician responses were retained as a benchmark. Later, the original full texts of the patient question were put into a new chat with ChatGPT (version GPT-3.5, OpenAI), without prior questions asked to avoid potential bias, and the corresponding ChatGPT responses were saved. To ensure the comparability between physician response and ChatGPT response, only consultations with one back-and-forth and pure texts were collected. As a result, 24 patient questions and corresponding physician responses and ChatGPT responses were gathered. An independent surgeon checked all the ChatGPT responses and found that ChatGPT did not provide inaccurate information.

The original patient questions, physician responses and ChatGPT responses were reviewed by 5 postgraduate students without medical background from 3 universities in China, including 3 females and 2 males (not coauthors). All evaluators were enlisted through social media channels, and they participated on a voluntary basis, without any monetary compensation. Responses were randomly ordered and labeled response A and B to blind evaluators. Any revealing information (e.g., statements such as '*I'm an artificial intelligence*') was removed. The evaluators were asked to read the patient question and both responses before answering the following questions:

1. Assuming you are encountering medical problems as the question described, which response are you more satisfied with, response A or response B? And why?
2. Using a seven-point Likert scale, where higher values indicate greater extent, how would you rate the level of usefulness of the information provided, the level of capturing intent of the question, the level of professionality, the level of empathy, and the level of helpfulness of the information provided, for responses A and B (Ayers et al., 2023; Singhal et al., 2023)?

We compared the length of physician responses and ChatGPT responses. The mean length of physician responses was 433.25 words, whereas ChatGPT responses had a mean length of 347.33 words. Recognizing the potential impact of response length on evaluators' choice preference, we performed a McNemar's chi-square test to examine the association between response length and response preference selection. McNemar's chi-square test is particularly useful in assessing significant differences in paired observations between two paired samples. The results of the McNemar's chi-square test indicated that there was no significant relationship between response text length and the evaluators' choice of preference (χ^2= 0.06, p = 0.808). Consequently, it can be inferred that response length did not compromise the reliability of the evaluation outcomes.

Preliminary findings

ChatGPT has shown great potential to be a patient-satisfied health information provider

ChatGPT demonstrated considerable potential as a health information source, contributing to patient satisfaction. Among the 24 patient questions examined, evaluators perceived ChatGPT responses to be more satisfactory than those provided by physicians in 13 instances (54%).

Among the 24 patient questions, on the one hand, ChatGPT responses garnered unanimous or near-unanimous approval from all 5 evaluators in 6 questions (25%), indicating its potential to surpass physicians in eliciting patient satisfaction. In 10 instances (42%), ChatGPT exhibited comparable performance to physicians, with 2 or 3 out of 5 evaluators endorsing its responses. This implies its capacity to address medical inquiries effectively.

On the other hand, in the rest 8 instances (33%) of patient questions, physician responses were preferred, as evidenced by either no evaluators or only one evaluator selecting ChatGPT. This suggests that ChatGPT may exhibit suboptimal performance in specific circumstances, emphasizing the need for careful consideration of its limitations in providing health information. Caution is warranted in relying solely on ChatGPT for health information.

In terms of quality, ChatGPT can be a dependable source for providing useful information

ChatGPT exhibited a performance closely aligned with that from a physician, with only marginal differentials (Figure 1). Especially in the realm of providing useful information, ChatGPT demonstrated a superior performance, earning an average rating of 5.58 compared to physicians at 5.52 (Figure 1(a)).

While ChatGPT lagged slightly behind in capturing the patients' intent and professionality. Regarding capturing patients' intent in medical consultations, ChatGPT responses received a rating 0.08 lower than physicians (Figure 1(b)). Evaluators assigned a slightly higher mean score to physician responses (5.89) in terms of perceived professionality compared to ChatGPT responses (5.79) (Figure 1(c)).

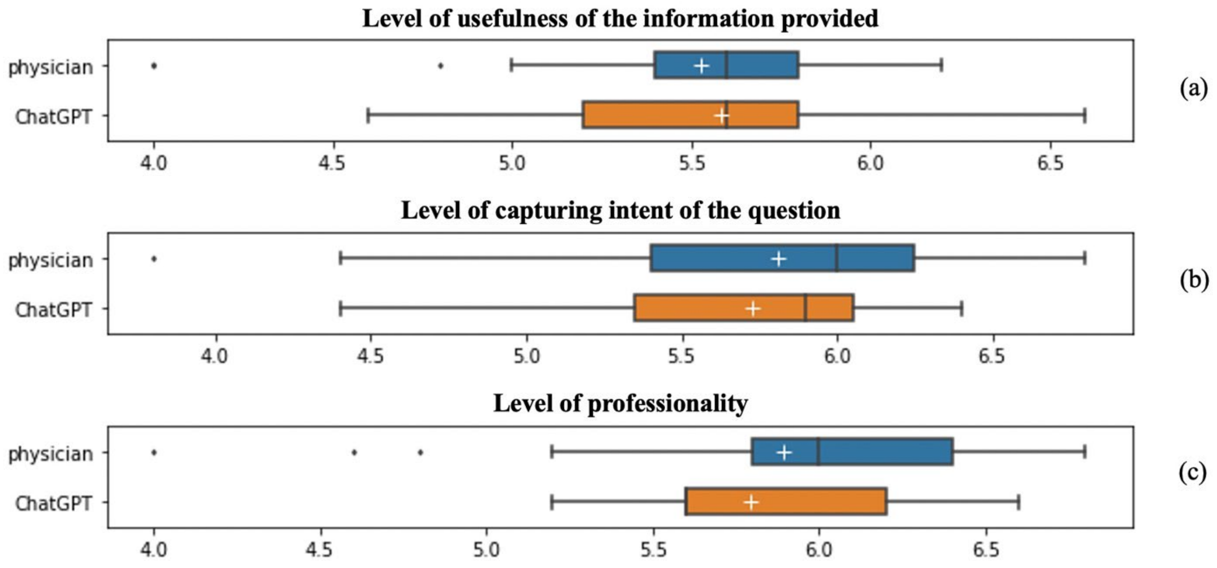

Figure 1. Evaluation of level of usefulness of the information provided, capturing intent of the question, and professionality

Note: The box in the plot represents the interquartile range (IQR), which spans from the 25th percentile to the 75th percentile of the data. Inside the box, a vertical line marks the median (50th percentile) of the data, and a white cross marks the average value of the data. The lines extending from the box, called whiskers, depict the range of the data outside the interquartile range. Individual data points that fall outside the whiskers are considered outliers and are plotted as individual points.

In terms of empathy, ChatGPT is still no match for physicians

Physicians demonstrated a greater degree of empathy and helpfulness in their responses to patient questions, when compared to ChatGPT (Figure 2). Specifically, ChatGPT responses received a rating in terms of empathy with a mean score of 5.38, 0.12 lower than physician responses (Figure 2(a)). What's more, when assessing the helpfulness of responses, ChatGPT lagged largely behind physicians, with a noticeable 0.3-point gap between the two (Figure 2(b)).

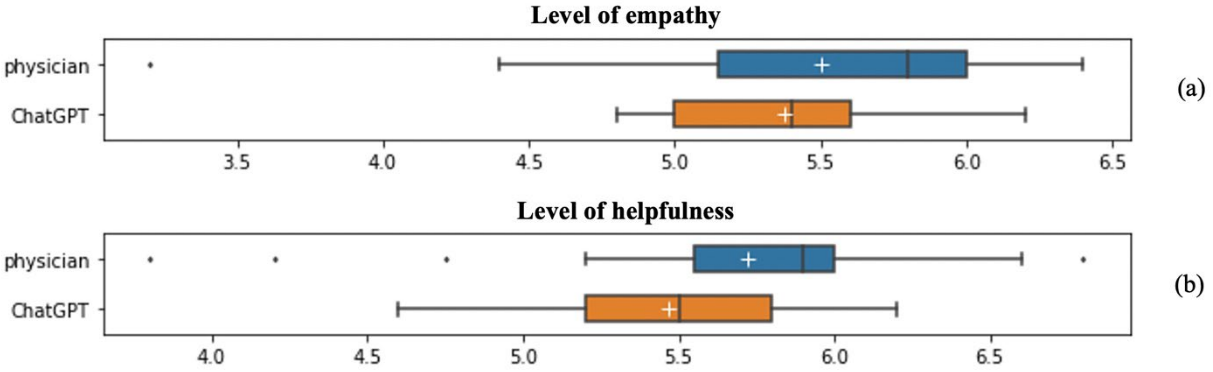

Figure 2. Evaluation of level of empathy and helpfulness

Discussion

This cross-sectional study evaluated the capability of ChatGPT from the perspective of the patients toward providing health information in the context of medical consultation. ChatGPT exhibited tremendous potential. In over fifty percent of patient questions, ChatGPT delivered responses that were more satisfactory. Regarding the quality of information, it paralleled the performance of human physicians, displaying only marginal disparities. Surprisingly, in the provision of useful information, ChatGPT even slightly surpassed human physicians. Nevertheless, in terms of empathy, ChatGPT still fell short of matching physicians.

Based on our findings, we posit that ChatGPT cannot serve as a replacement for a physician in the context of medical consultation currently, as AI is still unable to substitute for the full of warmth human interaction in healthcare. However, ChatGPT possesses the capability to gather targeted and useful health information for physicians. We suggest that physicians consider utilizing ChatGPT as a supplementary information source when addressing general medical consultations, which may save time for healthcare professionals, enhance efficiency, and contribute to an improved information experience for patients.

Limitations and outlook

This work-in-progress study subjects to limitations. First, the generalizability of study findings to other specific medical questions such as rare diseases is constrained. Only general patient questions were included for assessment in this study as these are the common medical problems in daily life, which are easy to understand by the evaluators. In addition, it would be premature to make conclusions without a broader range of patients' questions being analysed. At this point, the findings of our research serve as a directive indication. Second, the ideal evaluation of health information provided by ChatGPT from the perspective of patients should be accomplished by patients with real needs. However, due to the limit of the medical consultation platform where we acquired patient questions and physician responses, it was unable for us to reach out the patients who actually raised the original questions. Furthermore, our findings are based on only five evaluators drawn from a limited population of postgraduate students. This is not representative of the population as a whole.

Some of the limitations will be addressed in the ongoing research. The research focus can be extended to other medical questions to get a complete evaluation in the context of medical consultation. In terms of ideal evaluators, experimental approaches could be applied. In this way, participants are allowed to interact with ChatGPT about their real health concerns, and thus strengthen the objectivity of the evaluation. Besides, with the help of interview, new findings about the strengths and weaknesses of ChatGPT in providing health information may emerge.

ChatGPT applied in medicine is promising, but the bar for clinical applications is high (Singhal et al., 2023). Future research might evaluate ChatGPT performance from other dimensions like equity and bias.

Acknowledgement

We appreciate the anonymous reviewers for their valuable suggestions. We also thank our evaluators for their precious time and energy. This work is supported by the Key Program of the National Nature Science Foundation of China [grant number: 72134007] and the National Key R&D Program of China [grant number: 2022ZD0116203].

About the authors

Siqi Luo is a PhD candidate in the Department of Information Resources Management at Zhejiang University. Her research focuses on health information behaviour. She can be contacted at siqi_luo@zju.edu.cn.

Hongyi Qin is a PhD candidate in the School of Public Affairs at Zhejiang University. Her research focuses on digital health, particularly focusing on nudging early detection of disease using social experimental methods. She can be contacted at zoeqhy@zju.edu.cn.

Hanlin Li is a PhD candidate in the Department of Information Resources Management at Zhejiang University. His research interests include human-AI collaboration. He can be contacted at 12122095@zju.edu.cn.

Cui Huang is a professor in the School of Public Affairs at Zhejiang University. Her main area of research is in information resources management and digital governance. She can be contacted at huangcui@zju.edu.cn.

References

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA *Internal Medicine*, 183(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838

Beets, B., Newman, T. P., Howell, E. L., Bao, L., & Yang, S. (2023). Surveying public perceptions of artificial intelligence in health care in the United States: systematic review. *Journal of Medical Internet Research*, 25(1), e40337. https://doi.org/10.2196/40337

Bibault, J.-E., Chaix, B., Guillemassé, A., Cousin, S., Escande, A., Perrin, M., Pienkowski, A., Delamon, G., Nectoux, P., & Brouard, B. (2019). A chatbot versus physicians to provide information for patients with breast cancer: blind, randomized controlled noninferiority trial. *Journal of Medical Internet Research*, 21(11), e15787. https://doi.org/10.2196/15787

Bickmore, T. W., Trinh, H., Olafsson, S., O'Leary, T. K., Asadi, R., Rickles, N. M., & Cruz, R. (2018). Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. *Journal of Medical Internet Research*, 20(9), e11510. https://doi.org/10.2196/11510

Budler, L. C., Gosak, L., & Stiglic, G. (2023). Review of artificial intelligence-based question-answering systems in healthcare. *WIREs Data Mining and Knowledge Discovery*, 13(2), e1487. https://doi.org/10.1002/widm.1487

Choudhury, A., & Shamszare, H. (2023). Investigating the impact of user trust on the adoption and use of ChatGPT: survey analysis. *Journal of Medical Internet Research*, 25(1), e47184. https://doi.org/10.2196/47184

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Eriksen, A. V., Möller, S., & Ryg, J. (2023). Use of GPT-4 to diagnose complex clinical cases. *NEJM AI*, 1(1), AIp2300031. https://doi.org/10.1056/AIp2300031

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Mental Health*, 4(2), e7785. https://doi.org/10.2196/mental.7785

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), Article 14. https://doi.org/10.3390/app11146421

Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*, 388(13), 1233–1239. https://doi.org/10.1056/NEJMsr2214184

Li, Y., Liang, S., Zhu, B., Liu, X., Li, J., Chen, D., Qin, J., & Bressington, D. (2023). Feasibility and effectiveness of artificial intelligence-driven conversational agents in healthcare interventions: A systematic review of randomized controlled trials. *International Journal of Nursing Studies*, 143, 104494. https://doi.org/10.1016/j.ijnurstu.2023.104494

Liu, B., & Sundar, S. S. (2018). Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking*, 21(10), 625–636. https://doi.org/10.1089/cyber.2018.0110

Liu, J., Wang, C., & Liu, S. (2023). Utility of ChatGPT in clinical practice. *Journal of Medical Internet Research*, 25(1), e48568. https://doi.org/10.2196/48568

Rasu, R. S., Bawa, W. A., Suminski, R., Snella, K., & Warady, B. (2015). Health literacy impact on national healthcare utilization and expenditure. *International Journal of Health Policy and Management*, 4(11), 747–755. https://doi.org/10.15171/ijhpm.2015.151

Seitz, L., Bekmeier-Feuerhahn, S., & Gohil, K. (2022). Can we trust a chatbot like a physician? A qualitative study on understanding the emergence of trust toward diagnostic chatbots. *International Journal of Human-Computer Studies*, 165, 102848. https://doi.org/10.1016/j.ijhcs.2022.102848

Shah, A. M., Yan, X., Tariq, S., & Ali, M. (2021). What patients like or dislike in physicians: Analyzing drivers of patient satisfaction and dissatisfaction using a digital topic modeling approach. *Information Processing & Management*, 58(3), 102516. https://doi.org/10.1016/j.ipm.2021.102516

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), Article 7972. https://doi.org/10.1038/s41586-023-06291-2

Walker, H. L., Ghani, S., Kuemmerli, C., Nebiker, C. A., Müller, B. P., Raptis, D. A., & Staubli, S. M. (2023). Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. *Journal of Medical Internet Research*, 25(1), e47479. https://doi.org/10.2196/47479

Zhao, Y., & Zhang, J. (2017). Consumer health information seeking in social media: A literature review. *Health Information & Libraries Journal*, 34(4), 268–283. https://doi.org/10.1111/hir.12192

Zhou, X., Xiong, H., & Xiao, B. (2023). A physician recommendation algorithm based on the fusion of label and patient consultation text. *Information Science*, 41(3), 145–154. https://doi.org/10.13833/j.issn.1007-7634.2023.03.017