

Research Letter

Quality and Accountability of ChatGPT in Health Care in Low- and Middle-Income Countries: Simulated Patient Study

Yafei Si¹, PhD; Yuyi Yang², MSc; Xi Wang³, MSc; Jiaqi Zu⁴, MSc; Xi Chen^{5,6}, PhD; Xiaojing Fan⁷, PhD; Ruopeng An^{3,8}; Sen Gong⁹, PhD

¹UNSW Business School and CEPAR, The University of New South Wales, Kensington, Australia

²Division of Computational and Data Sciences, Washington University in St Louis, St. Louis, MO, United States

³Brown School, Washington University in St Louis, St Louis, MT, United States

⁴Global Health Research Center, Duke Kunshan University, Kunshan, China

⁵Department of Health Policy and Management, Yale University, New Haven, CT, United States

⁶Department of Economics, Yale University, New Haven, CT, United States

⁷School of Public Policy and Administration, Xi'an Jiaotong University, Xi'an, China

⁸Silver School of Social Work, New York University, New York, NY, United States

⁹Centre for International Studies on Development and Governance, Zhejiang University, Hangzhou, China

Corresponding Author:

Xiaojing Fan, PhD

School of Public Policy and Administration

Xi'an Jiaotong University

28 West Xianning Road

Xi'an, 710049

China

Phone: 86 15891725861

Email: emirada@163.com

Abstract

Using simulated patients to mimic 9 established noncommunicable and infectious diseases, we assessed ChatGPT's performance in treatment recommendations for common diseases in low- and middle-income countries. ChatGPT had a high level of accuracy in both correct diagnoses (20/27, 74%) and medication prescriptions (22/27, 82%) but a concerning level of unnecessary or harmful medications (23/27, 85%) even with correct diagnoses. ChatGPT performed better in managing noncommunicable diseases than infectious ones. These results highlight the need for cautious AI integration in health care systems to ensure quality and safety.

(*J Med Internet Res* 2024;26:e56121) doi: [10.2196/56121](https://doi.org/10.2196/56121)

KEYWORDS

ChatGPT; generative AI; simulated patient; health care; quality and safety; low- and middle-income countries; quality; LMIC; patient study; effectiveness; reliability; medication prescription; prescription; noncommunicable diseases; AI integration; AI; artificial intelligence

Introduction

The rise of generative artificial intelligence (AI) models like ChatGPT is transforming the health care landscape, especially in low- and middle-income countries (LMICs). These regions, often facing shortages of health care professionals, are increasingly turning to AI tools for medical consultation, aided by growing internet and smartphone access [1]. Research has highlighted generative AI use in the fields of cardiology [2] and orthopedic diseases [3]. However, there are concerns about the

accuracy and safety of AI models like ChatGPT [4] given their lack of legal or professional accountability. This is crucial in medical settings, where precise and reliable decision-making is vital. Our study focuses on assessing ChatGPT's performance in treatment recommendations for common diseases in LMICs, addressing a critical need for the responsible application of AI in health care.

Methods

Overview

We used the simulated patient (SP) method to create a realistic testing environment for ChatGPT with GPT-3.5 from August 8 to 19, 2023. SPs are healthy individuals trained to consistently mimic real patients and their symptoms [5]. We trained the SPs to present 9 common, previously validated diseases [5-8]. We asked ChatGPT to act as a doctor in an LMIC and offer consultations. The SPs detailed their primary concerns, gave standardized responses to every question, and recorded all diagnoses and medication recommendations, which were cross-referenced with clinical guidelines to assess their accuracy and appropriateness. For a robust analysis, we presented each disease to ChatGPT 3 times. We conducted descriptive analyses with the final sample of 27 independent trials.

Ethical Considerations

The Ethics Committee of the First Affiliated Hospital of Xi'an Jiaotong University approved the study (LLSBPJ-2024-WT-019).

Results

Surprisingly, ChatGPT's performance varied across trials for each disease (Figure 1). When aggregating the results (Figure 2), ChatGPT had a 67% (18/27) success rate in initial diagnoses and a 59% (16/27) success rate in medication recommendations. When considering all recommendations, these rates increased to 74% (20/27) for any correct diagnosis and 82% (22/27) for any appropriate medication recommendation. However, there was a high rate of unnecessary or harmful medication suggestions, occurring in 85% (23/27) of trials overall and in 59% (16/27) of trials after a correct diagnosis. Our study also highlighted ChatGPT's varying performance across different types of diseases. Specifically, the AI demonstrated a superior ability in handling noncommunicable diseases compared to infectious diseases, both in terms of diagnosis and medication recommendations.

Figure 1. Heatmap comparing ChatGPT's responses with clinical guidelines. The asterisks (*) indicate infectious diseases; green cells denote correct or appropriate diagnoses or drug prescriptions; blue cells denote incorrect or unnecessary diagnoses or drug prescriptions; and red cells denote harmful drug prescriptions. Each row represents an independent trial.

Disease cases	Diag. 1	Diag. 2	Diag. 3	Diag. 4	Diag. 5	Diag. 6	Drug 1	Drug 2	Drug 3	Drug 4	Drug 5	Drug 6	Advice
Unstable angina	Green	Blue	Blue				Green	Green	Green	Green	Green	Blue	Green
Postpartum depression	Green						Red	Red					Green
Child diarrhea*	Blue						Blue	Green	Blue				Green
Type II diabetes	Green						Green	Green	Blue	Blue	Blue		Green
Pharyngitis	Green	Blue	Blue	Blue	Blue		Green	Green	Green	Green	Green		Green
Asthma	Green	Blue	Blue				Green	Green	Red	Green			
Pulmonary tuberculosis*	Blue	Blue	Blue	Blue			Red	Blue	Blue	Blue			Green
Genital herpes*	Green	Blue	Blue	Blue	Blue		Green	Red	Blue	Blue			Green
Syphilis*	Green	Blue	Blue	Blue	Blue	Blue	Green	Green	Blue				Green

Figure 2. ChatGPT’s capability in diagnosing and treating 9 common diseases. The asterisks (*) indicate infectious diseases; green denotes socially desired outcomes; red denotes undesired outcomes; darker colors denote higher probabilities.

Case	Disease presentation	Correct diagnosis		Correct drug		Unnecessary/harmful drug	
		First recommendation	Any recommendation	First recommendation	Any recommendation	Unconditional	Conditional on correct diagnosis
1	Unstable angina	100%	100%	100%	100%	100%	100%
2	Postpartum depression	100%	100%	0%	0%	100%	100%
3	Child diarrhea*	0%	0%	0%	67%	100%	0%
4	Type II diabetes	100%	100%	100%	100%	100%	100%
5	Pharyngitis	100%	100%	100%	100%	0%	0%
6	Asthma	67%	100%	100%	100%	67%	67%
7	Pulmonary tuberculosis*	0%	0%	0%	0%	100%	0%
8	Genital herpes*	67%	67%	67%	67%	100%	67%
9	Syphilis*	67%	100%	67%	100%	100%	100%
Noncommunicable diseases		93%	100%	80%	80%	73%	73%
Infectious diseases		33%	42%	33%	58%	100%	42%
Overall		67%	74%	59%	82%	85%	59%

Discussion

Our findings reveal a high level of accuracy in both correct diagnoses (74%) and medication recommendations (82%) by ChatGPT. Previous studies using the SP method found that primary care providers in LMICs like China, India, and Kenya could only reach correct diagnoses in 12%-52% of SP visits [5,6]. Therefore, ChatGPT can potentially outperform traditional primary care providers in LMICs in diagnostic accuracy. Since ChatGPT with GPT-3.5 is free, the AI tool has the potential to offer affordable and far-reaching solutions in LMICs, particularly in rural and underserved areas.

However, ChatGPT tended to suggest more unnecessary or even harmful medications (in 85% of trials) than primary care providers (28%-64%) [5,6]. AI models work by analyzing available data using machine learning and deep learning techniques [9]. Their approach to drug prescription can be aggressive due to a lack of professional accountability or a motive to reduce medical expenses. ChatGPT also performed better in managing noncommunicable diseases than infectious diseases. This could be because more information on the former is available for AI training during development [10]. ChatGPT’s

performance also varied within each disease case, contrary to our expectation that this would be more standardized.

We acknowledge several limitations. First, a broader array of diseases, especially those specific to different regions, should be used in future studies. Second, we did not introduce more details (ie, location) to avoid the prompts becoming overcomplicated, and by default, ChatGPT’s responses reflect the average population to increase its generalizability. Third, we did not account for the relative importance of the AI’s questions and emotional communications. Fourth, a larger sample size may have enabled us to perform head-to-head comparisons between AI care and traditional care.

Despite the limitations, we present the first audit-study evidence to evaluate ChatGPT’s performance in diagnosing and treating common diseases in LMICs. A rich set of 9 established diseases makes our findings highly relevant to and widely applicable in LMICs. ChatGPT reaches high levels of accuracy in diagnosis and medication recommendations, but also recommends a concerning level of unnecessary or harmful medications. Integrating AI tools like ChatGPT into health care systems in LMICs may potentially improve diagnostic accuracy but also raises concerns about care safety.

Acknowledgments

No funding was available to support this study. XC acknowledges financial support from the Drazen scholarship and the Aden scholarship dedicated to research on Chinese health care systems. YS and SG acknowledge the support from the National Social Science Foundation of China (23AZD091). During the preparation of this work the authors used ChatGPT with GPT-4 in order to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

YS contributed to conceptualization, investigation, analysis, and writing (original draft); YY contributed to analysis, investigation, review, and editing; XW contributed to analysis, investigation, review, and editing; JZ contributed to analysis, investigation, review, and editing; XC contributed to review and editing; XF contributed to review and editing; RA contributed to conceptualization, investigation, analysis, and writing; and SG contributed to review and editing. All authors approved the final version of the paper.

Conflicts of Interest

None declared.

References

1. Howarth J. How many people own smartphones? (2023-2028). Exploding Topics. URL: <https://explodingtopics.com/blog/smartphone-stats> [accessed 2023-12-07]
2. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA*. Mar 14, 2023;329(10):842-844. [FREE Full text] [doi: [10.1001/jama.2023.1044](https://doi.org/10.1001/jama.2023.1044)] [Medline: [36735264](https://pubmed.ncbi.nlm.nih.gov/36735264/)]
3. Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, et al. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. *J Med Internet Res*. Sep 15, 2023;25:e47621. [FREE Full text] [doi: [10.2196/47621](https://doi.org/10.2196/47621)] [Medline: [37713254](https://pubmed.ncbi.nlm.nih.gov/37713254/)]
4. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of Using ChatGPT in health care. *J Med Internet Res*. Aug 11, 2023;25:e48009. [FREE Full text] [doi: [10.2196/48009](https://doi.org/10.2196/48009)] [Medline: [37566454](https://pubmed.ncbi.nlm.nih.gov/37566454/)]
5. Kwan A, Daniels B, Bergkvist S, Das V, Pai M, Das J. Use of standardised patients for healthcare quality research in low- and middle-income countries. *BMJ Glob Health*. 2019;4(5):e001669. [FREE Full text] [doi: [10.1136/bmjgh-2019-001669](https://doi.org/10.1136/bmjgh-2019-001669)] [Medline: [31565413](https://pubmed.ncbi.nlm.nih.gov/31565413/)]
6. Si Y, Bateman H, Chen S, Hanewald K, Li B, Su M, et al. Quantifying the financial impact of overuse in primary care in China: a standardised patient study. *Soc Sci Med*. Mar 2023;320:115670. [doi: [10.1016/j.socscimed.2023.115670](https://doi.org/10.1016/j.socscimed.2023.115670)] [Medline: [36669284](https://pubmed.ncbi.nlm.nih.gov/36669284/)]
7. Xue H, D'Souza K, Fang Y, Si Y, Liao H, Qin WA, et al. Direct-to-consumer telemedicine platforms in China: a national market survey and quality evaluation. Preprints with The Lancet. Preprint posted online Oct 18, 2021. [doi: [10.2139/ssrn.3944587](https://doi.org/10.2139/ssrn.3944587)]
8. Si Y, Xue H, Liao H, Xie Y, Xu D, Smith M, et al. The quality of telemedicine consultations for sexually transmitted infections in China. *Health Policy Plan*. Mar 12, 2024;39(3):307-317. [doi: [10.1093/heapol/czad119](https://doi.org/10.1093/heapol/czad119)] [Medline: [38113375](https://pubmed.ncbi.nlm.nih.gov/38113375/)]
9. Sellamuthu S, Vaddadi S, Venkata S, Petwal H, Hosur R, Mandala V, et al. AI-based recommendation model for effective decision to maximise ROI. *Soft Comput*. 2023:1-10. [FREE Full text] [doi: [10.1007/S00500-023-08731-7](https://doi.org/10.1007/S00500-023-08731-7)]
10. Sanders JW, Fuhrer GS, Johnson MD, Riddle MS. The epidemiological transition: the current status of infectious diseases in the developed world versus the developing world. *Sci Prog*. 2008;91(Pt 1):1-37. [FREE Full text] [doi: [10.3184/003685008X284628](https://doi.org/10.3184/003685008X284628)] [Medline: [18453281](https://pubmed.ncbi.nlm.nih.gov/18453281/)]

Abbreviations

AI: artificial intelligence

LMIC: low- and middle-income country

SP: simulated patient

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 06.01.24; peer-reviewed by D Simmons, N Domingues, W Yang; comments to author 06.04.24; revised version received 21.04.24; accepted 30.07.24; published 09.09.24

Please cite as:

Si Y, Yang Y, Wang X, Zu J, Chen X, Fan X, An R, Gong S

Quality and Accountability of ChatGPT in Health Care in Low- and Middle-Income Countries: Simulated Patient Study

J Med Internet Res 2024;26:e56121

URL: <https://www.jmir.org/2024/1/e56121>

doi: [10.2196/56121](https://doi.org/10.2196/56121)

PMID: [39250188](https://pubmed.ncbi.nlm.nih.gov/39250188/)

©Yafei Si, Yuyi Yang, Xi Wang, Jiaqi Zu, Xi Chen, Xiaojing Fan, Ruopeng An, Sen Gong. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 09.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.