I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17204

Quality and Accountability of Large Language Models (LLMs) in Healthcare in Low- And Middle-Income Countries (LMIC): A Simulated Patient Study Using ChatGPT

Yafei Si
Yuyi Yang
Xi Wang
Ruopeng An
Jiaqi Zu
Xi Chen
Xiaojing Fan
Sen Gong

AUGUST 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17204

Quality and Accountability of Large Language Models (LLMs) in Healthcare in Low- And Middle-Income Countries (LMIC): A Simulated Patient Study Using ChatGPT

Yafei Si
University of New South Wales

Xi Chen
Yale University and IZA

Yuyi Yang, Xi Wang, Ruopeng An
Washington University in St. Louis

Xiaojing Fan
Xi'an Jiaotong University

Jiaqi Zu
Duke Kunshan University

Sen Gong
Zhejiang University

AUGUST 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Quality and Accountability of Large Language Models (LLMs) in Healthcare in Low- And Middle-Income Countries (LMIC): A Simulated Patient Study Using ChatGPT

Using simulated patients to mimic nine established non-communicable and infectious diseases over 27 trials, we assess ChatGPT's effectiveness and reliability in diagnosing and treating common diseases in low- and middle-income countries. We find ChatGPT's performance varied within a single disease, despite a high level of accuracy in both correct diagnosis (74.1%) and medication prescription (84.5%). Additionally, ChatGPT recommended a concerning level of unnecessary or harmful medications (85.2%) even with correct diagnoses. Finally, ChatGPT performed better in managing non-communicable diseases compared to infectious ones. These results highlight the need for cautious AI integration in healthcare systems to ensure quality and safety.

JEL Classification: C0, I10, I11, C90

Keywords: ChatGPT, Large Language Models, generative AI, simulated patient, healthcare, quality, safety, low- and middle-income countries

Corresponding author:
Ruopeng An
Brown School
Washington University in St Louis
One Brookings Dr, St Louis
MO 63130
USA
E-mail: ruopeng@wustl.edu

Introduction

The rise of generative artificial intelligence (AI), exemplified by models like ChatGPT, is transforming healthcare landscapes, especially in low- and middle-income countries (LMICs). These regions, often facing healthcare professional shortages, are increasingly turning to AI tools for medical consultation, aided by growing internet and smartphone access [1,2]. Research has highlighted the effectiveness of generative AI in fields such as cardiology [3], anaesthesiology [4], orthopaedic diseases [5], and oncology [6]. However, there are concerns about the accuracy and safety of AI models like ChatGPT [7], given their lack of legal or professional accountability. This is particularly crucial in medical settings where precise and reliable decision-making is vital. Our study is focused on assessing the effectiveness and reliability of ChatGPT in diagnosing and treating common diseases in LMICs, addressing a critical need for responsible AI application in healthcare.

Methods

We employed the method of simulated patient (SP) to create a realistic testing environment for the free version of ChatGPT 3.5 from August 8 to 19 in 2023. SPs are healthy individuals trained to consistently mimic real patients and their symptoms. The SP method is increasingly recognised as a "gold standard" to evaluate the quality of care in LMICs [8], and it also has several comparative advantages for the project. First, SPs ensure uniform scenarios with the illness and optimal care pre-defined, allowing for direct comparison of physician practices against clinical guidelines. Second, SPs offer consistency in symptom presentation, controlling for variation in patient preferences and communication styles. Third, using SPs negates the risks associated with testing new AI technology on real patients.

We trained SPs to present nine common diseases, both non-communicable and infectious, which have been validated in previous research [8–10]. These diseases, often encountered in clinical settings, include unstable angina, postpartum depression, child diarrhoea, type II diabetes, pharyngitis, asthma, pulmonary tuberculosis (TB), genital herpes, and syphilis.

We asked ChatGPT to act as a doctor in LMICs and offer consultation to SPs. Each SP script detailed the patient's primary concern (e.g., experiencing chest pain recently) and standardised responses to every possible question posed by ChatGPT. SPs meticulously recorded all diagnoses, medication recommendations, and medical advice provided by ChatGPT. For a robust analysis, we presented each

disease case to ChatGPT three times, ensuring that the AI model did not carry over its understanding from one trial to another. This process resulted in 27 independent trials. To evaluate ChatGPT's performance, these responses were cross-referenced with standard clinical guidelines, assessing the accuracy and appropriateness of both diagnosis and treatment.

Results

It is surprising that ChatGPT's performance varied across trials for each disease (Figure 1). When aggregating the results (Table 1), ChatGPT had a 66.7% success rate (18 out of 27) in initial diagnoses and a 59.3% success rate (16 out of 27) in recommending appropriate medication. When considering all recommendations, these rates increased to 74.1% (20 out 27) for any correct diagnoses and 81.5% (22 out of 27) for any appropriate medication recommendations. However, there was a high incidence of unnecessary or harmful medication suggestions, occurring in 85.2% (23 out of 27) of the trials. Even among correct diagnoses, ChatGPT recommended such medications in 59.3% (16 out of 27) of trials. Our study also highlighted ChatGPT's varying performance across different types of diseases. Specifically, the AI demonstrated a superior ability in handling non-communicable diseases compared to infectious ones, both in terms of diagnosis and medication prescription.

Discussion

Our findings reveal a high level of accuracy in both correct diagnosis (74.1%) and medication prescription (81.5%) by ChatGPT. Using the similar SP method, previous studies found that primary care providers in LMICs like China, India, and Kenya can only reach correct diagnoses in 12-52% of SP visits [8,9]. Therefore, ChatGPT can potentially outperform traditional primary care providers in LMICs in diagnostic accuracy, although we cannot make more detailed comparisons at the current stage. ChatGPT could be a valuable healthcare tool, particularly in diagnostics and treatment planning. Since ChatGPT 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's tendency to suggest unnecessary or even harmful medications (85.2%) is also higher than the 28-64% found in previous similar SP studies [8,9]. The unnecessary care is often influenced by physicians' financial incentives within a fee-for-service system [9], while AI in medical consultation works by analysing patient records, medical literature, clinical trials, and drug databases,

employing techniques like natural language processing, machine learning, and deep learning [11]. Our findings suggest that the AI's approach to drug prescription can be very aggressive, possibly due to a lack of legal or professional accountability and presumably also lacking a sense of saving medical expenses.

Moreover, ChatGPT's performance varied across disease types, with better results in managing non-communicable diseases compared to infectious ones. One main explanation is that ChatGPT was trained in developed contexts, where infectious diseases are less common than non-communicable diseases [12]. It is more surprising that ChatGPT's performance varied within each disease case, since the answer from ChatGPT should be more standardised. These results emphasize the importance of tailoring AI tools to fit the unique health profiles and needs of different regions and underscore the necessity for stringent oversight and thorough validation in the clinical use.

We acknowledge several limitations in the study. First, the nine diseases, mostly selected for SP presentations, may not represent the scope of all common diseases in LMICs. Second, we did not introduce more details such as geographical locations and medical institutions of SP visits to make the pilot study over-complexed. By default, ChatGPT replied to SP presentations at the average level. Third, we did not account for the relative important of the AI's questions and emotional communications, while the two parts are important to systematically understand ChatGPT’s reasoning process and communication styles. Fourth, this pilot study yielded 27 independent trials between SPs and ChatGPT, while a larger sample size may enable us to perform head-to-head comparisons between AI care and traditional care. These highlights a need for future research to ensure broader applicability of the findings.

Despite the limitations, we present the first audit-study evidence to evaluate ChatGPT’s effectiveness and reliability in diagnosing and treating common diseases in LMICs. ChatGPT reaches a high level of accuracy in both correct diagnosis and medication prescription, and a concerning level of unnecessary or harmful medications even with correct diagnoses. Integrating AI tools like ChatGPT into healthcare systems in LMICs may potentially improve their diagnostic accuracy but also raise more concerns about care safety. Therefore, it would be valuable to emphasize the necessity of enhanced regulation and

rigorous validation of AI tools in healthcare, as well as encourage further investigation into the care of AI tools in various contexts to ensure their quality and safety in clinical practice.

Figure 1 Heatmap of comparing ChatGPT's responses with clinical guidelines.

Note: Green grids denote correct or appropriate diagnoses or drug prescriptions; blue grids denote incorrect or unnecessary diagnoses or drug prescriptions; red grids denote harmful drug prescriptions. Each row represents an independent trial.

Table 1 ChatGPT's capability in diagnosing and treating nine common diseases.

Case No.	Disease presentation	Correct diagnosis		Correct drug		Unnecessary / Harmful drug	
		The 1st recomm.	Any recomm.	The 1st recomm.	Any recomm.	Unconditional	Conditional on correct diag.
1	Unstable angina	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
2	Postpartum depression	100.0%	100.0%	0.0%	0.0%	100.0%	100.0%
3	Child diarrhoea*	0.0%	0.0%	0.0%	66.7%	100.0%	0.0%
4	Type II diabetes	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
5	Pharyngitis	100.0%	100.0%	100.0%	100.0%	0.0%	0.0%
6	Asthma	66.7%	100.0%	100.0%	100.0%	66.7%	66.7%
7	Pulmonary tuberculosis*	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
8	Genital herpes*	66.7%	66.7%	66.7%	66.7%	100.0%	66.7%
9	Syphilis*	66.7%	100.0%	66.7%	100.0%	100.0%	100.0%
Non-communicable diseases		93.3%	100.0%	80.0%	80.0%	73.3%	73.3%
Infectious diseases		33.3%	41.7%	33.3%	58.3%	100.0%	41.7%
Overall		66.7%	74.1%	59.3%	81.5%	85.2%	59.3%

Note: * indicates infectious disease; recomm. denotes recommendation; green colour denotes socially desired outcome while red colour undesired outcome; darker colours denote higher probabilities.

References

1. Howarth J. How Many People Own Smartphones (2023-2028). Explod Top. 2023. Available from: https://explodingtopics.com/blog/smartphone-stats [accessed Dec 7, 2023]

2. Guo J, Li B. The Application of Medical Artificial Intelligence Technology in Rural Areas of Developing Countries. Health Equity 2018 Aug;2(1):174–181. doi: 10.1089/heq.2018.0037

3. 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 2023 Mar 14;329(10):842–844. doi: 10.1001/jama.2023.1044

4. Aldridge MJ, Penders R. Artificial intelligence and anaesthesia examinations: exploring ChatGPT as a prelude to the future. Br J Anaesth Elsevier; 2023 Aug 1;131(2):e36–e37. PMID:37244834 https://doi.org/10.1016/j.bja.2023.04.033

5. Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, Fujita K. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. J Med Internet Res JMIR Publications Toronto, Canada; 2023;25:e47621. doi:10.2196/47621

6. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. Radiology 2023 May 1;307(4):e230424. doi: 10.1148/radiol.230424

7. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res JMIR Publications Toronto, Canada; 2023;25:e48009. doi: 10.2196/48009

8. 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 BMJ Specialist Journals; 2019;4(5):e001669. https://doi.org/10.1136/bmjgh-2019-001669

9. Si Y, Bateman H, Chen S, Hanewald K, Li B, Su M, Zhou Z. Quantifying the financial impact of overuse in primary care in China: A standardised patient study. Soc Sci Med Elsevier; 2023;115670. https://doi.org/10.1016/j.socscimed.2023.115670

10. Xue H, D'Souza K, Fang Y, Si Y, Liao H, Qin WA, Yip W, Xu DR, Gong W, Chen W. Direct-to-Consumer Telemedicine Platforms in China: A National Market Survey and Quality Evaluation. 2021; http://dx.doi.org/10.2139/ssrn.3944587

11. Sellamuthu S, Vaddadi SA, Venkata S, Petwal H, Hosur R, Mandala V, Dhanapal R, singh J. AI-based recommendation model for effective decision to maximise ROI. Soft Comput Springer; 2023;1–10. https://doi.org/10.1007/s00500-023-08731-7

12. 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 Mar;91(1):1–37. doi: 10.3184/003685008X284628

Funding

Xi Chen acknowledges financial support from the Drazen scholarship and the Aden scholarship dedicated to research on Chinese healthcare systems. The views expressed are those of the authors and not necessarily those of the funders. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication. The IZA Discussion Paper Series serves as a preprint server to deposit latest research for feedback.

Data Availability

Data are fully available in the supplementary for download. The SP Scripts (including background and dialog) have been published and are available in Xue et al. (2021).

Author Contributions

Yafei Si: Conceptualization, Investigation, Analysis, Writing – Original Draft; Yuyi Yang: Analysis, Investigation, Review & Editing; Xi Wang: Analysis, Investigation, Review & Editing; Jiaqi Zu: Analysis, Investigation, Review & Editing; Xi Chen: Review & Editing. Xiaojing Fan: Review & Editing. Ruopeng An: Conceptualization, Investigation, Analysis, Writing. Sen Gong: Review & Editing. During the preparation of this work the authors used ChatGPT 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. All authors approved the final version of the paper.

Declaration of Interests

The authors have no conflicts of interest to declare.

Patient and Public Involvement

Patients or the public WERE NOT involved in the design, or conduct, or reporting, or dissemination plans of our research.