Language-adaptive artificial intelligence: assessing CHATGPT'S answer to frequently asked questions on total hip arthroplasty questions

Muhammad Talal Ibrahim1, Sarah Ashraf Khaskheli2, Hania Shahzad3, Shahryar Noordin4

1,4Department of Surgery, Section of Orthopaedics, Aga Khan University Hospital, 2 Medical College, Aga Khan University Hospital, Karachi, Pakistan, 3Department of Orthopaedics, UC Davis Health, Sacramento, California, USA.

Correspondence: Muhammad Talal Ibrahim
Email: mtalalibrahim@gmail.com
ORCID ID. 0000-0002-6687-0418

Abstract
ChatGPT is reported to be an acceptable tool to answer a majority of frequently asked patient questions. ChatGPT also converses in other languages including Urdu, which offers immense potential for the education of Pakistani patients. Therefore, this study evaluated ChatGPT’s Urdu answers to the ten most frequently asked questions on Total Hip Arthroplasty, which were then rated by an expert. Out of 10 answers in English, 9 (90%) were satisfactory requiring minimal clarification and 1 (10%) was satisfactory requiring moderate clarification. In both Roman and Nastaliq script Urdu, 1 (10%) answer was satisfactory requiring moderate clarification, while 9 (90%) were unsatisfactory requiring substantial clarification. In conclusion, as opposed to ChatGPT English responses, Urdu responses were much less rigorous, generic, and lacked scientific rigor. We have a long way to go before Pakistani patients with limited English language skills could benefit from AI chatbots like ChatGPT.

Keywords: Arthroplasty, Replacement, Hip, Artificial Intelligence, ChatGPT, Patient knowledge, Language Barrier.

DOI: https://doi.org/10.47391/JPMA.AKU-9S-25

Introduction
Patients are increasingly seeking medical information online, which may be used for self-diagnosis, self-treatment, triaging the need to visit a doctor, have background information on their disease before a doctor’s visit, or to have a second opinion on a doctor’s advice. Most of this online medical information is available in the English language. This is particularly true for trusted sources of medical information, for example, websites of trusted hospitals, international organizations, and government agencies. Hence, patients who do not understand the English Language remain at a disadvantage. This, compounded by low health literacy in Pakistan,1 presents a humongous challenge in making high-quality health education material available to the Pakistani population. Access to health is a basic right, and providing high-quality health education material in the patient’s language is paramount for ensuring equitable healthcare access and good outcomes.

Moreover, language barriers are a major concern in providing high-quality equitable healthcare. These barriers increase the time required for consultation, have higher chances of miscommunication, and compromise patient safety, leading to poor quality of consultation and lower patient satisfaction.2 In high-income countries, hospitals arrange for qualified medical translators for patients with language barriers. However, due to the high cost and logistical issues, doctors in Pakistan rely on anyone having any level of command over the languages as translators. Not only are these ‘translators’ not trained for medical translation, but a lot of times they are relatives of the patients. This may potentially lead to the loss of important information in translation and compromise the patient’s privacy in the case of a relative translator.3

ChatGPT is an artificial intelligence (AI) powered chatbot that has been trained on large volumes of data. Multiple studies have found that a majority of ChatGPT answers to frequently asked patient questions were reliable enough.4 ChatGPT has also been found to pass the United States Medical Licensing Exam’s practice test,5 reflecting its technical abilities. Almost all studies evaluating the role and accuracy of ChatGPT in answering medical questions have been in English Language. However, the information in English can also be found on multiple trusted sources online, which might have been the source of ChatGPT’s reliability.

ChatGPT also converses in other languages including Urdu. This offers immense potential to solve the problem of language barrier6 and provision of healthcare educational material in our local language(s). However, no study has evaluated the efficacy of ChatGPT’s Urdu answers. Therefore, this study aimed to evaluate...
ChatGPT’s Urdu answers to frequently asked questions on Total Hip Arthroplasty (THA).

**Patients and methods**

This study was conducted in October and November 2023 at Aga Khan University, Karachi, Pakistan, and University of California Davis Health, Sacramento, California, USA. We did a qualitative text analysis of the answers provided by ChatGPT-3.5. The questions were adapted from the ten most frequently asked questions on THA from published literature, and translated the questions from English to Urdu, in both Roman and Nastaliq scripts. Since the purpose of this study was to simulate patients using ChatGPT to get these answers, only forward translation of the questions was done by the authors, who were all well-versed in English and Urdu. These questions were then used as prompts for OpenAI’s GPT-3.5 to generate free-text responses. To simulate real-life patient interaction, no context was provided to ChatGPT, and no restrictions were placed on the response. There were no follow-up questions or repetition. The response generated for each question was collected by copying and pasting it into a Word document. An expert Orthopaedic Surgeon with more than 15 years of experience rated the responses in the following categories: “excellent response not requiring clarification”, “satisfactory requiring minimal clarification”, “satisfactory requiring moderate clarification” and “unsatisfactory requiring substantial clarification.”

**Results**

Out of 10 answers in English, 9 (90%) were satisfactory requiring minimal clarification and 1 (10%) was satisfactory requiring moderate clarification. In both Roman and Nastaliq script Urdu, 1 (10%) answer was satisfactory requiring moderate clarification, while 9 (90%) were unsatisfactory requiring substantial clarification. (Table 1)

ChatGPT’s Urdu responses were relatively better to the question on the duration of THA. This could be because it is relatively a simpler question that could be answered without any technical medical terms. The worst-performing English answer was the question about the success rate of THA. The reason for poor performance on this question could be due to a great variation in the technical details of THA being performed (primary versus secondary, etc.) and the role of patient factors in determining the success of the procedure. (Table 2).

**Discussion**

In this first study evaluating ChatGPT’s Urdu responses, we found significant problems in the quality of information provided. As opposed to information provided in English, Urdu responses were much less rigorous, generic, and lacked scientific rigor. Moreover, we also recognized major errors of grammar and vocabulary in both the Roman and Nastaliq scripts. It also mixed vocabulary from the Hindi language in its responses, particularly in the Roman script.

As opposed to its success in providing good-quality medical information in English, as found in our study and previous literature, ChatGPT-3.5 cannot be recommended to patients who speak and understand Urdu. ChatGPT’s responses in Urdu for generic non-medical prompts also have errors in grammar and vocabulary, reflecting that the problem is not limited to technical questions. Although, with time, we anticipate that these errors are not very hard to overcome. AI’s transformative continuous self-learning will improve its ability to improve its Urdu language skills.

Another major challenge that we anticipate is the lack of Urdu vocabulary for a majority of medical terms. For example, it is impossible to technically translate “Intramedullary femur nailing” in Urdu, without losing scientific accuracy. In practice, clinicians in Pakistan use their judgment to translate medical terms based on the patient’s level of education, health literacy, and ability to understand the information provided. Oftentimes, this means generic information, forgoing specificity in medical terms. We anticipate that AI chatbots like ChatGPT will need a lot of time and training to be able to serve patients in Urdu Language. This will probably require collaborative initiatives with local clinicians, to

| Table-1: Evaluation of responses provided by ChatGPT on ten frequently asked questions about Total Hip Arthroplasty. |
|---|---|---|---|
| **Answer Evaluation** | **English (N = 10 Questions)** | **Urdu (Roman Script) (N = 10 Questions)** | **Urdu (Nastaliq Script) (N = 10 Questions)** |
| Excellent response not requiring clarification | 0 (0%) | 0 (0%) | 0 (0%) |
| Satisfactory requiring minimal clarification | 9 (90%) | 0 (0%) | 0 (0%) |
| Satisfactory requiring moderate clarification | 1 (1%) | 1 (1%) | 1 (1%) |
| Unsatisfactory requiring substantial clarification | 0 (0%) | 9 (90%) | 9 (90%) |
provide transcripts of clinical interactions, and mechanisms to gauge patients’ requirements in the context of their education level, health literacy, and health information requirements. Additionally, it will also require cultural and religious contextualization, especially for sensitive healthcare topics like contraception, genetic testing, abortion, and sexual history, to name a few.

The language barrier is a well-known reason for inequitable access to healthcare.8 AI is being proposed to find possible solutions to fight poverty and improve the conditions in low and middle-income countries9,10. Moreover, AI tools like ChatGPT have also been proposed as a bridge for language barriers in academic healthcare6. However, our study highlights that Urdu-speaking patients may not be the immediate beneficiary of this revolutionary solution. We need significant collaborative work to be done by the healthcare and AI industries to enable Pakistani patients and the healthcare system to benefit from this technology.

Future studies should include patients’ perspectives on ChatGPT, particularly in their local languages, and how patients interact with it. These may include observing real-time patients interacting with ChatGPT and assessing their pre- and post-interaction knowledge. Moreover, healthcare-specific AI tools might perform better in clinical settings, particularly with continuous self-learning and real-time data from patient interaction. Limitations of this study include single-expert and single-disease evaluation of responses and use of ChatGPT-3.5 version, which is not linked to the internet. However, issues identified in ChatGPT’s Urdu capabilities are less likely to be changed with the involvement of more experts or questions on more diseases. Secondly, ChatGPT-4.0, although connected to the internet and hence probably better than ChatGPT-3.5, is subscription-based and hence less likely to be used by patients who have limited English language skills.

**Conclusion**

In conclusion, ChatGPT promises a better experience to English speakers as opposed to Urdu speakers, and we have a long way to go before Pakistani patients with limited English language skills could benefit from AI.
chatbots like ChatGPT.

**Disclaimer:** None.

**Conflict of Interest:** None.

**Funding disclosure:** None.

**References**


