Evaluating the comprehension and accuracy of ChatGPT's responses to diabetes-related questions in Urdu compared to English

Seyreen Faisal[1] [ID] **, Tafiya Erum Kamran[1], Rimsha Khalid[1], Zaira Haider[1],
Yusra Siddiqui[1], Nadia Saeed[2], Sunaina Imran[1], Romaan Faisal[3]
and Misbah Jabeen[4]**

Abstract

Introduction: Patients with diabetes require healthcare and information that are accurate and extensive. Large language models (LLMs) like ChatGPT herald the capacity to provide such exhaustive data. To determine (a) the comprehensiveness of ChatGPT's responses in Urdu to diabetes-related questions and (b) the accuracy of ChatGPT's Urdu responses when compared to its English responses.

Methods: A cross-sectional observational study was conducted. Two reviewers experienced in internal medicine and endocrinology graded 53 Urdu and English responses on diabetes knowledge, lifestyle, and prevention. A senior reviewer resolved discrepancies. Responses were assessed for comprehension and accuracy, then compared to English.

Results: Among the Urdu responses generated, only two of 53 (3.8%) questions were graded as comprehensive, and five of 53 (9.4%) were graded as correct but inadequate. We found that 25 of 53 (47.2%) questions were graded as mixed with correct and incorrect/outdated data, the most significant proportion of responses being graded as such. When considering the comparison of response scale grading the comparative accuracy of Urdu and English responses, no Urdu response (0.0%) was considered to have more accuracy than English. Most of the Urdu responses were found to have an accuracy less than that of English, an overwhelming majority of 49 of 53 (92.5%) responses.

Conclusion: We found that although the ability to retrieve such information about diabetes is impressive, it can merely be used as an adjunct instead of a solitary source of information. Further work must be done to optimize Urdu responses in medical contexts to approximate the boundless potential it heralds.

Introduction

Type 2 diabetes (T2D) accounts for about 90% of diabetes cases in the world.[1] The global prevalence of T2D has increased fivefold from 108 million in 1980 to 537 million in 2022.[2] In 2022, it was revealed that 26.7% of the adult population in Pakistan have diabetes, amounting to a staggering 33 million people.[3] The South Asian populace, including Pakistan, hosts an ethnic group that has a

[1]Shifa College of Medicine, Shifa Tameer-e-Millat University, Islamabad, Pakistan
[2]Department of Internal Medicine, Shifa College of Medicine, Shifa Tameer-e-Millat University, Islamabad, Pakistan
[3]Islamabad Medical and Dental College, Shaheed Zulfiqar Ali Bhutto Medical University, Islamabad, Pakistan
[4]Department of Endocrinology, Shifa International Hospital, Islamabad, Pakistan

Corresponding author:
Seyreen Faisal, Shifa College of Medicine, Shifa Tameer-e-Millat University, NCBMS Tower, Sector H-8/4, Service Road S, Near Federal Board Office, H 8/4 H-8, Islamabad, Islamabad Capital Territory 46000, Pakistan.
Email: seyreenfaisal@gmail.com

predisposition to developing diabetes, with poorer disease management, diet higher body mass indexes, and sedentary lifestyles; the group also has a simultaneous and increased association with complications and mortality rates.[4]

Pakistan and its geographic region have been slated to have constrained resources, especially specific to the care of aged patients in terms of human resources, health funding, and social protection.[5] These factors coupled with the inattention to health awareness and deficiency of health literacy contribute to the exponential rise in diabetes prevalence within the region. Many studies have been dedicated to delineating barriers to diabetes awareness and self-help; much of which has been attributed to demographic and cultural details of patients from South Asian regions.[4]

These barriers tend to encompass a myriad of factors, including a need for more knowledge regarding the condition, its management, and the role of diet and medications in controlling sugar levels. Some sects of the population in the study described a hesitance to seek advice and disclose their status as people with diabetes, for fear of societal implications.[4] Patients have progressively turned to the internet for health information, with an astonishing 80% of US adults opting to garner information through this avenue.[1] With the plethora of easily available information, a considerable proportion of which can be inaccurate and harmful, this remains an imminent problem to resolve.[2]

Additionally, certain patients felt that language barriers and not being fluent in English impeded their ability to seek information regarding their disease. Even patients who articulated well in English found existing literature on diabetes to be far too complex to comprehend easily.[4] Given the intricacy of care required to manage a prevailing condition such as diabetes, it is imperative that patient health and the patient's capacity to solicit advice and information not be hindered. Literature has illustrated how a high level of compliance can substantially improve quality of life, as well as diminish long-term complications.[6] The extent of a patient's knowledge concerning their disease plays a role instrumental to the self-management of their condition. It is widely adjudged that a patient well versed in their disease course and management will be less prone to the more critical aggravations of diabetes.[7,8]

Developments within the scope of artificial intelligence and language processing have ushered in the era of large language models (LLMs) such as the widely used and easily accessible ChatGPT 3.5. LLMs have been utilized to answer various questions related to various diseases, such as diabetes.[9] The progression of LLMs like ChatGPT has developed to an extent wherein it can impart answers to an expansive list of questions related to health and diseases.[10] LLMs like ChatGPT can offer valuable discernment concerning clinical diagnostics and analysis. It can augment clinical practice and become an asset to individuals hesitant to consult doctors.[11] ChatGPT can retrieve relevance from the internet almost immediately, regurgitating information in a conversational and idiomatic way.[11] While considering the barriers illustrated prior, it would seem that LLMs can serve as an indispensable, easily comprehensible, and swiftly accessible source of information for patients with diabetes. However, while ChatGPT's abilities remain impressive, it necessitates the need to proceed with caution. ChatGPT cannot act as a singular source of professional medical advice, diagnosis, and treatment, requiring further study and clarity.[12]

The barriers patients face coupled with ChatGPT's emerging proficiency for disseminating information highlight a need for crucial evaluation for this tool. There has been research that has studied ChatGPT's proficiency in dispelling diabetes misconceptions in English.[13] There is limited research on the assessment of ChatGPT in its proficiency and accuracy in answering questions in languages other than English. There have been studies done analyzing ChatGPT's capabilities in answering cirrhosis-related questions in Arabic, concluding that ChatGPT has the potential to act as a supplementary source of information for Arabic-speaking individuals.[14]

To date, there is no literature prevalent in appraising ChatGPT's ability to assess the accuracy and comprehension in answering diabetes-related questions in Urdu. In regions like Pakistan, wherein there is a concerning rise in diabetes cases, limited resources, and information specific to patients in their language, it entails the requirement to conduct our study to investigate the utility of ChatGPT in our region.

Materials and methods

Ethical considerations

The research was conducted at Shifa International Hospital after approval was taken from the Institutional Review Board and Ethics Committee (IRB and EC) of the institute mentioned above (IRB: 121-24). The study was administered over 2 months from March to May 2024.

Data source

We set out to conduct a cross-sectional observational study. We curated and shortlisted 53 commonly asked diabetes-related questions by referring to other studies as a blueprint.[11,13] We also correlated these questions with those frequently asked in support groups found on Facebook and those frequented in clinical practice. Institutional Review Board and Ethics approval was attained for questions acquired from clinical practice on 22 March 2024.

These questions were tailored to the disease and regional population. Questions were modified in aspects of wording to make them more accurate. The same questions were then translated into Urdu by the authors who were fluent in both

languages. The Urdu translation was then approved independently by two individuals with a PhD in Urdu language.

Exclusion criteria included questions that were vague and had no specific answers (e.g. how does diabetes affect my body?), questions that would vary from person to person (e.g. what is the chance my diabetes worsens?), and non-medical questions about the condition (e.g. what support groups are available for diabetes in Pakistan?) (Figure 1).

The questions included were then stratified into three domains of inquiry: basic knowledge, lifestyle, and preventative.

ChatGPT and response generation

ChatGPT is a natural language processing model, and its first research preview prototype was released in November 2022.[15] It is a part of the GPT-3.5 LLM subtype and is programed to have information from an expanse of online sources updated till January 2022.[16] ChatGPT has been trained to appropriately respond to information that is fed to it that is apt within the context in which it is asked.[11] This is regarded as reinforcement learning from human feedback or reinforcement learning from human preference (RLHF/RLHP). Users can input any information, and ChatGPT will reply in a pertinent and conversational fashion, its response based on the information inputted into its database.

We entered questions into the March 2024 version of ChatGPT Model 3.5. A question was entered in English and Urdu, both done independently using the "New Chat" function.

Grading

Grading was done independently by two physicians who were fluent in both languages. One physician was an internal medicine specialist and the other an endocrinologist. Both of them were practicing doctors with fellowship degrees in medicine and had over a decade's worth of experience. They were asked to grade responses considering two aspects. They were first asked to assess the comprehensiveness of Urdu responses solitarily following a scale used in previous studies[11] (Figure 2).

The reviewers were then asked to grade the Urdu response accuracy with its English counterpart using another scale as formed by Samaan et al.[14] (Figure 3).

For questions with vastly varying responses (e.g. response graded as relatively comprehensive by one reviewer and incomprehensive or inaccurate by the second reviewer), reviewers were asked to reevaluate them separately. Further disparity was resolved by involving a third board-certified senior internal medicine specialist with over 60 years of experience in diabetes-related management and treatment within Pakistan.

The grading for comprehension of Urdu responses was stratified into two groups: grades 1 and 2 versus grades 3

and 4. If answers from both reviewers were classed as being in two separate categories, they were determined to be significantly different from each other, requiring a third reviewer to resolve.

Evaluating ChatGPT's response to questions on diabetes-related questions on basic knowledge, lifestyle, and preventative management

The questions included were stratified into three domains of inquiry: basic knowledge, lifestyle, and preventative. We assessed ChatGPT's ability to respond to questions concerning all three stratifications separately. Proportions of each grade, comprehension of Urdu response, and comparison of responses were calculated for each domain and were divulged as percentages. The analyses were calculated using the latest available online version of Google Sheets as of April 2024.

Results

The research data were quantitatively presented, and a total of 53 questions were included in the study. Of the Urdu responses generated, in terms of comprehensiveness, only two of 53 (3.8%) responses were graded as comprehensive and five of 53 responses (9.4%) were graded as correct but inadequate. Overall, only seven of 53 (13.2%) answers were graded as comprehensive or correct but inadequate. Twenty-five of 53 (47.2%) questions were graded as mixed with correct and incorrect/outdated data, the most significant proportion of responses being graded as such as seen in Table 1 (Supplementary Table 1). The Urdu responses for ChatGPT pertaining to diabetes have been translated to English and presented in Supplementary Table 2.

We found that 23 of 53 (43.4%) questions were graded differently, i.e. comprehensive or correct but inadequate versus mixed with correct and incorrect/outdated data or completely incorrect by two reviewers and resolved by the third reviewer.

When considering the comparison of response scale grading the comparative accuracy of Urdu and English responses, no Urdu responses (0.0%) were considered to have more accuracy than English. Most of the Urdu responses were found to have an accuracy less than that of English, an overwhelming majority of 49 of 53 (92.5%) responses. We found that ChatGPT was significantly less able to give responses as extensive as English. This was substantiated by how the mean word count in responses for English was 378 and 223 in Urdu.

We stratified the questions into three domains, i.e. basic knowledge, lifestyle, and preventative medicine. In these three domains, we found the lifestyle section to have the most comprehensive responses, with two of 10 (20.0%)
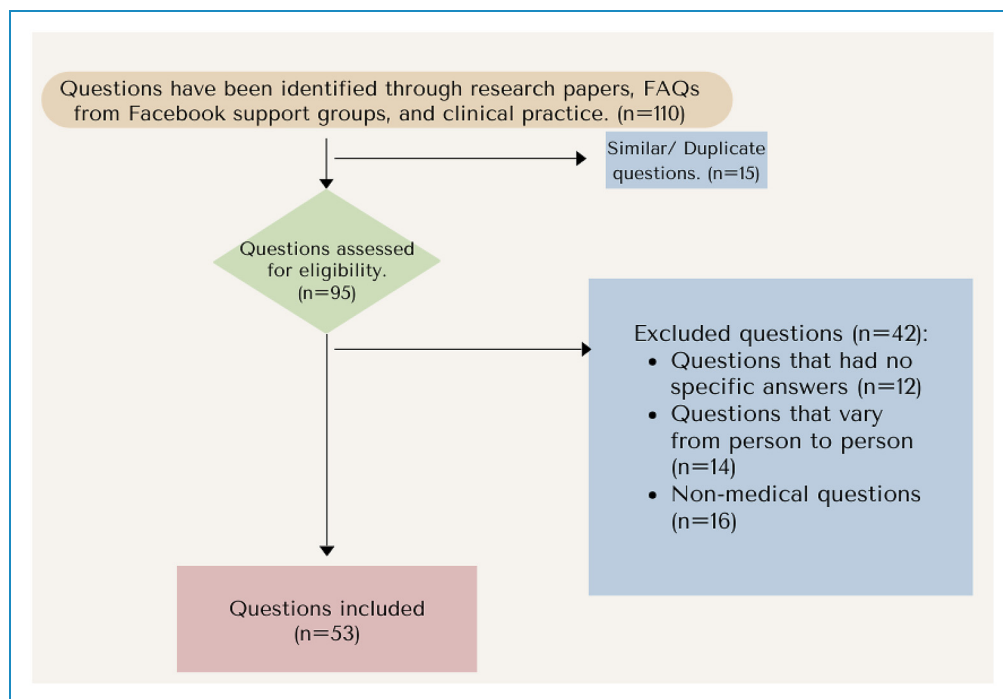
Figure 1. Flowchart of question selection for diabetes-related questions. Frequently asked questions about the knowledge, lifestyle, and preventative management of diabetes were collected from questionnaires, Facebook support groups, and clinical practice.

Figure 2. Comprehensiveness of Urdu Responses Grading Scale.

Figure 3. Comparative accuracy of Urdu and English Responses Grade.

questions graded as comprehensive or correct but inadequate. We found that three of 16 (18.8%) questions of the preventative domain were graded similarly. Basic knowledge was found to be the least comprehensive, with only two of 27 (7.4%) answers to be graded as such (Table 1).

Basic knowledge was found to be the most incomprehensive in Urdu responses, and 25 of 27 (92.6%) answers were graded as mixed with correct and incorrect/outdated data or completely inaccurate. Thirteen of 16 (81.3%) questions in the preventative section were graded similarly. The lifestyle domain was found to be the least incomprehensive, four out of 10 (80.0%) of which were graded as stated prior.

Discussion

ChatGPT's responses are confined to the data that it has been trained on, the knowledge it has based on the substantial volume of text data that has been encoded into it. A study done evaluating and comparing ChatGPT's responses in Arabic and English postulates that the sources utilized by the program are not known. It is not apparent, however, whether information reiterated in Arabic is cited from Arabic sources or directly translated from English.[14] We pose similar questions, unaware of what sources exactly have been used to generate Urdu responses. These sources are instrumental to ChatGPT's performance. There is an abundance of medical literature present in English about diabetes and its basic knowledge, lifestyle, and preventative management. This may have contributed to the stark difference in the comprehensiveness of answers between English and Urdu. For questions specific to the lifestyle management of

Table 1. Comparison of comprehension and accuracy of responses between English and Urdu overall and in three domains.

	Basic knowledge	Lifestyle	Preventative	
Total	27	10	16	
Accuracy				Total
Comprehensive	0 (0%)	1 (10.0%)	1 (6.3%)	2 (3.8%)
Correct but inadequate	2 (7.4%)	1 (10.0%)	2 (12.5%)	5 (9.4%)
Mixed correct and incorrect	15 (55.6%)	4 (40.0%)	6 (37.5%)	25 (47.2%)
Completely incorrect	10 (37.0%)	4 (40.0%)	7 (43.8%)	21 (39.6%)
Comparison of response				Total
Urdu response is more accurate	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Similar accuracy	2 (7.4%)	1 (10.0%)	1 (6.3%)	4 (7.6%)
Urdu response is less accurate	25 (92.6%)	9 (90.0%)	15 (93.8%)	49 (92.5%)

ChatGPT, we found that its answers in Urdu were more relevant to the question asked than in the other subtypes. We ascertain this to be true as much of the information regurgitated for lifestyle questions was analogous responses in basic knowledge and preventative questions. We feel that the relevance of its responses was better suited to lifestyle, hence improving the grading in terms of comprehension.

Our study highlights the compelling requirement for developing non-English evaluation methods for LLMs. English is undoubtedly the most broadly spoken language in the world.[17] The widened schism in proficiency and comprehension of Urdu and English responses as ascertained by us illustrates how ChatGPT is presently unable to expound upon and cater to linguistic heterogeneity across multiple ethnicities and cultures. ChatGPT should be further programed and developed in a fashion that overcomes these existing barriers. Although it remains an impressive feat to articulate in Urdu despite multiple factors, as further work is done specific to Urdu, there should be special attention given to understanding and navigating societal and cultural constructs in the context of medical practice.

The AI model itself refers to its reliance on the standard of data available directly influencing the quality of its responses.[18] This further elucidates why responses in Urdu were found to be lacking both in comprehensiveness and accuracy. There is little medical literature present in the form of research and books completely in Urdu. Specific Urdu medical terminologies for technical English terms have not been aptly formulated, and ChatGPT thence communicates in English instead. The dearth of Urdu medical literature, it seems, enforces ChatGPT to resort to directly translating those English terminologies to Urdu. The resulting sentence ends up being too complicated, even though there are far superior alternatives present in the Urdu language to impart the same information. We found many of the Urdu sentences to need to be corrected in syntax. The innumerable mistakes in word structures and grammar have made interpreting ChatGPT's response particularly difficult, even more so for a regular patient with no prior knowledge of diabetes. When asked whether diabetes is painful, in Urdu, ChatGPT made basic spelling mistakes, changing one letter to the other and completely altering the meaning of the word it was trying to use. The significant difference is quantitatively highlighted by the lower mean word count of Urdu responses. This indicated a distinct lack of terminology specific to diabetes in Urdu present in ChatGPT's arsenal. For example, when asked which demographic develops diabetes, in comparison to English, although ChatGPT posed similar headings, it was unable to elaborate as exhaustively in Urdu as in English. Furthermore, the explanation it gave for gestational diabetes was completely inaccurate and did not touch upon other widely regarded conditions associated with diabetes, for example, polycystic ovarian syndrome.

ChatGPT's inadequate performance in Urdu necessitates the requirement to develop evidence-based literature specific to diabetes in Pakistan that is denoted completely in the Urdu language.[18] Although there is plentiful information available on diabetes disease management, many Pakistanis will find it difficult to access and comprehend due to existing language barriers. None of ChatGPT's responses in Urdu were more accurate than in English, and barely any responses heralded a similar accuracy. The results of our study pose a worrying advent regarding

how scant medical literature is available in Urdu. The importance of having medical literature in Urdu is evidenced by 7% of the population speaking purely Urdu and others considering it a language that is widely spoken and understood.[19] Henceforth, we find that with the development of different conduits of medical information, healthcare providers in South Asia, and especially in Pakistan, should work to publish the most accurate diabetes responses to commonly asked diabetes-related questions in Urdu online. This is required to allow ChatGPT to retrieve relevant data on diabetes-related questions in Urdu effectively. We hope that our study galvanizes other healthcare providers to take steps to develop Urdu literature on diabetes and publish it online and in local journals.

Furthermore, Pakistan boasts a low doctor-to-patient ratio, which, according to the World Health Organization stands at 1:1300, subsequently leading to significant patient burden on existing healthcare frameworks.[20] Should ChatGPT development be catered to developing in Urdu, it could revolutionize healthcare and relieve the aforementioned burden. As it stands presently, however, ChatGPT's ability to personalize responses, allowing for almost human-like interaction with its user,[21] is not apparent in Urdu. The potential for it to do so and provide critically clear and concise communication of medical information remains boundless.[22]

Educating diabetic patients regarding their disease allows them to properly understand their condition, manage their symptoms, and improve their quality of life. Patient-centered education has long since played a role that is intrinsic to a patient's ability to live with the disease.[18] For a country where patients are often hesitant to ask for help, having an AI model like ChatGPT widely accessible and easy to use can be paramount to patient support and management. Although many patients understand how necessary dietary control and patient management are, adherence to a healthier lifestyle remained lacking, particularly due to patients being unclear on the information provided to them.[22] We found that ChatGPT could not be used as a source of information in perpetuity particularly due to rapid updates in the medical field.

Patients, with chronic medical illnesses like diabetes, may be unable to retain the hordes of information provided to them, hence expounding on how LLMs like ChatGPT can act as conduits of information. Having a constant source of information can alter the way we manage such patients. ChatGPT can potentially act as a repository from the comfort and convenience of their homes. This can be even more beneficial in impoverished areas lacking proper infrastructure in Pakistan.

In terms of study limitations, 43.4% of responses were marked differently between the two reviewers, indicating that the scale used seemed open to interpretation and prone to being perceived differently by independent reviewers. ChatGPT's ability to respond in Urdu has several limitations and requires to be utilized with much care as even an adjunct source of information. Further research could be done by blinding the fact that ChatGPT is the source of responses, allowing for stricter assessment, hence preventing the under- or overestimation of grading.

Conclusion

We have found that although ChatGPT's ability to retrieve such information is impressive, within scientific and medical frameworks particular to diabetes, it can merely be used as an adjunct instead of a solitary source of information. Having had responses graded by consultants with years of experience in the medical field, answers from ChatGPT about diabetes in Urdu were considered subpar lacking reliability for information that was more technical in a medical context. We ascertained that instead of improving patient comprehension of their condition, it could give rise to further confusion through these subpar responses. The responses were found to be difficult to understand and embroiled in grammatical and spelling errors. Pakistani healthcare providers must work to publish gold standard medical and diabetes information in Urdu online. This would enable ChatGPT and other LLMs to source clinically accurate and easily understandable information in Urdu. These systems have the potential to revolutionize healthcare and alleviate the medical health burden specific to diabetes, especially in communities with limited access to reliable information.

References

1. World Health Organization. Diabetes. WHO Regional Office for Europe, https://www.who.int/europe/news-room/fact-sheets/item/diabetes (2023, accessed 17 August 2024).
2. World Health Organization. Diabetes. https://www.who.int/news-room/fact-sheets/detail/diabetes (2023, accessed 17 August 2024).
3. Azeem S, Khan U and Liaquat A. The increasing rate of diabetes in Pakistan: A silent killer. *Ann Med Surg (Lond)* 2022; 79: 103901.
4. Pardhan S, Nakafero G, Raman R, et al. Barriers to diabetes awareness and self-help are influenced by people's demographics: Perspectives of South Asians with type 2 diabetes. *Ethn Health* 2020; 25: 843–861.
5. Matthews NR, Porter GJ, Varghese M, et al. Health and socioeconomic resource provision for older people in South Asian countries: Bangladesh, India, Nepal, Pakistan and Sri Lanka evidence from NEESAMA. *Glob Health Action* 2023; 16: 2110198.
6. Sękowski K, Grudziąż-Sękowska J, Pinkas J, et al. Public knowledge and awareness of diabetes mellitus, its risk factors, complications, and prevention methods among adults in Poland: A 2022 nationwide cross-sectional survey. *Front Public Health* 2022; 10: 1029358.
7. Waheedi M, Awad A, Hatoum HT, et al. The relationship between patients' knowledge of diabetes therapeutic goals and self-management behavior, including adherence. *Int J Clin Pharm* 2017; 39: 45–51.
8. Chavan GM, Waghachavare VB, Gore AD, et al. Knowledge about diabetes and relationship between compliance to the management among the diabetic patients from rural area of Sangli District, Maharashtra, India. *J Family Med Prim Care* 2015; 4: 439–443.
9. Biswas S, Davies LN, Sheppard AL, et al. Utility of artificial intelligence-based large language models in ophthalmic care. *Ophthalmic Physiol Opt* 2024; 44: 641–671.
10. Liu J, Wang C and Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res* 2023; 25.
11. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol* 2023; 29: 721–732.
12. Dave T, Athaluri SA and Singh S. ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell* 2023; 6: 1169595.
13. Huang C, Chen L, Huang H, et al. Evaluate the accuracy of ChatGPT's responses to diabetes questions and misconceptions. *J Transl Med* 2023; 21: 502.
14. Samaan JS, Yeo YH, Ng WH, et al. ChatGPT's ability to comprehend and answer cirrhosis related questions in Arabic. *Arab J Gastroenterol* 2023; 24: 145–148.
15. OpenAI. ChatGPT: Optimizing language models for dialogue. OpenAI website, https://openai.com/blog/chatgpt (2022, accessed 9 January 2024).
16. ChatGPT. https://chat.openai.com/ (2022, accessed 9 January 2024).
17. Temsah O, Khan SA, Chaiah Y, et al. Overview of early ChatGPT's presence in medical literature: Insights from a hybrid literature review by ChatGPT and human experts. *Cureus* 2023; 15.
18. Mahmood S, Khan MO and Hameed I. With great power comes great responsibility—Deductions of using ChatGPT for medical research. *EXCLI J* 2023; 22: 499–501.
19. World Health Organization. Health system profile of Pakistan: regional health system observatory. Cairo, Egypt: WHO Regional Health System Observatory-EMRO, http://apps.who.int/medicinedocs/documents/s17305e/s17305e.pdf (2006, accessed 17 August 2024).
20. Muhammad Q, Ahmed S, Khan Z, et al. Healthcare in Pakistan: Navigating challenges and building a brighter future. *Cureus* 2023; 15.
21. Torab-Miandoab A, Samad-Soltani T, Jodati A, et al. Interoperability of heterogeneous health information systems: A systematic literature review. *BMC Med Inform Decis Mak* 2023; 23: 18.
22. Ranasinghe P, Pigera AS, Ishara MH, et al. Knowledge and perceptions about diet and physical activity among Sri Lankan adults with diabetes mellitus: A qualitative study. *BMC Public Health* 2015; 15: 1160.