



Evaluating the accuracy and adequacy of ChatGPT in responding to queries of diabetes patients in primary healthcare

İrem Şenoymak¹ · Nuriye Hale Erbatur² · Mustafa Can Şenoymak² · Memet Taşkın Egici³

Received: 15 March 2024 / Accepted: 29 August 2024

© The Author(s), under exclusive licence to Research Society for Study of Diabetes in India 2024

Abstract

Objective This study evaluates the accuracy and adequacy of Chat Generative Pre-trained Transformer (ChatGPT) in responding to common queries formulated by primary care physicians based on their interactions with diabetic patients in primary healthcare settings.

Methods Thirty-two frequently asked questions were identified by experienced primary care physicians and presented systematically to ChatGPT. Responses underwent evaluation by two endocrinology and metabolism physicians which utilized a 3-point Likert scale for accuracy (1, inaccurate; 2, partially accurate; 3, accurate) and a 6-point Likert scale for adequacy (1, completely inadequate to 6, completely adequate). Questions were categorized into groups including general information, diagnostic processes, treatment procedures, and complications.

Results The median accuracy score was 3.0 (IQR, 3.0–3.0), and the adequacy score was 4.5 (IQR, 4.0–5.8). None of the questions received an inaccurate rating, and the lowest accuracy score assigned by both evaluators was 3. Significant agreement was observed between the evaluators, demonstrated by a weighted κ of 0.61 ($p < .0001$) for accuracy and substantial agreement with a weighted κ of 0.62 ($p < 0.0001$) for adequacy. The Kruskal–Wallis tests revealed no statistically significant differences among the groups of $p = .71$) and adequacy ($p = .57$).

Conclusions ChatGPT demonstrated commendable accuracy and adequacy in addressing diabetes-related queries in primary healthcare.

Keywords Diabetes mellitus · Artificial intelligence · ChatGPT · Primary healthcare

Introduction

Diabetes mellitus, a chronic disease, affects approximately 10.5% of the global population, with its prevalence steadily increasing [1]. Disease management advocates a personalized treatment paradigm involving active collaboration between patients and healthcare professionals, delineating specific responsibilities for each [2, 3]. The awareness and knowledge depth of patients regarding this complex condition exhibit a direct and significant correlation with the

effectiveness of disease management [4]. Thus, acquiring comprehensive knowledge about diabetes emerges as an inevitable prerequisite for achieving successful therapeutic outcomes.

Primary healthcare institutions serve as the initial point of contact for many patients with diabetes, playing a pivotal role in their healthcare journey [5]. Understanding the concerns and questions frequently raised by these individuals is indispensable for enhancing the quality of care provided.

In the contemporary landscape, artificial intelligence applications have evolved into widely used and easily accessible sources of information [6]. Patients, on occasion, utilize these applications to inquire about their illnesses and health conditions [7, 8]. It has been observed that patients benefit from such applications in several ways by gaining insights into their conditions, managing treatment through reminders, dosage instructions, and information about side effects, and also by asking questions in a conversational manner [7, 8]. The aim of our study is to scrutinize the

✉ Mustafa Can Şenoymak
senoymak@gmail.com

¹ Family Medicine, Üsküdar State Hospital, Istanbul, Turkey

² Endocrinology and Metabolism, University of Health Sciences Sultan Abdulhamid Han Training and Research Hospital, Istanbul, Turkey

³ Family Medicine, University of Health Sciences Haydarpaşa Numune Training and Research Hospital, Istanbul, Turkey

accuracy and adequacy of responses provided by the artificial intelligence program Chat Generative Pre-trained Transformer (ChatGPT) to the most common questions posed by patients with diabetes attending primary healthcare institutions, thereby contributing to establishing a bridge between primary healthcare and artificial intelligence (AI).

Materials and methods

Study design

Experienced primary care physicians who had worked in primary healthcare institutions were involved in the study. These physicians were tasked with identifying the most frequently asked questions by diabetic patients seeking assistance at primary healthcare facilities.

Following the establishment of the list, the identified 32 questions were formulated and systematically presented to ChatGPT by primary care physicians in English, an artificial intelligence program (<https://chat.openai.com>) (GPT3.5, Nov 12, 2023 version). ChatGPT version 3.5 was chosen for its accessibility to all users recognizing that the latest versions of such technologies may not be universally accessible. Each question was presented to ChatGPT twice to assess the reproducibility of responses for the same question, and the recorded responses were recorded for further analysis. The temperature of the model for sampling responses was set to a default value of 0.7 for ChatGPT 3.5, following common practices in natural language processing tasks.

The responses recorded from ChatGPT were subjected to assessment by two independent endocrinology and metabolism physicians. The evaluation process employed a scoring system, considering both accuracy and adequacy as primary criteria. In this study, accuracy is defined as the extent to which ChatGPT's responses align with accurate medical information and guidelines recognized in the field of diabetes management. Adequacy, on the other hand, refers to the extent to which ChatGPT's responses meet the informational needs of diabetes patients in a primary healthcare setting, providing sufficiently detailed and understandable information to support their understanding and management of their condition. The responses were graded according to the international guideline of the American Diabetes Association Professional Practice Committee, Standards of Care in Diabetes 2024. The scoring scale encompassed a range of values, the accuracy scale was a 3-point Likert scale (with 1 indicating inaccurate; 2 partially accurate; and 3 accurate). Additionally, the adequacy scale was a 6-point Likert scale (with 1 indicating completely inadequate; 2 more inadequate than adequate; 3 approximately equal adequate and inadequate; 4 more adequate than inadequate; 5 nearly

all adequate; and 6 completely adequate). Furthermore, the questions were categorized into four distinct groups: general information, diagnostic process, treatment process, and complications of diabetes mellitus. An examination was undertaken to assess whether there were discernible differences in both accuracy and adequacy exhibited by ChatGPT across these delineated question categories. This systematic scoring approach was employed to derive an evaluation of ChatGPT's efficacy in responding to the identified queries posed by patients with diabetes in a primary healthcare context.

Statistical analyses

IBM SPSS Statistics version 25.0 software was used for data analyses. Outcome scores were presented in a descriptive manner, including median [interquartile range (IQR)] values and mean [standard deviation (SD)] values, and were subjected to group-wise comparisons using either the Mann–Whitney *U* test or the Kruskal–Wallis test (SPSS, version 25). Inter-rater concordance was assessed employing the weighted κ statistic across the entire spectrum of scores, ranging from 1 to 3 for accuracy and 1 to 6 for adequacy. Responses to the repeated queries were subjected to comparison using the Wilcoxon signed rank test in order to assess reproducibility. A significance threshold of $p < 0.05$ was deemed indicative of statistical significance.

Results

In this study, artificial intelligence addressed frequently asked questions about diabetes mellitus in routine primary care practice. Subsequently, two physicians working in endocrinology and metabolism assessed and scored the provided answers (Table 1). Sample ChatGPT responses to the questions are presented in Figs. 1 and 2.

For all 32 questions examined, the median average accuracy score was 3.0 (interquartile range, 3.0–3.0), suggesting a consistently accurate performance. The average mean (standard deviation) score of 2.8 (0.3) fell within the range between accuracy and partial accuracy (Table 2).

Regarding adequacy, the average median score was 4.5 (interquartile range, 4.0–5.8), indicating a level between more adequate than inadequate and nearly all adequate. The mean (standard deviation) score of 4.6 (1.0) further supports this assessment (Table 3).

Reproducibility test demonstrated that the responses to the original and repeated questions did not differ significantly in terms of both accuracy and adequacy (with respective *p* values of 0.2 and 0.43 determined by the Wilcoxon signed-rank test). The responses to repeated questions garnered a median accuracy score of 3 (IQR, 2.5–3.0; mean [SD] score, 2.9 [0.1]) and a median adequacy score

Table 1 Questions and average accuracy and adequacy scores given by doctors to the AI's answers^a

Questions	Accuracy score ^b	Adequacy score ^c
What is diabetes mellitus?	3	5
How is diabetes mellitus diagnosed?	3	4.5
How many types of diabetes mellitus are there?	2.5	4
How is the differentiation made between Type 1 and Type 2 diabetes mellitus?	3	5.5
What causes diabetes mellitus?	3	5.5
Is diabetes mellitus genetic, and can it be passed on to my children?	3	6
Can diabetes mellitus be prevented?	3	6
What are the symptoms of diabetes mellitus?	3	6
Is diabetes mellitus permanent?	3	6
Can diabetes mellitus be treated?	3	5
What is the treatment for diabetes mellitus?	2	3
Should every diabetic patient use insulin?	2.5	3
Is insulin usage mandatory for Type 1 diabetes?	3	3
Can insulin be discontinued for Type 1 diabetes?	3	4
What is a diabetic diet?	3	4.5
Should exercises be done for diabetes mellitus, and if so, how should exercise be implemented?	3	4
What are the medications for diabetes mellitus?	3	5.5
What are the side effects of diabetes medications?	3	6
Can diabetes medications be stopped later on?	2	4
What should be the fasting blood sugar level for a diabetic patient?	3	4
What should be the postprandial blood sugar level for a diabetic patient?	3	4
What is HbA1c?	3	6
What should be the HbA1c level for a diabetic patient?	3	4
Which organs does diabetes mellitus affect?	3	5
What are the complications of diabetes mellitus?	3	5.5
What can be done to prevent complications of diabetes mellitus?	3	6
Is there a diabetes surgery?	2	3
How often should a diabetic patient visit the doctor?	3	4
Should blood sugar measurements be done at home, and if yes, how often?	3	3.5
Can a diabetic patient fast?	3	3
Does diabetes mellitus hinder childbirth?	3	6
In which situations should a diabetic patient seek emergency care?	2.5	3.5

^aAI indicates artificial intelligence


^bThe accuracy scale was a 3-point Likert scale (with 1 indicating inaccurate; 2 partially accurate; and 3 accurate)


^cThe adequacy scale was a 6-point Likert scale (with 1 indicating completely inadequate; 2 more inadequate than adequate; 3 approximately equal adequate and inadequate; 4 more adequate than inadequate; 5 nearly all adequate; and 6 completely adequate)

of 4.5 (IQR, 3.5–5.8; mean [SD] score, 4.7 [1.0]). While there were differences in how sentences were structured and some minor changes, there were no significant alterations in content found among the responses. The scores given by the evaluators in the second assessment varied by a maximum of 1 point. Notably, the only question where the adequacy median score changed by up to 1 point was question 22 (“What is HbA1c?”). In the second evaluation, ChatGPT provided more detailed information compared to

the initial evaluation, stating that the normal HbA1c level in the general population should be below 5.7, explained that HbA1c reflects the average blood glucose level over the past 3 months due to the 120-day lifespan of erythrocytes, and offered more detailed information about its formation through glycation. These additional details contributed to a higher adequacy rating in the second assessment. Similarly, in terms of accuracy, the only question where the median score changed by up to 1 point was question 29

Fig. 1 ChatGPT response to the question “How is diabetes mellitus diagnosed?”

 **Anonymous**
2) How is diabetes mellitus diagnosed?


 **ChatGPT**
Diagnosing diabetes mellitus typically involves several tests and evaluations to assess blood sugar levels and the body's response to glucose. The primary diagnostic tests for diabetes include:


1. **Fasting Plasma Glucose (FPG) Test:** This test measures blood sugar after an overnight fast (usually at least 8 hours). A fasting blood sugar level of 126 milligrams per deciliter (mg/dL) or higher on two separate tests indicates diabetes.
2. **Oral Glucose Tolerance Test (OGTT):** This test involves fasting overnight and then drinking a sugary solution. Blood sugar levels are checked periodically over the next few hours. A blood sugar level of 200 mg/dL or higher after two hours indicates diabetes.
3. **Random Plasma Glucose Test:** This test measures blood sugar levels at any time of the day, regardless of when you last ate. If your blood sugar level is 200 mg/dL or higher and you're experiencing symptoms of diabetes (like increased thirst or urination), it may indicate diabetes.
4. **Hemoglobin A1c Test (HbA1c):** This test measures the average blood sugar level over the past 2-3 months. An HbA1c level of 6.5% or higher usually indicates diabetes.

Healthcare professionals may use one or a combination of these tests to confirm a diabetes diagnosis. Sometimes, additional tests or repeated tests may be necessary to ensure accuracy and determine the type of diabetes (Type 1, Type 2, gestational, etc.).

If you suspect you might have diabetes due to symptoms such as increased thirst, frequent urination, unexplained weight loss, or fatigue, it's essential to consult a healthcare provider. They can perform the necessary tests, evaluate your symptoms and medical history, and guide you through proper diagnosis and management. Early detection and management of diabetes are crucial in preventing complications.

Fig. 2 ChatGPT response to the question “Is diabetes mellitus genetic, and can it be passed on to my children?”

 **Anonymous**
6) Is diabetes mellitus genetic, and can it be passed on to my children?

 **ChatGPT**
Yes, there's a genetic component to diabetes mellitus, particularly in Type 1 and Type 2 diabetes.

1. **Type 1 Diabetes:** Genetics play a significant role in the development of Type 1 diabetes. Having family members with Type 1 diabetes slightly increases the risk of developing the condition. However, the inheritance pattern is complex and not solely determined by genetics. Environmental factors also play a role in triggering the autoimmune response that leads to Type 1 diabetes.
2. **Type 2 Diabetes:** Family history and genetics significantly influence the risk of developing Type 2 diabetes. If a parent or sibling has Type 2 diabetes, it can increase an individual's risk. However, lifestyle factors, such as unhealthy eating habits, lack of physical activity, and obesity, also play a crucial role in its development.

While genetics can predispose someone to diabetes, it's not a certainty that children will develop diabetes if their parents have it. The risk is increased but not guaranteed. Lifestyle factors and environmental influences also contribute significantly to the development of diabetes.

It's important to note that gestational diabetes in a mother can also increase the risk of the child developing Type 2 diabetes later in life.

Maintaining a healthy lifestyle, including a balanced diet, regular exercise, and weight management, is essential in reducing the risk of developing diabetes, especially for individuals with a family history of the condition. Regular check-ups and discussions with healthcare professionals can help monitor and manage any increased risk factors for diabetes.

Table 2 Scoring accuracy^a and comparing responses generated by artificial intelligence based on categorization into question groups

	General information	Diagnostic process	Treatment process	Complications of DM	Total	<i>p</i> value
Rater 1						
Median (IQR)	3.0 (3.0–3.0)	3.0 (3.0–3.0)	3.0 (3.0–3.0)	3.0 (2.5–3.0)	3.0 (3.0–3.0)	0.78
Mean (SD)	2.9 (0.2)	3.0 (0.0)	2.9 (0.3)	2.8 (0.4)	2.9 (0.2)	
Rater 2						
Median (IQR)	3.0 (3.0–3.0)	3.0 (3.0–3.0)	3.0 (2.0–3.0)	3.0 (2.5–3.0)	3.0 (3.0–3.0)	0.69
Mean (SD)	2.8 (0.3)	3.0 (0.0)	2.7 (0.4)	2.8 (0.4)	2.8 (0.3)	
Average score						
Median (IQR)	3.0 (3.0–3.0)	3.0 (3.0–3.0)	3.0 (2.5–3.0)	3.0 (2.5–3.0)	3.0 (3.0–3.0)	0.71
Mean (SD)	2.8 (0.3)	3.0 (0.0)	2.8 (0.3)	2.8 (0.4)	2.8 (0.3)	

The Kruskal–Wallis (> 2 variables) tests were used for non-parametric variables. *p* values less than 0.05 were considered statistically significant.

^aThe accuracy scale was a 3-point Likert scale (with 1 indicating inaccurate; 2 partially accurate; and 3 accurate).

Table 3 Scoring adequacy^a and comparing responses generated by artificial intelligence based on categorization into question groups

	General information	Diagnostic process	Treatment process	Complications of DM	Total	<i>p</i> value
Rater 1						
Median (IQR)	5.5 (4.0–6.0)	4.0 (3.25–5.5)	4.0 (3.0–6.0)	4.0 (3.5–5.5)	4.5 (3.2–6.0)	0.62
Mean (SD)	5.0 (1.2)	4.2 (1.2)	4.4 (1.3)	4.4 (1.1)	4.6 (1.2)	
Rater 2						
Median (IQR)	5.0 (4.2–6.0)	4.5 (4.0–5.7)	4.0 (4.0–6.0)	4.0 (3.5–4.5)	5.0 (4.0–5.7)	0.37
Mean (SD)	4.9 (1.0)	4.7 (0.9)	4.5 (1.1)	4.0 (0.7)	4.6 (1.0)	
Average score						
Median (IQR)	5.2 (4.1–6.0)	4.0 (4.0–5.5)	4.0 (3.5–6.0)	4.0 (3.5–5.0)	4.5 (4.0–5.8)	0.57
Mean (SD)	4.9 (1.1)	4.5 (1.0)	4.5 (1.2)	4.2 (0.9)	4.6 (1.0)	

The Kruskal–Wallis (> 2 variables) tests were used for non-parametric variables, and data were given median (minimum–maximum). *p* values less than 0.05 were considered statistically significant.

^aThe adequacy scale was a 6-point Likert scale (with 1 indicating completely inadequate; 2 more inadequate than adequate; 3 approximately equal adequate and inadequate; 4 more adequate than inadequate; 5 nearly all adequate; and 6 completely adequate).

(“Should blood sugar measurements be done at home, and if yes, how often?”). In the second evaluation, ChatGPT emphasized the importance of blood sugar measurement during changes in medical treatment for type 2 diabetes, which was not mentioned in the initial evaluation. This additional information led to a higher accuracy rating in the second assessment.

Significant agreement was observed between the evaluators, demonstrated by a weighted κ of 0.61 ($p < 0.0001$) for accuracy and substantial agreement with a weighted κ of 0.62 ($p < 0.0001$) for adequacy.

Evaluator #1 assigned the highest accuracy score (3.0) to 29 questions (90.6%), while 3 questions (9.4%) received a partially accurate score (2.0). In contrast, evaluator #2 awarded the highest accuracy score (3.0) to 26 questions (81.3%), with 6 questions (18.8%) deemed partially accurate (2.0). It is noteworthy that both evaluators assigned the lowest accuracy score of 2.0, and none of the questions were rated as inaccurate.

Regarding adequacy, evaluator #1 scored the answers to 12 questions (37.5%) as completely adequate, 8 (26.1%) as adequate, and 8 (25%) as approximately equal adequate and inadequate. On the other hand, evaluator #2 scored the answers to 8 questions (25%) as completely adequate, 9 (28.1%) as nearly all adequate, 10 (31.3%) as more adequate than inadequate, and 5 (15.6%) as approximately equal adequate and inadequate.

The lowest adequacy score assigned by both evaluators was 3. Notably, none of the questions were scored as inadequate or more inadequate than adequate. A modest correlation between accuracy and adequacy was observed, as evidenced by a Spearman correlation coefficient (r) of 0.53 ($p < 0.01$; $\alpha = 0.002$) for all questions, indicating a positive but not strong relationship between the two variables.

Questions were categorized into groups such as general information, diagnostic process, treatment process, and complications of diabetes mellitus. The average median accuracy scores for these groups were 3.0 (interquartile range,

2.6–3.0), 3.0 (IQR, 2.2–3.0), 3.0 (IQR, 3.0–3.0), and 3.0 (IQR, 3.0–3.0), respectively. Correspondingly, the average mean [SD] scores were 2.8 [0.3], 2.75 [0.5], 2.8 [0.3], and 3.0 [0.0], respectively. The Kruskal–Wallis test ($p=0.71$) revealed no statistically significant differences among the groups, indicating comparable accuracy across various question types (Table 2).

Similarly, for questions grouped into general information, diagnostic process, treatment process, and complications of diabetes mellitus, the average median adequacy scores were 5.25 (IQR, 4.1–6.0), 4.0 (IQR, 4.0–5.4), 4.0 (IQR, 3.5–6.0), and 4.0 (IQR, 3.5–5.0), respectively. The average mean [SD] scores were 4.9 [1.1], 4.5 [1.0], 4.5 [1.2], and 4.2 [0.9], respectively. The Kruskal–Wallis test ($p=0.57$) indicated no statistically significant differences among the groups, signifying consistent adequacy across various question types (Table 3).

In summary, the results suggest that there are no major differences in the accuracy and adequacy of artificial intelligence-generated answers among different question-type groups.

Discussion

This study is of significance as it investigates the adequacy of artificial intelligence programs in meeting the needs of patients seeking primary healthcare services. It also contributed to the comprehensive understanding of prevalent concerns among diabetic patients in primary care settings. Furthermore, the study stands as one of the limited examinations assessing the accuracy and efficacy of artificial intelligence programs in the diagnosis and management of diabetes.

In this study, it was determined that the artificial intelligence program ChatGPT generally provided accurate and adequate responses to questions from diabetic patients attending primary healthcare institutions. Numerous investigations in the literature have addressed this particular subject. In a study conducted by Sagstad et al. in Norway, ChatGPT was found to provide answers to 88.51% of questions related to gestational diabetes mellitus; however, the accuracy and adequacy of the responses were not evaluated [9]. In our study, ChatGPT demonstrated the capability to respond to all posed questions. Similar to our study, Hernandez and colleagues found in their investigation that ChatGPT provided appropriate responses to 98.5% of the 70 questions related to diabetes [10]. However, the questions were not categorized, and additionally, they were not evaluated for adequacy. In another study where five expert physicians examined 12 questions answered by ChatGPT, it was once again determined that the questions were responded to with

high accuracy [11]. In the investigation by Mondal et al., which explored lifestyle-related diseases, including diabetes among 20 cases, questions were posed to ChatGPT. The program was observed to analyze the cases with high accuracy and proficiency. Consequently, it has been suggested that patients could utilize ChatGPT in situations where they cannot access medical professionals [12].

In the study conducted by Meo and colleagues, responses provided by ChatGPT to multiple-choice questions derived from textbooks regarding diabetes were examined. The results indicated that ChatGPT outperformed another artificial intelligence program, Google Bard, but still accurately answered only 23 out of 50 multiple-choice questions [13]. Therefore, it was reported that, at this stage, it is not suitable for the use of medical students and requires further improvement. At this point, similar to our study, it might be considered that information is more useful for patients rather than medical professionals. Additionally, in studies comparing ChatGPT responses with guideline recommendations, the artificial intelligence program did not exhibit significant success. For instance, in a study where 20 questions about the assessment and treatment of obesity in type 2 diabetes mellitus were posed, the compatibility of ChatGPT responses with the American Diabetes Association and American Association of Clinical Endocrinology guidelines was examined. Evaluation of the answers revealed good alignment with guideline recommendations in the assessment section but insufficient alignment in the treatment section. Consequently, it was emphasized that ChatGPT should not be used as a substitute for healthcare professionals [14]. In another study, the synthesis and adaptation ability of ChatGPT regarding diabetic ketoacidosis were assessed based on three different guidelines. The artificial intelligence application was reported to be not very successful in this regard, highlighting the necessity of careful interpretation and verification of content generated by artificial intelligence in the medical field [15]. It is evident that despite the promising results demonstrated by ChatGPT in providing accurate and adequate responses to diabetes-related queries, it is crucial to emphasize that the use of artificial intelligence programs by patients to self-manage their illnesses or medical conditions can lead to potentially harmful outcomes. Disease management processes should be strictly overseen and conducted by medical professionals. The guidance and expertise of healthcare providers are essential to ensure safe and effective treatment and to mitigate risks associated with self-diagnosis and self-treatment.

Diabetes self-management and education (DSME) have been consistently reported as an indispensable factor in improving patient outcomes in numerous studies [4, 16–18]. Hildebrand et al.'s meta-analysis demonstrated a significant reduction in A1C levels as a result of DSME

interventions [17]. Furthermore, a review of 44 studies revealed that DSME contributes to behavioral improvement and significant enhancement in clinical outcomes for patients. However, it was emphasized that the individualization of education methods is crucial [18]. In our study, while contemplating the successful implementation of DSME by artificial intelligence programs, we acknowledge that the constraint of uniformity and standardization in responses, regardless of individuals' educational and cultural levels, poses a limiting factor.

The main limitation of our study and a key challenge with artificial intelligence programs is the variability in responses based on question formulation. The variability of questions posed by patients due to differences in their ability to articulate queries may lead to disparate responses. This study did not systematically analyze how variations in question structure or wording might influence ChatGPT's responses. Exploring how responses vary with different formulations presents a valuable area for future research, highlighting both the limitations of our current study and potential directions for further investigation. Another point is that while the responses generated by ChatGPT are evaluated by healthcare professionals, it remains uncertain whether these responses will have the same impact on the primary target audience, namely patients. As a result, patients may not fully benefit from accurate and sufficient information which was observed in such studies. To address this, multicenter studies encompassing diverse sociocultural backgrounds, educational levels, and languages should be conducted in the future, wherein data evaluating healthcare professionals are also assessed by patients. By involving patients in the evaluation process, we can gain valuable insights into the practical effectiveness and usability of ChatGPT as a tool for patient education and support in managing diabetes mellitus.

Our study has some other limitations that warrant consideration. Despite implementing a systematic scoring approach, the inherent subjectivity in using Likert scales to assess the accuracy and adequacy of ChatGPT's responses introduces a potential source of variability. Individual perceptions and biases can affect the ratings, impacting the consistency and reliability of the results. This variability limits the generalizability of the findings, even with the significant agreement observed between evaluators. Additionally, the temporal restriction of ChatGPT's knowledge to 2022 imposes limitations on the timeliness and inclusiveness of the information it offers, potentially overlooking recent developments in diabetes research and treatment modalities. Another limitation of the study is that the use of ChatGPT version 3.5, may not represent the full capabilities of newer versions. As ChatGPT evolves, it is essential to assess how updates and newer versions impact the quality of responses. Future studies

should compare results across different versions to evaluate improvements in response quality. This will include both objective evaluations, using standardized benchmarks and automated metrics, and patient-referenced evaluations, gathering feedback from diabetic patients on clarity, relevance, and usability. This longitudinal approach will help determine whether advancements in the model enhance its accuracy, adequacy, and overall usefulness in providing diabetes-related information. Another point is that while our study utilized questions derived from common queries encountered by primary care physicians to reflect real-world scenarios, incorporating more complex and nuanced questions would offer a more thorough evaluation of the model's capabilities and limitations. Finally, the evaluation of questions by endocrinologists instead of primary care physicians may not fully capture the perspectives and priorities of primary care patients, thereby potentially overlooking aspects crucial to meeting the needs of individuals at the primary care level.

Conclusions

The study evaluated the accuracy and adequacy of artificial intelligence, exemplified by ChatGPT, in responding to queries that healthcare professionals created based on their experiences of caring for people with diabetes mellitus. The program provided a notable in addressing a spectrum of questions across categories encompassing general information, diagnostic processes, treatment procedures, and complications related to diabetes mellitus. The positive outcomes underscore its potential value in supplementing patient education and supporting disease management within primary healthcare settings, emphasizing the need for further research to address existing limitations and explore the lasting impact of AI applications in enhancing healthcare outcomes. However, it is crucial to emphasize that the use of artificial intelligence programs by patients to self-manage their illnesses or medical conditions can lead to potentially harmful outcomes given current conditions and existing scientific evidence. Disease management processes should be strictly overseen and conducted by medical professionals.

Acknowledgments The authors have no acknowledgments to declare.

Data availability Data generated and/or analyzed during the current study are available from the corresponding author on request.

Declarations

Conflict of interests The authors declare no competing interests.

Ethical clearance This observational study did not involve the use of human or animal subjects, and patient data were not utilized; hence, ethical committee approval was not required.

References

- GBD 2021 Diabetes Collaborators. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021 [published correction appears in *Lancet*. 2023 Sep 30;402(10408):1132]. *Lancet*. 2023;402(10397):203–234. [https://doi.org/10.1016/S0140-6736\(23\)01301-6](https://doi.org/10.1016/S0140-6736(23)01301-6).
- Da Rocha RB, Silva CS, Cardoso VS. Self-care in adults with type 2 diabetes mellitus: a systematic review. *CDR*. 2020;16:598–607.
- American Diabetes Association Professional Practice Committee. 5. Facilitating positive health behaviors and well-being to improve health outcomes: standards of care in diabetes-2024 [published correction appears in *Diabetes Care*. 2024 Feb 05;:]. *Diabetes Care*. 2024;47(Suppl 1):S77–S110. <https://doi.org/10.2337/dc24-S005>.
- Butayeva J, Ratan ZA, Downie S, Hosseinzadeh H. The impact of health literacy interventions on glycemic control and self-management outcomes among type 2 diabetes mellitus: a systematic review. *J Diabetes*. 2023;15(9):724–35. <https://doi.org/10.1111/1753-0407.13436>.
- Call JT, Cortés P, Harris DM. A practical review of diabetes mellitus type 2 treatment in primary care. *Rom J Intern Med*. 2022;60(1):14–23. Published 2022 Mar 17. <https://doi.org/10.2478/rjim-2021-0031>.
- Ramezani M, Takian A, Bakhtiari A, Rabiee HR, Ghazanfari S, Mostafavi H. The application of artificial intelligence in health policy: a scoping review. *BMC Health Serv Res*. 2023;23(1):1416. <https://doi.org/10.1186/s12913-023-10462-2>. (Published 2023 Dec 15).
- Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell*. 2023;6:1169595. <https://doi.org/10.3389/frai.2023.1169595>. (Published 2023 May 4).
- Elsabagh AA, Elhadary M, Elsayed B, et al. Artificial intelligence in sickle disease. *Blood Rev*. 2023;61:101102. <https://doi.org/10.1016/j.blre.2023.101102>.
- Sagstad MH, Morken NH, Lund A, Dingsør LJ, Nilsen ABV, Sorbye LM. Quantitative user data from a chatbot developed for women with gestational diabetes mellitus: observational study. *JMIR Form Res*. 2022;6(4):e28091. <https://doi.org/10.2196/28091>. (Published 2022 Apr 18).
- Hernandez CA, Vazquez Gonzalez AE, Polianovskaia A, et al. The future of patient education: AI-driven guide for type 2 diabetes. *Cureus*. 2023;15(11):48919. <https://doi.org/10.7759/cureus.48919>. (Published 2023 Nov 16).
- 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(1):502. <https://doi.org/10.1186/s12967-023-04354-6>. (Published 2023 Jul 26).
- Mondal H, Dash I, Mondal S, Behera JK. ChatGPT in answering queries related to lifestyle-related diseases and disorders. *Cureus*. 2023;15(11):e48296. <https://doi.org/10.7759/cureus.48296>. (Published 2023 Nov 5).
- Meo SA, Al-Khlaiwi T, AbuKhalaf AA, Meo AS, Klonoff DC. The scientific knowledge of bard and ChatGPT in endocrinology, diabetes, and diabetes technology: multiple-choice questions examination-based performance. *J Diabetes Sci Technol* 19322968231203987 (2023) <https://doi.org/10.1177/19322968231203987>.
- Barlas T, Altinova AE, Akturk M, Toruner FB. Credibility of ChatGPT in the assessment of obesity in type 2 diabetes according to the guidelines. *Int J Obes (Lond)*. 2024;48(2):271–5. <https://doi.org/10.1038/s41366-023-01410-5>.
- Hamed E, Eid A, Alberry M. Exploring ChatGPT's potential in facilitating adaptation of clinical guidelines: a case study of diabetic ketoacidosis guidelines. *Cureus*. 2023;15(5):e38784. <https://doi.org/10.7759/cureus.38784>. (Published 2023 May 9).
- Funnell MM, Brown TL, Childs BP, et al. National standards for diabetes self-management education. *Diabetes Care*. 2010;33 Suppl 1(Suppl 1):S89–S96. <https://doi.org/10.2337/dc10-S089>.
- Hildebrand JA, Billimek J, Lee JA, et al. Effect of diabetes self-management education on glycemic control in Latino adults with type 2 diabetes: a systematic review and meta-analysis. *Patient Educ Couns*. 2020;103(2):266–75. <https://doi.org/10.1016/j.pec.2019.09.009>.
- Camargo-Plazas P, Robertson M, Alvarado B, Paré GC, Costa IG, Duhn L. Diabetes self-management education (DSME) for older persons in Western countries: a scoping review. *PLoS One*. 2023;18(8):e0288797. <https://doi.org/10.1371/journal.pone.0288797>. (Published 2023 Aug 9).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.