Clin. Lab. 2026;72:XXX-XXX

ORIGINAL ARTICLE

Comparison of Artificial Intelligence and Guidelines in Answering Questions on Gestational Diabetes: a CLEAR Tool Analysis

Gökhan Köker [1], Lütfullah Z. Koç [1], Muhammed A. Coşkuner [2], Yasin Şahintürk [1], Bilgin B. Başgöz [2]

[1] *Antalya Training and Research Hospital, Department of Internal Medicine, Antalya, Türkiye*
[2] *Antalya State Hospital, Department of Internal Medicine, Antalya, Türkiye*

SUMMARY

Background: **Gestational diabetes mellitus (GDM) affects millions of people worldwide. Patients often turn to the internet and artificial intelligence (AI)-based conversational models for information. The CLEAR tool evaluates the quality of health-related content produced by AI-based models. This study assessed the responses provided by medical guidelines, ChatGPT, and Google Bard to the ten most frequently asked online questions about GDM, utilizing the CLEAR tool for evaluation.**
Methods: **The most common online questions about GDM were identified using Google Trends, and the top 10 questions were selected. Answers were then gathered from two experienced physicians, ChatGPT 4.0o-mini, and Google Bard, with responses categorized into 'Guide,' 'ChatGPT,' and 'Bard' groups. Answers from the AI models were obtained using two computers and two separate sessions to ensure consistency and minimize bias.**
Results: **ChatGPT received higher scores than the medical guidelines, while Bard scored lower than ChatGPT. The medical guidelines provided more accessible answers for the general audience, while ChatGPT and Bard required higher literacy levels. Good reliability (0.781) was observed between the two reviewers. Regarding readability, the medical guidelines were the easiest to read, while Bard provided the most challenging text.**
Conclusions: **ChatGPT and Google Bard perform well in content completeness and relevance but face challenges in readability and misinformation. Future research should improve accuracy and readability, integrate AI with peer-reviewed sources, and ensure healthcare professionals guide patients to reliable AI information.**
(Clin. Lab. 2026;72:xx-xx. DOI: 10.7754/Clin.Lab.2025.250544)

Correspondence:
Muhammed A. Coşkuner
Antalya State Hospital
Department of Internal Medicine
Göçerler 5379 Street
07080, Kepez, Antalya
Türkiye
Phone: + 90 5362596330
Email: coskunermd@gmail.com

Manuscript accepted June 30, 2025

KEYWORDS

gestational diabetes mellitus, artificial intelligence, health information accuracy, CLEAR tool evaluation

INTRODUCTION

Gestational diabetes mellitus (GDM) is defined as diabetes diagnosed during the second or third trimester of pregnancy that was not present before conception [1]. Its global prevalence during pregnancy is approximately 16.9%, affecting millions of expectant mothers worldwide [2]. GDM is associated with a substantial increase in pregnancy-associated complications [3].

The internet is a primary resource for individuals seeking detailed information on this condition. Recently, artificial intelligence (AI)-based conversational models, such as Chat Generative Pre-trained Transformer (ChatGPT) and Google Bard, have gained prominence as sources of information [4]. However, the ease of accessing information through such platforms has also heightened the risk of misinformation. To this concern, various tools have been developed to assess the accuracy and reliability of online information. Among these, the CLEAR tool is specifically designed to evaluate the quality of health-related content produced by AI-based models. CLEAR is an acronym representing the Completeness of the content, Lack of false information, Evidence supporting the content, Appropriateness of the content, and Relevance. Each criterion is scored on a scale from 1 to 5, yielding a maximum total score of 25 [5].

This study assessed the responses provided by medical guidelines, ChatGPT, and Google Bard to the 10 most frequently asked online questions about GDM, utilizing the CLEAR tool for evaluation.

MATERIALS AND METHODS

This study did not involve human participants, human tissue, or identifiable personal data; therefore, institutional ethics committee approval was not required. The analysis was based exclusively on publicly available online content generated by artificial intelligence and medical guidelines. The study was conducted in accordance with the ethical principles of the Declaration of Helsinki.

Determination of questions and answers

We determined the most prevalent online GBM-relevant questions by using Google Trends, searching the keyword "gestational diabetes," setting the region to "worldwide," and setting the time frame to the last 30 days. Among the retrieved questions, we selected the 10 most prevalent ones that can be answered according to the online information published in the American Diabetes Association (ADA)'s website's section "Diabetes and Pregnancy" (https://diabetes.org/living-with-diabetes/pregnancy). We expressed selected questions in Figure 1. After identifying the questions, we decided on the answers by ChatGPT 4.0o-mini, Google Bard, and two experienced physicians according to the ADA guideline. The reference answers based on the ADA guidelines were created by two internal medicine specialists with expertise in diabetes care. To avoid bias, two other internal medicine physicians - who were blinded to the source of each answer - independently evaluated the clarity, reliability, and guideline consistency of the AI-generated responses using the CLEAR tool. To ensure consistency in response structure across all sources, answers derived from the 2024 ADA guidelines were manually rewritten in a clear, question-answer format using natural language. These responses were limited to approximately 60 - 120 words, similar to the expected length and tone of answers generated by large language models (LLMs). This standardization aimed to minimize potential structural bias during evaluation. For illustrative purposes, example responses from the ADA guideline, ChatGPT, and Bard for three representative questions (Q3, Q6, and Q10) are provided in Supplementary Table 1. We listed answers obtained by these three different methods in 3 distinct groups, named "Guide," "ChatGPT," and "Bard." We received answers to both AI models using two computers and in two refreshed sessions each time to minimize potential bias, ensure consistency, and account for differences.

CLEAR tool and comparison of answers

The CLEAR Tool is designed to provide a standardized framework for evaluating the quality of AI-generated health content. This tool assesses content across five distinct criteria: 1) completeness, 2) lack of false information, 3) evidence, 4) appropriateness, and 5) relevance [5]. Each criterion is rated on a 5-point Likert scale, ranging from 1 (poor) to 5 (excellent).

Each response's readability was evaluated using the Flesch-Kincaid grade level (FKGL) [6] and the Flesch-Kincaid reading ease score (FRES) [7]. The FKGL is calculated based on the number of syllables, words, and sentences. Its score reflects a United States grade level, indicating the years of education required to comprehend the text. In contrast, the FRES measures readability ease using the same linguistic components, with higher scores denoting simpler and more accessible text. Two physicians with at least 15 years of experience diagnosing and treating GBM blindly evaluated each answer in three groups individually according to the CLEAR TOOL, FKGL, and FRES without access to their sources. While the CLEAR tool assessed the quality of AI-generated health information, the FKGL and FRES evaluated the readability of the responses. Figure 2 shows the flowchart of the study design.

Statistical analysis

All statistical analyses were performed using SPSS version 26 (IBM SPSS Statistics for Windows). Shapiro-Wilk test was applied to determine the normality of the data distribution. Variables following a normal distribution are presented as mean ± standard deviation, and differences between these groups were assessed using ANOVA, while pairwise comparisons were performed using Student's *t*-test. For skewed variables, they are expressed as median (interquartile range) and minimum-maximum values, and the Kruskal-Wallis test was employed, with pairwise comparisons performed using the Mann-Whitney U test adjusted by the Bonferroni correction. The correlation between the two authors was assessed using a two-way, mixed-effects intraclass correlation. The Flesch reading ease score (FRES) was calculated for each individual answer using the standard formula: FRES = 206.835 - (1.015 × average sentence

Table 1. The comparison of CLEAR tool total and subgroup scores across ChatGPT, Bard, and Guideline.

	CLEAR tool scores			p
CLEAR tool subgroup scores	ChatGPT	Bard	Guideline	
Completeness, median (min-max)	4 (4 - 5)	4 (4 - 4)	3 (3 - 4)	< 0.001
Lack, median (min-max)	4.5 (4 - 5)	4 (4 - 5)	3 (3 - 3)	< 0.001
Evidence, median (min-max)	5 (4 - 5)	4 (4 - 4.75)	3 (3 - 4)	< 0.001
Appropriateness, median (min-max)	4 (4 - 5)	4 (4 - 5)	3 (3 - 3)	< 0.001
Relevance, median (min-max)	4.5 (4 - 5)	4 (4 - 4)	3.5 (3 - 4)	< 0.001
CLEAR tool total score	22.5 (20.25 - 24.5)	21 (20 - 22)	16.5 (15 - 18)	< 0.001

Table 2. The pairwise comparison of total CLEAR score and subgroup scores.

CLEAR tool subgroups	ChatGPT vs. Guide	Bard vs. Guide	ChatGPT vs. Bard
Completeness	< 0.001	< 0.001	0.096
Lack	< 0.001	< 0.001	0.036
Evidence	< 0.001	< 0.001	0.096
Appropriateness	< 0.001	< 0.001	0.277
Relevance	< 0.001	< 0.001	0.081
CLEAR tool total	< 0.001	< 0.001	0.017

1. Does gestational diabetes go away?
2. What are the symptoms of gestational diabetes?
3. When is the gestational diabetes test performed?
4. What are the treatments for gestational diabetes?
5. How can you lower blood sugar during pregnancy?
6. What are the normal ranges for the 1-hour and 2-hour glucose tolerance tests during pregnancy?
7. What are the dietary recommendations for gestational diabetes?
8. What are the best snacks for gestational diabetes?
9. How does gestational diabetes affect the baby?
10. What factors increase the risk of gestational diabetes?

Figure 1. The 10 most common gestational diabetes mellitus-related questions searched online.

This figure presents the most frequently searched questions related to gestational diabetes mellitus (GDM) based on Google Trends data. Topics include diagnostic timing, symptoms, treatment strategies, dietary recommendations, and potential risks to the baby and mother.

length) - (84.6 × average syllables per word). A p-value < 0.05 was accepted as significant.

RESULTS

The median total CLEAR scores for the three groups were as follows: ChatGPT: 22.5 (20.25 - 24.5); Guide: 16.5 (15 - 18); Bard: 21 (20 - 22). A statistically significant difference was observed among the three groups in

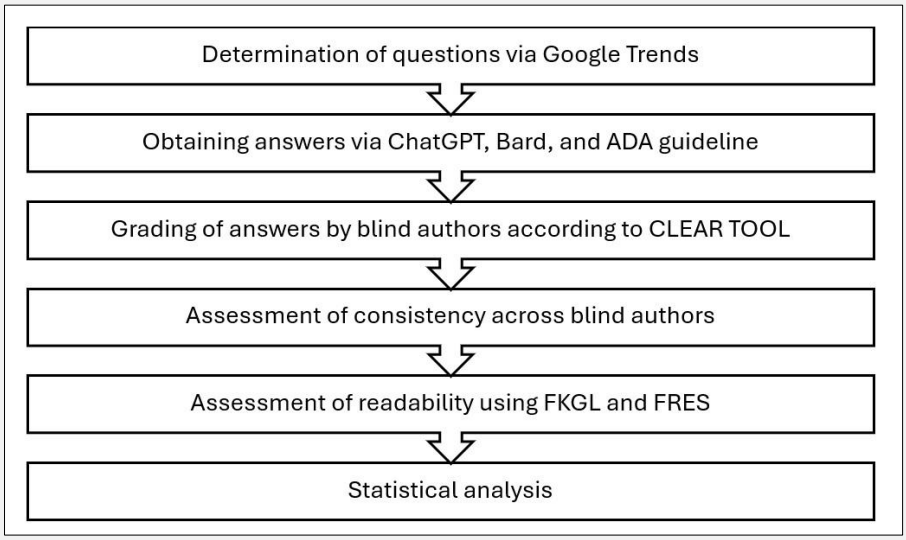

Figure 2. Flowchart of the study design.

This flowchart outlines the methodological process of the study, including selection of online questions, response collection from artificial Intelligence (AI) tools and medical guidelines, and scoring.
ChatGPT Chat Generative Pre-Trained Transformer, ADA American Diabetes Association, FKGL Flesch-Kincaid grade level, FRES Flesch-Kincaid reading ease score.

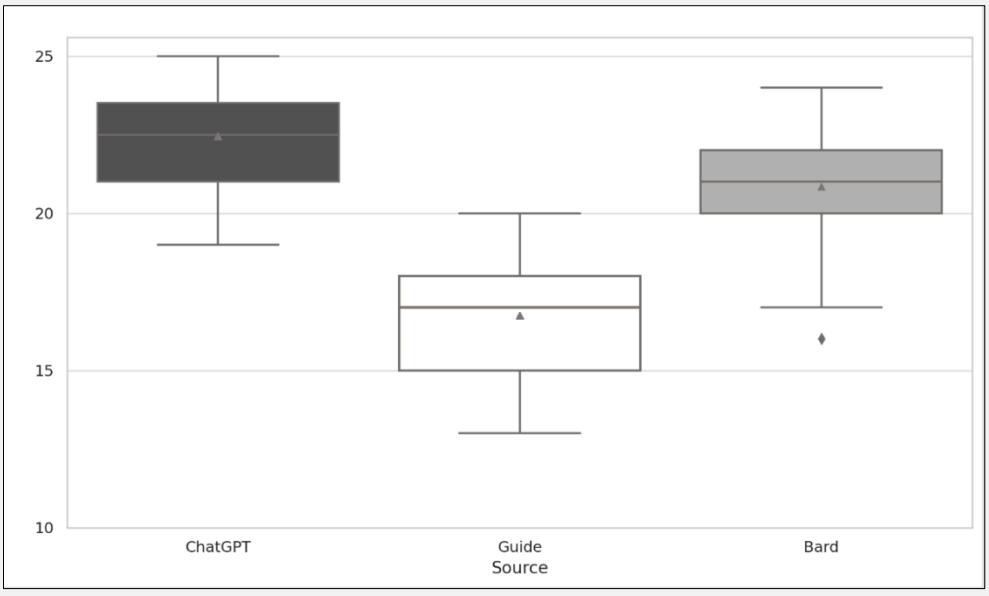

Figure 3. Total CLEAR TOOL scores of groups.

Boxplot comparing the total CLEAR Tool scores for responses generated by AI models (ChatGPT, Bard) and by official medical guidelines. The CLEAR Tool assesses answers across five domains: Correctness, Logical flow, Evidence support, Applicability, and Readability. Higher scores reflect better clinical reliability and overall quality.

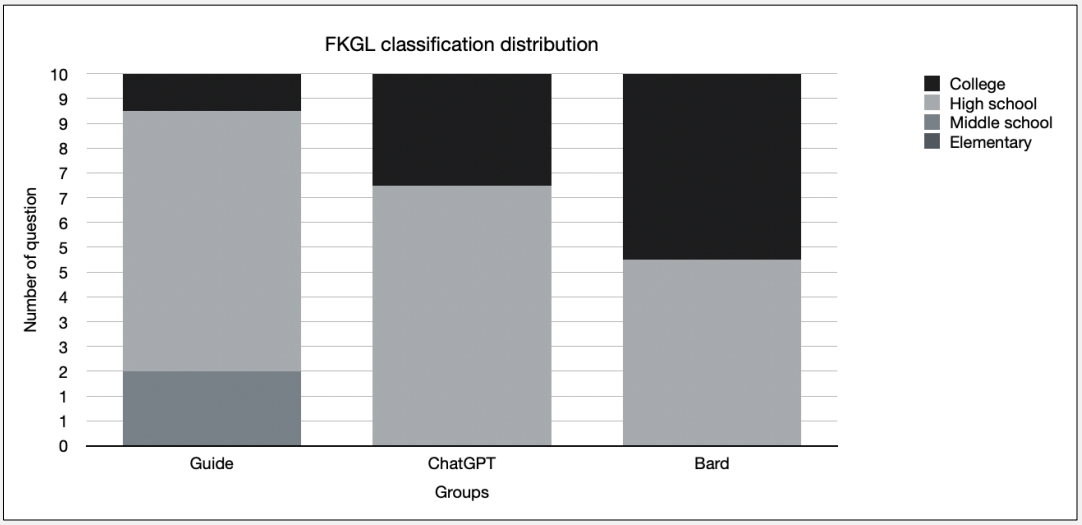

Figure 4. The comparison of FKGL across Guide, ChatGPT and Bard.

Bar graph illustrating the FKGL scores of responses generated by medical guidelines, ChatGPT, and Bard. The FKGL metric indicates the U.S. school grade level required to understand a given text. Lower scores indicate easier readability and simpler language.

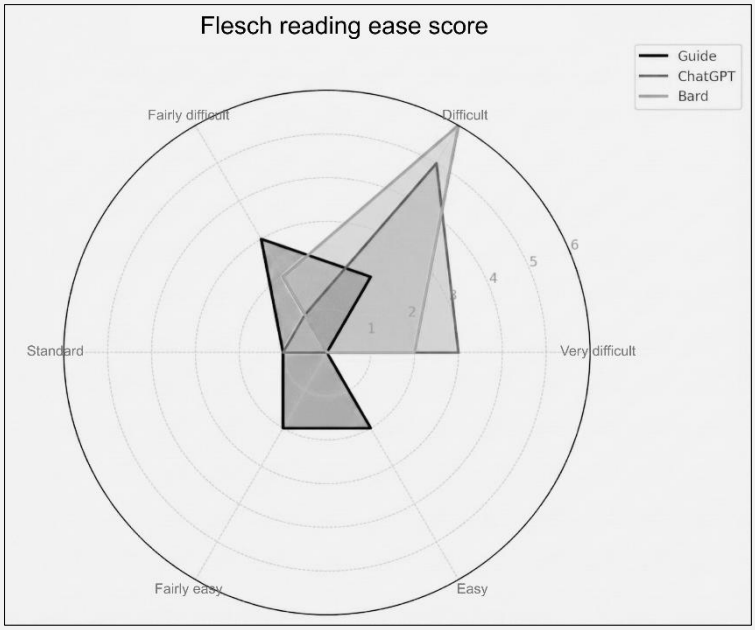

Figure 5. Spider web plot of FRES score distribution across readability categories.

This spider web (radar) plot displays the distribution of responses across the Flesch reading ease classification categories for each source (Guideline, ChatGPT, and Bard). The axes represent the standard FRES readability levels: very difficult, difficult, fairly difficult, standard, fairly easy, and easy. The distance from the center corresponds to the number of responses falling into each category. The plot allows for a visual comparison of the readability profile of each source, with a higher value indicating more responses in that category.

total scores and all subcategories ($p < 0.001$) (Figure 3; Table 1). Post-hoc analysis of CLEAR tool subcategories are reported in Table 2. In pairwise comparisons, we found a significant difference between Guide and ChatGPT in total scores and each subcategory ($p < 0.001$). Similarly, we observed a considerable difference between Guide and Bard in total scores and all subcategories ($p < 0.001$). In comparing ChatGPT and Bard, we noticed a difference only in the total score after the Bonferroni correction ($p = 0.011$, corrected $p < 0.017$). However, we did not see a considerable difference between the groups in any subcategory ($p > 0.05$) other than 'Lack of false information' ($p = 0.036$). Representative examples of incorrect or misleading chatbot responses, along with explanations based on ADA guidelines, are provided in Supplementary Table 2. The intraclass correlation coefficient between the two reviewers was 0.781 (95% CI, 0.539 - 0.896), indicating good reliability between them.

The readability assessment of the Guide, ChatGPT, and Bard responses revealed significant differences in their FKGL (Figure 4) and FRES (Figure 5). The guide had an FKGL mean of 7.20 (3.38), indicating a 7th-grade reading level suitable for students aged 12 - 13 years, and an FRES mean of 65.10 (19.51), categorized as reasonably easy to read. In contrast, ChatGPT displayed an FKG mean of 10.18 (2.07), corresponding to a 10th-grade level appropriate for students aged 15 - 16 years, and an FRES mean of 40.62 (12.69), suggesting a problematic text. Bard demonstrated the most advanced readability, with an FKGL mean of 10.89 (1.88), reflecting an 11th-grade level for students aged 16 - 17 years, and an FRES mean of 37.69 (10.24), classified as very difficult. These results indicate that Guide responses are more accessible to a general audience. At the same time, ChatGPT and Bard require higher reading skills, aligning more closely with high school or early college-level proficiency.

DISCUSSION

The rapid advancement of AI in healthcare has increased interest in its potential role in patient education. Our findings align significantly with existing literature, highlighting the strengths and weaknesses of AI-generated health information. This study evaluated the accuracy, completeness, and readability of responses from ChatGPT, Google Bard, and medical guidelines to ten frequently asked questions about GDM using the CLEAR tool. The findings showed that AI-generated responses were more comprehensive and evidence-based than guideline-based content, but they had significantly lower readability, making them harder for patients to understand. While AI models offer valuable health information, their complexity and potential for misinformation highlight the need for expert oversight and improved readability in AI-assisted patient education.

Our study is consistent with the work of Hulman et al. [8], who evaluated ChatGPT's responses to frequently asked questions about diabetes. Their study found that AI responses often resembled human-generated content but required expert verification. Similarly, Soto-Chavez et al. [9] examined the reliability of ChatGPT in patient education on chronic diseases. They concluded that while AI-generated responses were generally accurate, they lacked readability, presenting a notable limitation. Furthermore, our study aligns with Onder et al.'s study [10], which investigated ChatGPT-4's reliability and readability concerning hypothyroidism. They found that ChatGPT provided largely reliable responses; however, the complexity of the language posed challenges for patient comprehension. Similarly, in our study, AI-generated responses received higher scores on the CLEAR evaluation tool than official guidelines, yet they were more challenging to understand.

Additionally, Cheong et al. [11] compared ChatGPT and Google Bard in providing patient education on obstructive sleep apnea. Their findings indicated that ChatGPT outperformed Google Bard in terms of understandability and actionability. This finding parallels our results, where ChatGPT provided more comprehensive and evidence-based information on GDM.

Campbell et al. [12] assessed ChatGPT's responses regarding obstructive sleep apnea and found that 71.9% of responses were at least partially correct. They also noted that patient-friendly prompts improved readability, yet all responses remained above the recommended literacy threshold for patient education. This finding aligns with our observation that AI-generated responses were less readable than guideline-based materials.

Another study [13] evaluated ChatGPT's responses concerning thyroid nodules, finding that 69.2% of responses were at least partially correct. Furthermore, instructing ChatGPT to include references improved the inclusion of cited sources. This outcome supports our finding that AI models can provide evidence-based content when explicitly prompted.

Shah et al. [14] compared ChatGPT-generated patient education materials with traditional sources, reporting that while ChatGPT achieved higher readability scores, its content was sometimes oversimplified, omitting critical details. This finding contrasts our results, where AI-generated responses scored higher in comprehensiveness but were more challenging to understand.

Unlike previous studies in the literature, our study incorporated a blinded evaluation of official guidelines alongside AI-generated responses. This methodology provided an unbiased comparison between AI-based information and traditional guideline content, eliminating potential reviewer bias. By including blinded evaluations, our study ensured a more objective assessment of the completeness, accuracy, and readability of AI-generated versus guideline-based responses. This unique approach strengthened the validity of our findings and underscores the importance of standardized, blinded

methodologies in evaluating AI applications in patient education.

Furthermore, our study demonstrated a high inter-rater correlation among evaluators, ensuring consistency and reliability in the assessment process. Notably, both the AI-generated responses and the guideline-based information were assessed in a blinded manner, further minimizing bias and increasing the robustness of our findings. This dual-blinded evaluation approach set our study apart from previous research and reinforced the credibility of our conclusions.

A key strength of AI-generated responses is their comprehensiveness, evidence-based content, and high relevance. These models process vast amounts of medical information and present structured responses in an accessible manner. Barlas et al. [15] demonstrated that ChatGPT provides a systematic approach to type 2 diabetes and obesity management, reinforcing its potential for patient education.

However, a significant limitation of AI-generated content is the risk of misinformation or "hallucinations" – instances where AI produces inaccurate or misleading information. Arslan [16] emphasized that while ChatGPT holds promise in obesity treatment, its ethical and safety concerns must be considered.

One of the most critical limitations of AI-generated responses is readability and accessibility. Our study found that while ChatGPT and Google Bard provided complete and evidence-based conclusions, their readability scores were significantly lower than those of official guidelines. This situation poses a challenge, particularly for patients with low health literacy, who may struggle to comprehend AI-generated content.

This study has several limitations. The evaluation was restricted to 10 frequently asked questions about GDM, which may not fully capture the breadth of AI-generated medical information. Expanding the question to include a broader range of topics would enhance generalizability. The study focused solely on ChatGPT-4.0o-mini and Google Bard, excluding other AI models that may offer different levels of accuracy and readability. Future research should include additional models, such as Med-PaLM or Claude, for a more comprehensive comparison. This study reflects AI performance at a single point, while these models are continuously updated and refined. A longitudinal study assessing AI responses over time would offer insights into the evolution of AI-generated medical information and its reliability.

This study provides a comprehensive analysis of the strengths and weaknesses of AI-based models in providing health information on GDM. Our findings indicate that ChatGPT and Google Bard perform well in completeness, evidence-based content, and relevance, yet they present challenges regarding readability and misinformation risk.

Future research should focus on improving the readability and accuracy of AI-based medical information while exploring mechanisms to mitigate AI hallucinations through integration with peer-reviewed medical sources. Additionally, healthcare professionals are crucial in guiding patients toward reliable AI-based attributes and emphasizing the importance of direct medical consultation in clinical decision-making.

Acknowledgment:

The authors declare that they have no acknowledgments to disclose. No individuals or organizations provided support that requires formal recognition, and no financial or material assistance was received for this study. No AI-assisted technologies were used in the preparation or writing of this manuscript.

Source of Funds:

The authors declare that they did not receive any external support.

Declaration of Interest:

The authors declare that they have no conflicts of interest.

References:

1. American Diabetes Association. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2019. Diabetes Care 2019;42(Suppl 1):S13-28. (PMID: 30559228)

2. Guariguata L, Linnenkamp U, Beagley J, Whiting DR, Cho NH. Global estimates of the prevalence of hyperglycaemia in pregnancy. Diabetes Res Clin Pract 2014;103(2):176-85. (PMID: 24300020)

3. Ye W, Luo C, Huang J, Li C, Liu Z, Liu F. Gestational diabetes mellitus and adverse pregnancy outcomes: systematic review and meta-analysis. BMJ 2022;377:e067946. (PMID: 35613728)

4. De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023;11:1166120. (PMID: 37181697)

5. Sallam M, Barakat M, Sallam M. Pilot testing of a tool to standardize the assessment of the quality of health information generated by artificial intelligence-based models. Cureus 2023;15(11): e49373. (PMID: 38024074)

6. Flesch R. A new readability yardstick. J Appl Psychol 1948; 32(3):221-33. (PMID: 18867058)

7. Kincaid JP, Fishburne RP, Rogers RL, Chissom BS. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Institute for Simulation and Training 1975. https://stars.library.ucf.edu/istlibrary/56

8. Hulman A, Dollerup OL, Mortensen JF, et al. ChatGPT- versus human-generated answers to frequently asked questions about diabetes: A Turing test-inspired survey among employees of a Danish diabetes center. PLoS One 2023;18(8):e0290773. (PMID: 37651381)

9. Soto-Chávez MJ, Bustos MM, Fernández-Ávila DG, Muñoz OM. Evaluation of information provided to patients by ChatGPT about chronic diseases in Spanish language. Digit Health 2024;10: 20552076231224603. (PMID: 38188865)

10. Onder CE, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz SM. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. Sci Rep 2024;14(1):243. (PMID: 38167988)

11. Cheong RCT, Unadkat S, Mcneillis V, et al. Artificial Intelligence chatbots as sources of patient education material for obstructive sleep apnoea: ChatGPT versus Google Bard. Eur Arch Otorhinolaryngol 2024;281(2):985-93. (PMID: 37917165)

12. Campbell DJ, Estephan LE, Mastrolonardo EV, Amin DR, Huntley CT, Boon MS. Evaluating ChatGPT responses on obstructive sleep apnea for patient education. J Clin Sleep Med 2023;19(12):1989-95. (PMID: 37485676)

13. Campbell DJ, Estephan LE, Sina EM, et al. Evaluating ChatGPT responses on thyroid nodules for patient education. Thyroid 2024; 34(3):371-7. (PMID: 38010917)

14. Shah YB, Ghosh A, Hochberg AR, et al. Comparison of ChatGPT and traditional patient education materials for men's health. Urol Pract 2024;11(1):87-94. (PMID: 37914380)

15. 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. (PMID: 37951982)

16. Arslan S. Exploring the potential of ChatGPT in personalized obesity treatment. Ann Biomed Eng 2023;51(9):1887-8. (PMID: 37145177)

Additional material can be found online at:
http://supplementary.clin-lab-publications.com/250544/