

New Horizons in Digital Health: The Role of ChatGPT in Knowledge Quality and Patient Education in AI-Assisted Menopause Counseling

 Mehmet Ali Çiftçi

¹ Department of Obstetrics and Gynecology, Menderes State Hospital, İzmir, Türkiye

Correspondence: Mehmet Ali Çiftçi, M.D., mdmehmetaliciftci@gmail.com

Abstract

Aim: This study aimed to assess the accuracy, completeness, and comprehensiveness of large language model based artificial intelligence responses in menopause counseling. Specifically, it evaluated the clinical accuracy and guideline consistency of ChatGPT's answers to commonly asked menopause-related questions. **Methods:** In this cross-sectional descriptive study, twenty frequently asked questions about menopause were developed by five experts and submitted to ChatGPT (GPT-5, OpenAI). Responses generated between June 15 and June 30, 2024, were independently evaluated by 30 obstetricians and gynecologists from five countries using a 5-point Likert scale for accuracy, completeness, and comprehensiveness. Data were analyzed using SPSS version 25. Group differences were tested using ANOVA or Kruskal–Wallis analysis, and inter-rater reliability was calculated via intraclass correlation coefficient. **Results:** ChatGPT achieved an overall mean score of 3.82 (out of 5), with 74.2% of responses rated ≥ 4 . Accuracy had the highest mean score (4.12), followed by comprehensiveness (3.70) and completeness (3.65) ($p:0.004$ and $p:0.007$). The best-performing questions were “How is menopause diagnosed?” (4.45) and “When does menstruation completely stop?” (4.32), while “Do herbal products work?” (3.58) and “Is hormone therapy harmful?” (3.65) scored lowest. Thematic analysis showed the highest rating for “Definition and Physiology of Menopause” (12.85 ± 1.05) and lowest for “Hormone Therapy and Alternatives” (10.85 ± 1.32 , $p:0.005$). Inter-rater reliability was strong (ICC:0.84). **Conclusions:** ChatGPT provided accurate, guideline-consistent menopause information, particularly for diagnostic and physiological topics, though less comprehensive regarding individualized and therapeutic counseling.

Keywords: Menopause, Artificial Intelligence, ChatGPT

1. Introduction

Menopause is a physiological transition marked by hormonal and psychosocial changes that affect women's quality of life¹. Common symptoms such as hot flashes, sleep disturbances, and mood changes necessitate individualized counseling and reliable health information^{2,3}.

Recently, artificial intelligence (AI)-based language models like ChatGPT have emerged as accessible tools for patient education^{4,5}. While they can generate coherent medical explanations, concerns remain regarding their accuracy, completeness, and

Received:

05.01.2026

Accepted:

09.06.2026

Published:

30.06.2026

CC-BY-NC-ND

contextual reliability^{6,7}. However, prior studies indicate that ChatGPT may provide correct yet incomplete information, especially on hormone therapy and contraindications^{8,9}.

This study therefore aimed to systematically evaluate the accuracy, completeness, and comprehensiveness of ChatGPT's responses to common menopause-related questions, as rated by obstetrics and gynecology specialists.

2. Materials and Methods

Patients This cross-sectional descriptive study was designed to evaluate the informational accuracy of an AI-based language model in menopause counseling. The primary objective was to assess ChatGPT's responses to frequently asked menopause-related questions in terms of accuracy, completeness, and comprehensiveness, as rated by obstetrics and gynecology specialists.

2.1. Question Pool Development

In the first phase, a panel of five obstetricians and gynecologists identified the most frequently asked questions by menopausal patients in clinical practice. Based on expert consensus, 20 questions were selected. These covered a broad range of topics including the physiological process of menopause, diagnostic criteria, symptoms, treatment options, sexual health, psychological effects, and lifestyle recommendations.

2.2. Acquisition of ChatGPT Responses

The 20 selected questions were formulated in natural, patient-style English and submitted sequentially to ChatGPT (GPT-5, OpenAI, San Francisco, CA, USA). Responses were obtained between June 15 and June 30, 2024, through the official OpenAI web interface using a standard user account. All sessions were conducted under identical conditions (same browser and internet connection). Each question was phrased as if posed by a patient consulting a physician. Responses were transcribed verbatim without any modification. For each question, a standardized evaluation form was created, including three assessment dimensions:

1. Accuracy: Consistency of the response with current medical evidence and clinical guidelines.
 2. Completeness: Extent to which the response covered all essential aspects of the topic.
 3. Comprehensiveness: Clinical explanatory depth and practical guidance provided.
- Each dimension was rated on a 5-point Likert scale (1=very inadequate, 5=fully adequate).

2.3. Expert Evaluation

The ChatGPT responses were distributed via an online survey to 50 obstetrics and gynecology specialists from five different countries. The evaluations of the first 30 respondents were included in the analysis. Each specialist independently rated ChatGPT's responses for all 20 questions across the three predefined dimensions. In addition to numerical ratings, qualitative feedback regarding ChatGPT's potential role in menopause counseling was also collected.

2.4. Thematic Classification of Questions

To facilitate systematic evaluation, the 20 questions were categorized into nine thematic domains:

1. Definition and Physiology of Menopause
2. Bone and Cardiovascular Health

3. Lifestyle and Exercise
4. Screening and Follow-up
5. Fertility and Reproductive Function
6. Sexual Health and Vaginal Symptoms
7. Symptoms and Psychological Changes
8. Duration and Course
9. Hormone Therapy and Alternatives

This thematic classification enabled comparative analysis of ChatGPT's response quality across different content areas.

2.5. Data Analysis

Data were analyzed using IBM SPSS Statistics software (Version 25.0; IBM Corp., Armonk, NY, USA). Descriptive statistics (mean, standard deviation, minimum–maximum values) were calculated for each question and category. Mean Likert scores were computed separately for accuracy, completeness, and comprehensiveness. Depending on data distribution, one-way ANOVA or Kruskal–Wallis tests were used to compare differences between categories. A p-value <0.05 was considered statistically significant.

2.6. Ethical Considerations

The study was conducted in accordance with the principles of the Declaration of Helsinki and approved by the Ethics Committee of the Republic of Türkiye Ministry of Health, Tepecik Training and Research Hospital (Approval No: 08-01/2025). All participating specialists were informed about the purpose of the study and provided online informed consent prior to participation.

3. Results

A total of 50 obstetrics and gynecology specialists were invited to participate in the survey, and responses from the first 30 participants were included in the analysis. The participants had an average professional experience of 12±6 years; 60% were employed in state hospitals and 40% in university or training and research hospitals. All respondents were actively involved in menopause counseling in their daily clinical practice. Inter-rater reliability analysis showed an overall ICC=0.84 (95% CI: 0.78–0.89), indicating good reliability. Sub-dimension ICCs were 0.87 for accuracy, 0.82 for comprehensiveness, and 0.79 for completeness.

Table 1. Overall Evaluation of ChatGPT Responses

Evaluation Criterion	Mean Score (1–5)±SD	% of Responses Scoring ≥4	p value (vs. Accuracy)
Overall (All Responses)	3.82±0.65	74.2	-
Accuracy	4.12±0.58	84.5	Reference
Comprehensiveness	3.65±0.72	70.1	0.004
Completeness	3.70±0.68	68.3	0.007

SD: Standard Deviation; paired t-test used.

Overall, ChatGPT demonstrated a high level of performance in providing information related to menopause. The mean overall score for all responses was 3.82 (on a 1–5 scale), with 74.2% of responses rated 4 or higher. Among the evaluation criteria, accuracy received the highest mean score (4.12), with 84.5% of responses rated 4 or higher in this domain. Accuracy was followed by comprehensiveness (mean:3.70) and completeness

(mean:3.65), for which 70.1% and 68.3% of the responses, respectively, scored 4 or above. When compared across the three evaluation criteria, ChatGPT's performance showed significant variability. Accuracy (4.12 ± 0.58) was significantly higher than both completeness (3.65 ± 0.72) and comprehensiveness (3.70 ± 0.68) ($p:0.004$ and $p:0.007$, paired t-test) (Table 1).

Significant variations in ChatGPT's performance were observed depending on the question topic. The highest mean scores were achieved for the questions "How is menopause diagnosed?" (4.45) and "When does menstruation completely stop during menopause?" (4.32), whereas the lowest were recorded for "Do herbal products work?" (3.58) and "Is hormone therapy harmful?" (3.65) (Table 2).

Table 2. Topic-Based Performance of ChatGPT Responses

Question Title	Mean Score (1-5)	% of Responses Scoring ≥ 4
How is menopause diagnosed?	4.45	95
When does menstruation completely stop during menopause?	4.32	92
What is early menopause?	4.28	90
How can I tell if I am in menopause?	4.25	88
Does exercise help?	4.18	85
Does menopause affect heart health?	4.15	83
Does menopause affect bone health?	4.12	82
Which tests should I have?	4.08	80
What can be done for vaginal dryness or low libido?	4.05	78
Is sexual intercourse harmful during menopause?	4.02	76
Can I still get pregnant during menopause?	3.98	74
Is weight gain normal during menopause?	3.95	72
Does menopause affect mood?	3.92	70
What causes sleep disturbance?	3.88	68
What are the symptoms of menopause?	3.85	65
How long do menopause symptoms last?	3.82	63
Does menopause cause skin/hair changes?	3.78	60
Is hormone therapy necessary?	3.72	55
Is hormone therapy harmful?	3.65	50
Do herbal products work?	3.58	45

Questions are ranked in descending order of mean score.

Subgroup analysis revealed a consistent pattern: responses to questions explaining diagnostic and physiological mechanisms received significantly higher scores (mean 4.33 ± 0.12), while those related to treatment options and alternative approaches scored significantly lower (mean 3.65 ± 0.18) ($p < 0.001$, ANOVA).

Detailed categorical analysis demonstrated that the "Definition and Physiology of Menopause" category achieved the highest total score (12.85 ± 1.05). According to one-way ANOVA and post-hoc Tukey tests, all other categories had significantly lower total scores compared to this reference group ($p < 0.05$). The "Hormone Therapy and Alternatives" category showed the lowest overall performance (10.85 ± 1.32 ; accuracy 3.75, comprehensiveness 3.55), with a statistically significant difference from the reference category ($p:0.005$) (Table 3).

Table 3. Distribution of ChatGPT Performance by Question Category

Category	Total Score (Mean $\pm$ SD)	Accuracy	Comprehensiveness	Completeness	p value
Definition and Physiology of Menopause	12.85 ± 1.05	4.35	4.25	4.25	Reference
Bone and Cardiovascular Health	12.15 ± 1.12	4.12	4.02	4.01	0.038

Lifestyle and Exercise	12.08±1.08	4.08	4.00	4.00	0.042
Screening and Follow-up	11.95±1.15	4.05	3.95	3.95	0.028
Fertility and Reproductive Function	11.82±1.18	4.02	3.90	3.90	0.025
Sexual Health and Vaginal Symptoms	11.75±1.20	3.98	3.89	3.88	0.021
Symptoms and Psychological Changes	11.45±1.22	3.92	3.77	3.76	0.015
Duration and Course	11.28±1.25	3.88	3.70	3.70	0.012
Hormone Therapy and Alternatives	10.85±1.32	3.75	3.55	3.55	0.005

Total score: sum of the three sub-criteria (Accuracy + Comprehensiveness + Completeness). Categories are ranked by total score.

4. Discussion

Bone This study represents one of the first clinician-based evaluations of LLMs in menopause counseling, assessing the accuracy and content depth of AI-generated medical information. Our findings demonstrate that ChatGPT provides generally accurate responses to menopause-related questions (mean accuracy:4.12), although its comprehensiveness (3.70) and completeness (3.65) were significantly lower ($p:0.004$, $p:0.007$). The fact that 74.2% of responses scored 4 or higher indicates that the model can deliver clinically reliable information within the context of menopause counseling. These findings align with systematic reviews showing that ChatGPT performs well in general medical accuracy but remains limited in contextual depth and clinical nuance^{10,11}.

Recent reviews have reported ChatGPT's accuracy in medical domains to range between 60% and 80%, with stronger performance on factual questions ("diagnosis," "physiology") but lower accuracy on questions involving treatment recommendations or decision-making¹². Our study mirrored this pattern: "Definition and Physiology of Menopause" achieved the highest category score (12.85±1.05), whereas "Hormone Therapy and Alternatives" performed significantly worse (10.85±1.32, $p:0.005$). These results support the notion that ChatGPT is proficient in factual knowledge transfer but less capable in guiding individualized clinical decisions. Similarly, Shahsavari et al. found that ChatGPT was more reliable for "what"-type factual questions but more inconsistent for "how" and "why" explanatory questions¹³, consistent with our finding that questions about hormone therapy and herbal treatments received lower ratings.

Hormone therapy (HT) and its alternatives remain controversial topics in menopausal medicine. The 2023 North American Menopause Society (NAMS) guideline emphasizes that HT is the most effective treatment for suitable candidates but must be individualized¹⁴. Conversely, evidence supporting herbal or "natural" products remains weak, and such therapies are not recommended in clinical guidelines¹⁵. ChatGPT's cautious and generalized tone in these areas likely preserved factual accuracy but limited content depth and clinical applicability, explaining the lower scores for the questions "Is hormone therapy harmful?" and "Do herbal products work?"

The higher accuracy relative to comprehensiveness and completeness suggests that ChatGPT tends to avoid factual errors yet provides shallow explanations. Similar findings were reported by Beheshti et al., who noted that ChatGPT's medical responses were mostly correct but superficial¹⁶. Balci et al. likewise found that while ChatGPT's answers were generally accurate, they lacked completeness and contextual relevance¹⁷. These observations are consistent with our characterization of the model as "accurate but incomplete."

The top-scoring questions in our study related to diagnosis and physiology demonstrate that ChatGPT possesses strong lexical and conceptual knowledge of medical terminology and physiological mechanisms. Conversely, questions requiring patient-specific or therapeutic reasoning elicited weaker responses, reflecting a behavior pattern widely documented in LLM performance literature¹³. This highlights the need for cautious use of

ChatGPT in individualized counseling contexts such as menopause management, where clinical judgment and patient-tailored approaches are essential.

Our study is among the few clinician-based assessments of LLM performance in women's health. Similarly, Peled et al. reported an accuracy rate of around 70% for ChatGPT's obstetric responses and emphasized the importance of cautious interpretation of AI-generated information in patient safety contexts¹¹. Our findings parallel these results, suggesting that while ChatGPT provides satisfactory informational support, it remains insufficient as an autonomous clinical decision-making tool.

5. Conclusions

ChatGPT demonstrates high informational accuracy in menopause counseling, particularly regarding physiological and diagnostic topics. However, its responses remain limited in depth and scope for subjects such as hormone therapy, alternative approaches, and individualized clinical decision-making. These results indicate that AI-based models should be regarded as guideline-aligned informational aids rather than direct substitutes for professional clinical judgment in complex, risk-dependent domains such as menopause management.

Author Contributions: Conceptualization, M.A.Ç.; methodology, M.A.Ç.; validation, M.A.Ç.; formal analysis, M.A.Ç.; investigation, M.A.Ç.; resources, M.A.Ç.; data curation, M.A.Ç.; writing—original draft preparation, M.A.Ç.; writing—review and editing, M.A.Ç.; visualization, M.A.Ç.; supervision, M.A.Ç.; project administration, M.A.Ç. The author has read and agreed to the published version of the manuscript.

GenAI Statement: No generative artificial intelligence tools were used in the preparation of this manuscript. ChatGPT was evaluated solely as the subject of the research and was not used for manuscript writing, data analysis, or interpretation.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study because no human participants, patient data, biological materials, or animal subjects were involved.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated or analyzed during this study are included in this published article. Additional information is available from the corresponding author upon reasonable request.

Acknowledgments: The author would like to thank all clinicians and researchers whose expertise contributed to the development of the study questions and evaluation framework.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Akdemir C, Balcı MF, Şanlı M, et al. Is It Time to Revise Cervical Cancer Screening Guidelines? *Ankara Medical Journal*. 2025;25(2):182-192. doi:10.5505/amj.2025.37891.
2. Santoro N, Epperson CN, Mathews SB. Menopausal symptoms and their management. *Endocrinol Metab Clin North Am*. 2015;44(3):497-515. doi:10.1016/j.ecl.2015.05.001.
3. Hess R, Thurston RC, Hays RD, et al. The impact of menopause on health-related quality of life: results from the STRIDE longitudinal study. *Qual Life Res*. 2012;21(3):535-544. doi:10.1007/s11136-011-9959-7.

4. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2). doi:10.1371/journal.pdig.0000198.
5. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health*. 2023;5(3). doi:10.1016/S2589-7500(23)00021-3.
6. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388(13):1233-1239. doi:10.1056/NEJMs2214184.
7. Bolgova O, Shypilova I, Sankova L, Mavrych V. How well did ChatGPT perform in answering questions on different topics in gross anatomy? *Eur J Med Health Sci*. 2023;5(6):94-100. doi:10.24018/ejmed.2023.5.6.1989.
8. Sütçüoğlu BM, Güler M. Appropriateness of premature ovarian insufficiency recommendations provided by ChatGPT. *Menopause*. 2023;30(10):1033-1037. doi:10.1097/GME.0000000000002246.
9. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*. 2025. doi:10.1001/jama.2024.21700.
10. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. *J Biomed Inform*. 2024;151:104620. doi:10.1016/j.jbi.2024.104620.
11. Peled T, Sela HY, Weiss A, Grisaru-Granovsky S, Agrawal S, Rottenstreich M. Evaluating the validity of ChatGPT responses on common obstetric issues: Potential clinical applications and implications. *Int J Gynecol Obstet*. 2024;166(3):1127-1133. doi:10.1002/ijgo.15501.
12. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. *J Med Internet Res*. 2024;26. doi:10.2196/60807.
13. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. *JMIR Hum Factors*. 2023;10(1). doi:10.2196/47564.
14. The 2023 nonhormone therapy position statement of The North American Menopause Society. *Menopause*. 2023;30(6):573-590. doi:10.1097/GME.0000000000002200.
15. Compounded Bioidentical Menopausal Hormone Therapy: ACOG Clinical Consensus No. 6. *Obstet Gynecol*. 2023;142(5):1266-1273. doi:10.1097/AOG.0000000000005395.
16. Beheshti M, Toubal IE, Alaboud K, et al. Evaluating the reliability of ChatGPT for health-related questions: A systematic review. *Informatics*. 2025;12(1):9. doi:10.3390/informatics12010009.
17. Balcı M, Akdemir C, Yıldırım F. Is Artificial Intelligence-Assisted Pregnancy Counseling Feasible? An Evaluation of the Quality of ChatGPT Responses. *Cukurova Anestesi ve Cerrahi Bilimler Dergisi*. 2025;8:336-340. doi:10.36516/jocass.1753846.