DOI: 10.4274/anatoljmed.2026.97759
Anatol J Gen Med Res 2026;36(1):84-9

CLINICAL RESEARCH / KLİNİK ARAŞTIRMA

ANATOLIAN JOURNAL of GENERAL MEDICAL RESEARCH

Patients Ask, AI Answers: An Evaluation of ChatGPT-4o Responses to the Most Frequently Googled Questions on Laparoscopic Hysterectomy

Hasta Eğitiminde Yapay Zeka: Laparoskopik Histerektomi Hakkında Google'da En Sık Aranan Sorulara ChatGPT-4o Yanıtlarının Değerlendirilmesi

Alaattin Karabulut[1], Sercan Kantarcı[1], Uğurcan Dağlı[1], Ahmet Güdüklü[2], İlker Uçar[1], Volkan Karataşlı[3], Abdurrahman Hamdi İnan[1]

[1]University of Health Sciences Türkiye, İzmir Tepecik Education and Research Hospital, Department of Obstetrics and Gynecology, İzmir, Türkiye
[2]Foça State Hospital, Clinic of Obstetrics and Gynecology, İzmir, Türkiye
[3]University of Health Sciences Türkiye, Balıkesir Atatürk City Hospital, Department of Gynecologic Oncology, Balıkesir, Türkiye

Cite as: Karabulut A, Kantarcı S, Dağlı U, et al. Patients ask, AI answers: an evaluation of ChatGPT-4o responses to the most frequently googled questions on laparoscopic hysterectomy. Anatol J Gen Med Res. 2026;36(1):84-9

Abstract

Objective: Patients increasingly use search engines and artificial intelligence-based tools to obtain medical information prior to surgical consultations. This study aimed to evaluate the quality of ChatGPT-4o's responses to the patient questions most frequently searched on Google regarding laparoscopic hysterectomy.

Methods: This expert-based evaluation identified the most frequently searched patient questions on Google regarding laparoscopic hysterectomy using a newly created Google account with no prior search history. After eliminating duplicates, 24 unique questions were included. Each question was independently presented to ChatGPT-4o using a new account with no prior conversation history. Ten experienced gynecologic surgeons evaluated each response across four domains: accuracy, clarity, completeness, and relevance, using a five-point Likert scale. Descriptive statistics were calculated, and interrater reliability was assessed using intraclass correlation coefficients.

Results: Mean scores across all questions were 4.36 for accuracy, 4.55 for clarity, 4.21 for completeness, and 4.59 for relevance. No responses received a score below 3 in any domain. Interrater reliability varied across evaluation domains. Overall, ChatGPT-4o responses were perceived as highly relevant and easy to understand, with comparatively lower scores for completeness.

Conclusion: ChatGPT-4o provides responses to common online queries about laparoscopic hysterectomy that are generally accurate, relevant, and clearly presented. However, limitations in completeness highlight the need for clinician-led counseling. Artificial intelligence-generated information may serve as a supplementary resource for patient education but should not replace professional medical guidance.

Keywords: Laparoscopic hysterectomy, artificial intelligence, patient education, ChatGPT, minimally invasive surgical procedures

Address for Correspondence/Yazışma Adresi: Alaattin Karabulut, MD, University of Health Sciences Türkiye, İzmir Tepecik Education and Research Hospital, Department of Obstetrics and Gynecology, İzmir, Türkiye
E-mail: alaattin_karabulut@hotmail.com
ORCID ID: orcid.org/0000-0002-0244-4401

Received/Geliş tarihi: 24.01.2026
Accepted/Kabul tarihi: 23.02.2026
Published date/Yayınlanma tarihi: 30.04.2026

Öz

Amaç: Hastalar, cerrahi girişimler öncesinde tıbbi bilgi edinmek amacıyla giderek artan şekilde arama motorları ve yapay zeka tabanlı araçlara başvurmaktadır. Bu çalışmanın amacı, laparoskopik histerektomi ile ilgili Google'da en sık aranan hasta sorularına ChatGPT-4o tarafından verilen yanıtların kalitesini değerlendirmektir.

Yöntem: Bu uzman temelli değerlendirme çalışmasında, laparoskopik histerektomiye ilişkin en sık aranan hasta soruları, daha önce arama geçmişi bulunmayan yeni bir Google hesabı kullanılarak belirlendi. Tekrar eden sorular çıkarıldıktan sonra toplam 24 soru çalışmaya dahil edildi. Her soru, daha önce konuşma geçmişi bulunmayan yeni bir hesap kullanılarak ChatGPT-4o'ya ayrı ayrı yöneltildi. Elde edilen yanıtlar, en az beş yıllık cerrahi deneyime sahip on jinekolog tarafından doğruluk, netlik, eksiksizlik ve alaka başlıkları altında beşli Likert ölçeği kullanılarak bağımsız olarak değerlendirildi. Tanımlayıcı istatistikler hesaplandı ve değerlendiriciler arası uyum sınıf içi korelasyon katsayısı ile analiz edildi.

Bulgular: Tüm sorular için ortalama puanlar doğruluk için 4,36, netlik için 4,55, eksiksizlik için 4,21 ve alaka için 4,59 olarak bulundu. Hiçbir yanıt herhangi bir değerlendirme alanında 3 puanın altında skor almadı. Değerlendiriciler arası güvenilirlik, değerlendirme alanlarına göre değişiklik göstermiştir. Yanıtlar genel olarak yüksek alaka ve anlaşılabilirlik düzeyi göstermesine karşın, eksiksizlik alanında görece daha düşük puanlar aldı.

Sonuç: ChatGPT-4o, laparoskopik histerektomi ile ilgili sık sorulan hasta sorularına genel olarak doğru, ilgili ve anlaşılır yanıtlar sunmaktadır. Bununla birlikte, eksiksizlik konusundaki sınırlılıklar, yapay zeka temelli bilgilerin hasta eğitimi açısından tamamlayıcı bir araç olarak değerlendirilmesi gerektiğini ve hekim danışmanlığının yerini alamayacağını göstermektedir.

Anahtar Kelimeler: Laparoskopik histerektomi, yapay zeka, hasta eğitimi, ChatGPT, minimal invaziv cerrahi

Introduction

Patients increasingly seek medical information online before consulting a physician, particularly for elective surgical procedures that raise concerns about safety, recovery, and long-term outcomes. Early studies demonstrated that a substantial proportion of patients use the internet as a primary source of health information, often prior to clinical encounters[1]. However, the quality, reliability, and interpretability of online medical content remain highly variable, and patients frequently struggle to distinguish accurate information from misleading or incomplete sources[2]. This phenomenon, commonly referred to as "Dr. Google", has reshaped the physician-patient relationship by influencing expectations, decision-making, and preoperative counseling[3,4].

With recent advances in artificial intelligence, large language models (LLMs) such as ChatGPT have emerged as a new generation of information tools that differ fundamentally from traditional web searches. Instead of directing users to static webpages, these models generate conversational, context-aware responses that may be perceived as personalized and authoritative. Consequently, LLMs have rapidly gained popularity among patients seeking answers to medical questions, including those related to surgical procedures. In obstetrics and gynecology, where patient education is often challenged by complex terminology and variable health literacy, AI-generated responses may offer both opportunities and risks[5-7]. While recent studies suggest that ChatGPT can provide generally accurate and readable information across various medical domains, concerns remain regarding consistency, completeness, and potential omissions of clinically relevant details[8,9].

Laparoscopic hysterectomy is one of the most commonly performed minimally invasive gynecologic surgeries worldwide and is associated with numerous patient concerns related to perioperative safety, postoperative recovery, sexual function, and long-term quality of life. These concerns frequently drive patients to search online for information prior to surgery. Despite the growing use of LLMs by patients, limited data exist on the quality of ChatGPT-generated responses to frequently searched questions about laparoscopic hysterectomy. This study aimed to evaluate the accuracy, clarity, completeness, and relevance of ChatGPT-4o responses to the most frequently Googled patient questions about laparoscopic hysterectomy, as assessed by experienced gynecologic surgeons.

Materials and Methods

This expert evaluation study assessed the quality of artificial intelligence-generated responses to patient-centered questions regarding laparoscopic hysterectomy. The study did not involve direct patient participation or the collection of personal health data. The ChatGPT-generated responses were independently evaluated by expert gynecologic surgeons using standardized assessment forms, and written informed consent was obtained from all expert gynecologic surgeons who served as evaluators in the study. Ethical approval was obtained from the University of Health Sciences

Türkiye, İzmir Tepecik Education and Research Hospital Non-Interventional Research Ethics Committee (approval no: 2025/08-25, date: 11.09.2025), prior to study initiation, and the study was conducted in accordance with the principles of the Declaration of Helsinki.

A new Google account without prior search history was created to minimize personalization bias when identifying patient-centered questions. The "people also ask" and related question features were reviewed using predefined search terms related to laparoscopic hysterectomy to generate an initial pool of commonly searched questions. Approximately 100 questions were collected during this phase. Two investigators independently reviewed the questions to remove duplicates, merge conceptually overlapping items, and exclude questions unrelated to patient education. Any disagreements were resolved by consensus. This process resulted in a final list of 24 unique questions reflecting patients' most frequently searched concerns on Google regarding laparoscopic hysterectomy. The full list of the most frequently Googled patient questions about laparoscopic hysterectomy that were included in the analysis is presented in Table 1.

Table 1. Frequently Googled patient questions about laparoscopic hysterectomy included in the study

Question number	Question
1	What are the do's and don'ts after a laparoscopic hysterectomy?
2	What are the disadvantages of laparoscopic hysterectomy?
3	How long is bed rest after laparoscopic hysterectomy?
4	What food to avoid after laparoscopic hysterectomy surgery?
5	Which surgery is best for a hysterectomy?
6	What can prolapse after a laparoscopic hysterectomy?
7	What to know before having a laparoscopic hysterectomy?
8	What is the most common complication of a laparoscopic hysterectomy?
9	How much blood is lost during a laparoscopic hysterectomy?
10	What helps you heal faster after a laparoscopic hysterectomy?
11	How much walking is okay after a laparoscopic hysterectomy?
12	What is the best position to sit after a laparoscopic hysterectomy?
13	Why is the stomach swollen after a laparoscopic hysterectomy?
14	What are the risk factors for a laparoscopic hysterectomy?
15	What to expect 3 months after a laparoscopic hysterectomy?
16	What size uterus can be removed laparoscopically?
17	Is it okay to push to poop after a laparoscopic hysterectomy?
18	How soon can I fly after a laparoscopic hysterectomy?
19	How can I naturally lubricate after a laparoscopic hysterectomy?
20	Where do eggs go after a laparoscopic hysterectomy?
21	How many days bleeding after laparoscopic hysterectomy?
22	What is the best alternative to a laparoscopic hysterectomy?
23	What is the best age to get a laparoscopic hysterectomy?
24	What is the regret rate for a laparoscopic hysterectomy?

All questions were presented individually to ChatGPT (version 4o) via a newly created account with no prior conversation history to avoid contextual carryover. No additional prompts, clarifications, or follow-up questions were provided. Each question was entered separately, and responses were recorded verbatim. To reduce temporal and sequential biases, questions were administered at different time points rather than during a single session. The generated responses were not edited, summarized, or reformatted prior to evaluation.

Ten board-certified gynecologic surgeons with at least five years of independent clinical experience in minimally invasive gynecologic surgery participated as evaluators. All evaluators were blinded to the study design details and informed only that they were assessing written responses to patient questions related to laparoscopic hysterectomy. Evaluators were not informed that the responses had been generated by an artificial intelligence model.

Each response was independently evaluated by all ten surgeons using a five-point Likert scale across four predefined

domains: accuracy, clarity, completeness, and relevance. Accuracy reflected the factual correctness of the information provided; clarity assessed how easily the response could be understood by a lay audience; completeness evaluated whether key aspects of the question were adequately addressed; and relevance measured how well the response aligned with the patient's original question. Scores ranged from 1 (very poor) to 5 (excellent) for each domain.

The primary outcomes of the study were the mean scores for accuracy, clarity, completeness, and relevance across all questions. The secondary outcome was interrater reliability among evaluators within each evaluation domain.

Statistical Analysis

Descriptive statistics were used to summarize scores for each evaluation domain and were reported as mean ± standard deviation. Interrater reliability was assessed using intraclass correlation coefficients (ICCs) estimated from a two-way random-effects model with absolute agreement and average measures. This model was selected to reflect agreement among multiple independent raters evaluating the same set of responses. ICC values were interpreted according to established guidelines.

Results

A total of 24 ChatGPT-4o-generated responses corresponding to the most frequently Googled patient questions about laparoscopic hysterectomy were evaluated independently by 10 experienced gynecologic surgeons. Each response was scored across four predefined domains: accuracy, clarity, completeness, and relevance.

The mean accuracy score across all questions was 4.36±0.51. No response received a score below 3 from any evaluator. Several questions received a perfect mean score of 5.0, indicating unanimous agreement that the information provided was factually correct. Interrater reliability for accuracy yielded an ICC of 0.43 (95% confidence interval, 0.27-0.54).

Clarity received high overall ratings among the four evaluation domains. The mean clarity score was 4.55±0.32, with many responses showing minimal variability between evaluators. Multiple questions exhibited complete agreement, as reflected by a standard deviation of zero. Interrater reliability for clarity yielded an ICC of 0.48 (95% confidence interval, 0.29-0.64), which was the highest ICC observed among the evaluated domains.

The mean completeness score across all responses was 4.21±0.50. Although this domain yielded the lowest average score compared with those for accuracy, clarity, and relevance, each response was still rated above 3 by all evaluators. Interrater reliability for completeness yielded an ICC of 0.42 (95% confidence interval, 0.23-0.55).

Relevance achieved the highest mean score overall, with an average of 4.59±0.38. Responses were consistently judged to be well aligned with the corresponding patient questions, and no response was considered irrelevant. Interrater reliability analysis for relevance yielded an ICC of 0.33 (95% confidence interval, 0.17-0.45).

Across all four domains, ChatGPT-4o responses were rated favorably by expert evaluators, with mean scores exceeding 4.0 in each category. Mean scores for each evaluation domain, along with interrater reliability results, are summarized in Table 2.

Discussion

This study evaluated the quality of ChatGPT-4o responses to the most frequently Googled patient questions about laparoscopic hysterectomy using expert surgeon assessments across four domains: accuracy, clarity, completeness, and relevance. The findings demonstrate that ChatGPT-4o generally provides responses that are perceived as accurate, highly relevant, and easily understandable, with mean scores exceeding 4.0 in all evaluated domains. These results support the growing body of literature suggesting that LLMs can serve as a valuable adjunct source of medical information for patients seeking preliminary guidance before clinical consultation.

Among the evaluated domains, relevance and clarity achieved the highest mean scores. This finding indicates that ChatGPT-4o is particularly effective at aligning its responses with the intent of patients' questions and presenting information in a manner accessible to a lay audience. Similar observations have been reported in previous evaluations of

Table 2. Mean scores of ChatGPT-4o responses across evaluation domains

Evaluation domain	Mean ± SD	ICC
Accuracy	4.36±0.51	0.43
Clarity	4.55±0.32	0.48
Completeness	4.21±0.50	0.42
Relevance	4.59±0.38	0.33

SD: Standard deviation, ICC: Intraclass correlation coefficient

ChatGPT in surgical and gynecologic contexts, where clarity and relevance consistently outperformed other qualitative dimensions[8-10]. From a patient education perspective, these strengths are clinically meaningful, as misunderstandings related to surgical procedures often arise from unclear or overly technical explanations.

Accuracy scores were also high, with no response receiving a rating below 3 from any evaluator and several responses achieving unanimous agreement. This finding aligns with earlier studies demonstrating that ChatGPT frequently provides factually correct information across a range of medical and surgical topics[10,11]. However, accuracy alone does not equate to clinical adequacy. While responses were largely correct, they were not uniformly comprehensive, underscoring the importance of clinician involvement in patient counseling.

Completeness emerged as the lowest-scoring domain, although mean scores remained above 4.0. This pattern has been consistently reported in prior studies and reflects an inherent characteristic of LLMs, which tend to prioritize concise and broadly applicable responses rather than exhaustive clinical detail[8,12]. In the context of laparoscopic hysterectomy, this tendency may result in the omission of nuanced considerations such as individualized risk factors, rare complications, or institution-specific practices. While brevity may enhance readability for patients, it also highlights a key limitation of relying on AI-generated information for comprehensive preoperative counseling. This observation is consistent with broader evaluations of LLMs in clinical settings, which emphasize that although AI systems can support information delivery and patient engagement, they lack the capacity to replace clinician judgment and individualized decision-making[13].

According to established guidelines, the observed ICCs fall within the poor reliability range. However, ICC values should not be interpreted in isolation. As highlighted in recent methodological literature, studies characterized by restricted score variability and clustering of ratings at the upper end of the scale are expected to yield lower ICC estimates, despite consistent evaluator judgments. In the present study, the high mean scores and limited dispersion across evaluation domains suggest a ceiling effect, which likely contributed to reduced ICC values rather than true disagreement among evaluators[9,10,14-16].

The results of this study have practical implications for patient education in minimally invasive gynecologic surgery. As patients increasingly turn to search engines and conversational AI tools for health-related information, ChatGPT-4o may function as an accessible entry point for understanding common aspects of laparoscopic hysterectomy. However, the observed limitations in completeness reinforce the notion that AI-generated responses should complement, rather than replace, physician-led counseling. Clinicians should be aware of the types of information patients may encounter online and proactively address gaps or misconceptions during preoperative discussions.

Study Limitations

Several limitations warrant consideration. First, the study evaluated responses generated by a single version of ChatGPT at a specific time point, and model performance may evolve with future updates. Second, assessments were conducted exclusively by expert surgeons, which may not reflect patients' perceptions of quality or usefulness. Third, the study focused on commonly searched questions and may not capture less frequent but clinically important concerns. Another important limitation of this study is the absence of a direct comparator. ChatGPT-4o responses were not evaluated against established patient education resources, such as professional society guidelines, institutional websites, or standardized patient information leaflets. Therefore, the present findings should be interpreted as an assessment of AI-generated information in isolation rather than a comparative evaluation against existing educational standards. Future studies incorporating direct comparisons with validated patient education materials may provide additional insights into the relative strengths and limitations of AI-based tools. Despite these limitations, the standardized methodology, expert blinding, and use of multiple evaluative domains strengthen the robustness of the findings.

Conclusion

ChatGPT-4o provides responses to commonly searched questions about laparoscopic hysterectomy that are generally accurate, relevant, and clear, though occasionally incomplete. These findings suggest that while LLMs hold promise as supplementary tools for patient education, their outputs should be interpreted within the context of professional medical guidance.

Ethics

Ethics Committee Approval: Ethical approval was obtained from the University of Health Sciences Türkiye, İzmir Tepecik

Education and Research Hospital Non-Interventional Research Ethics Committee (approval no: 2025/08-25, date: 11.09.2025).

Informed Consent: Written informed consent was obtained from all expert gynecologic surgeons who participated in the study by completing the evaluation forms.

Footnotes

Authorship Contributions

Concept: A.K., A.H.İ., Design: A.K., A.G., V.K., Data Collection or Processing: A.K., S.K., U.D., A.G., İ.U., V.K., A.H.İ., Analysis or Interpretation: U.D., İ.U., Literature Search: A.K., S.K., A.G., İ.U., V.K., Writing: A.K., S.K., U.D., A.H.İ.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Diaz JA, Griffith RA, Ng JJ, Reinert SE, Friedmann PD, Moulton AW. Patients' use of the internet for medical information. J Gen Intern Med. 2002;17:180-5.
2. Eysenbach G, Köhler C. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. BMJ. 2002;324:573-7.
3. Van Riel N, Auwerx K, Debbaut P, Van Hees S, Schoenmakers B. The effect of Dr Google on doctor-patient encounters in primary care: a quantitative, observational, cross-sectional study. BJGP Open. 2017;1:bjgpopen17X100833.
4. Cocco AM, Zordan R, Taylor DM, et al. Dr Google in the ED: searching for online health information by adult emergency department patients. Med J Aust. 2018;209:342-7.
5. Fahimuddin FZ, Sidhu S, Agrawal A. Reading level of online patient education materials from major obstetrics and gynecology societies. Obstet Gynecol. 2019;133:987-93.
6. Oliveira JA, Eskandar K, Kar E, de Oliveira FR, Filho ALDS. Understanding AI's role in endometriosis patient education and evaluating its information and accuracy: systematic review. JMIR AI. 2024;3:e64593.
7. Daram NR, Maxwell RA, D'Amato J, Massengill JC. Can artificial intelligence improve the readability of patient education information in gynecology? Am J Obstet Gynecol. 2025;233:640.e1-9.
8. Sparks CA, Fasulo SM, Windsor JT, et al. ChatGPT is moderately accurate in providing a general overview of orthopaedic conditions. JB JS Open Access. 2024;9:e23.00129.
9. West M, Alsaidi A, Siddiqi R, et al. Artificial intelligence in obstetrics and gynecology: evaluating ChatGPT and Google Gemini in answering patient questions. Int J Gynaecol Obstet. 2026;173:328-35.
10. Ayık G, Ercan N, Demirtaş Y, Yıldırım T, Çakmak G. Evaluation of ChatGPT-4o's answers to questions about hip arthroscopy from the patient perspective. Jt Dis Relat Surg. 2025;36:193-9.
11. Zhou G, Wang Y, Che X. A comparative study of AI chatbots and traditional medical sources for hysterectomy patient education: assessing professionalism, readability, and patient education quality. Medicine (Baltimore). 2025;104:e44403.
12. Shao CY, Li H, Liu XL, et al. Appropriateness and comprehensiveness of using ChatGPT for perioperative patient education in thoracic surgery in different language contexts: survey study. Interact J Med Res. 2023;12:e46900.
13. Büyüktoka RE, Salbas A. Multimodal large language models for pediatric bone-age assessment: a comparative accuracy analysis. Acad Radiol. 2025;32:6905-12.
14. Taşkum İ, Sınacı S, Sucu S, Yücel Yetişkin FD. Evaluating the reliability and clinical utility of artificial intelligence in first trimester prenatal screening and noninvasive prenatal testing. Sci Rep. 2025;15:41331.
15. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15:155-63.
16. Ten Hove D, Jorgensen TD, van der Ark LA. Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. Psychol Methods. 2024;29:967-79.