

Quality and Safety of ChatGPT Responses to Anesthesia-Related Patient Questions: An Expert Evaluation

Engin Çetin^{1*}, Tunahan Çevik¹

¹ Department of Anesthesiology and Reanimation, Kocaeli City Hospital, Kocaeli, Türkiye

*Correspondence: Engin Çetin, MD; doccetin52@gmail.com

Cite as: Çetin E, Çevik T. Quality and Safety of ChatGPT Responses to Anesthesia-Related Patient Questions: An Expert Evaluation. J Cukurova Anesth Surg. 2026;9(3):746-751. doi: 10.36516/jocass.1934689

Received: 20.04.2026, Accepted: 17.09.2026, Published: 30.09.2026, CC-BY-NC-4.0

Abstract

Aim: Large language models such as ChatGPT are increasingly used by patients to obtain perioperative information. Given the complexity and potential risks associated with anesthetic management, the accuracy and safety of such information are of critical importance. This study aimed to evaluate the quality and potential clinical risks of ChatGPT-generated responses to anesthesia-related questions using a structured scoring system. **Materials and Methods:** A total of 100 anesthesia-related questions were evaluated, including 50 guideline-derived questions and 50 frequently asked public questions. All questions were submitted to the ChatGPT-4 model in Turkish using the premium version. Responses were independently evaluated by three anesthesiology specialists using the Global Quality Score (GQS; 1–5). In addition, each response was assessed for the presence of potentially harmful or misleading information. Comparisons were conducted between question categories. **Results:** The median Global Quality Score (GQS) differed significantly between question types, with higher scores observed for frequently asked questions than for guideline-based questions (4 [IQR: 3–5] vs. 3 [IQR: 2–4], $p < 0.001$). Potentially harmful or misleading information was identified in 7% of responses and was more frequent in guideline-based questions than in frequently asked questions (12% vs. 2%, $p = 0.02$). **Conclusion:** ChatGPT achieved higher quality scores for frequently asked anesthesia-related questions than for guideline-based questions. Potentially harmful or misleading information was identified in 7% of responses and occurred more frequently in guideline-based questions. Although ChatGPT may support patient education, clinician oversight remains essential when discussing guideline-dependent or clinically critical topics.

Keywords: ChatGPT; anesthesiology; artificial intelligence; patient education; global quality score

1. Introduction

The widespread availability of internet-based medical information has profoundly changed how patients prepare for surgical and anesthetic procedures¹. Due to its technical complexity and perceived risks, anesthesia is among the most frequently searched medical topics online². Previous studies have shown that internet-based anesthesia information

is often incomplete, lacks adequate risk disclosure, and may contain misleading statements, particularly regarding complications and contraindications^{2,3}.

Recent advances in artificial intelligence have led to the development of large language models (LLMs) such as ChatGPT, which are capable of generating fluent, context-aware medical explanations⁴. Several studies across different medical specialties have reported that ChatGPT performs well in addressing general patient questions but demonstrates reduced accuracy when responding to guideline-based or highly technical inquiries⁵.

In urology, Ozgor et al. reported that ChatGPT achieved high Global Quality Score (GQS) values for frequently asked questions, whereas responses to European Association of Urology guideline-derived questions were significantly poorer⁶. Similar findings have been reported in hepatology, spine surgery, and perioperative medicine, raising concerns about the reliability of AI-generated medical advice in guideline-dependent scenarios⁷.

Despite the growing interest in artificial intelligence applications in medicine, systematic evaluations of ChatGPT responses within anesthesiology remain limited⁸. Given the potential consequences of inaccurate anesthesia-related information, such evaluation is particularly important. This study aims to assess the quality and potential risks of ChatGPT responses to anesthesia-related questions commonly encountered by patients and derived from international guidelines.

2. Materials and Methods

2.1. Study Design

This was an observational, cross-sectional, *in silico* study evaluating AI-generated responses. No patient data were involved.

2.2. Question Selection

A total of 100 questions were included. Fifty questions were developed from patient-relevant recommendations contained in current guidelines published by the American Society of Anesthesiologists (ASA), American Society of Regional Anesthesia and Pain Medicine (ASRA), European Society of Regional Anaesthesia and Pain Therapy (ESRA), European Society of Anaesthesiology and Intensive Care (ESAIC), and Enhanced Recovery After Surgery (ERAS) Society. Guideline recommendations were reformulated into patient-oriented questions while preserving their original clinical content.

The remaining 50 questions consisted of frequently asked questions (FAQs) identified through review of patient information pages from academic hospitals, publicly available anesthesia education websites, and common Google search queries related to anesthesia.

The final question set consisted of 100 questions equally distributed across five anesthesia-related content domains: general anesthesia, regional anesthesia, sedation, preoperative evaluation, and postoperative care (20 questions per category).

2.3. ChatGPT Query Procedure

All questions were submitted between 15 and 18 February 2026 using ChatGPT-4 (OpenAI, San Francisco, CA, USA) accessed through the premium subscription version. Each question was entered in a separate chat session to avoid contextual influence from previous conversations. The following standardized prompt was used for all queries: "Please answer the following question as if explaining it to a patient. Use clear, understandable, and balanced language. Avoid technical terminology whenever possible and provide a concise but informative explanation." No additional prompts, follow-up questions, or clarifications were provided.

2.4. Evaluation Process

Global Quality Score (GQS)

Each response was evaluated using the validated Global Quality Score (GQS)⁶, a five-point scale assessing the overall quality, accuracy, completeness, balance, and usefulness of health-related information for patients. The GQS has been widely used in previous studies evaluating online health information.

Risk Assessment

In addition to quality scoring, responses were assessed for the presence of potentially harmful or misleading information. Such content was defined as statements that could lead to unsafe clinical interpretations, including inaccurate fasting recommendations, inappropriate reassurance regarding neuraxial anesthesia in patients receiving anticoagulation therapy, or oversimplified explanations of anesthesia-related risks.

Before evaluation, the three reviewers agreed on predefined criteria for classification of potentially harmful or misleading information. Responses were considered potentially harmful when they contained inaccurate fasting recommendations, inappropriate advice regarding anticoagulant management and neuraxial anesthesia, misleading perioperative instructions, incorrect risk communication, or omissions that could reasonably affect patient safety.

When disagreement occurred regarding risk classification, consensus was reached through discussion among the reviewers. Representative examples of responses categorized as potentially harmful or misleading are presented in Supplementary Table 2.

Three anesthesiology specialists independently evaluated all responses. In cases of disagreement, the median score was used for analysis. Inter-rater reliability of GQS assessments was evaluated using the intraclass correlation coefficient (ICC). A two-way random-effects model with absolute agreement definition (ICC [2,1]) was used to determine agreement between raters.

2.5. Ethics Approval

This study did not involve human participants or patient data. The study protocol was reviewed by the Scientific Research Ethics Committee of Kocaeli City Hospital, which determined that formal ethical approval was not required (decision date: 12 February 2026; protocol number: 2026-4).

2.6. Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics for Windows, Version 27.0 (IBM Corp., Armonk, NY, USA). As the Global Quality Score (GQS) represents ordinal data, results were expressed as median (interquartile range [IQR]). Categorical variables were presented as number and percentage (n, %).

Between-group comparisons (guideline-based vs. FAQ sources) were performed using the Mann–Whitney U test for GQS scores and the Chi-square or Fisher's exact test for categorical variables, including the proportion of high-quality responses (GQS ≥ 4) and the presence of potentially harmful or misleading information.

Comparisons across anesthesia content categories were conducted using the Kruskal–Wallis test with Dunn–Bonferroni post hoc correction when appropriate.

Inter-rater reliability of GQS assessments was evaluated using the intraclass correlation coefficient (ICC) based on a two-way random-effects model with absolute agreement definition (ICC [2,1]).

All tests were two-tailed, and a p-value < 0.05 was considered statistically significant.

3. Results

A total of 100 anesthesia-related questions were evaluated, including 50 guideline-based questions and 50 frequently asked questions (FAQs).

ChatGPT responses differed significantly according to question type. Frequently asked questions yielded higher Global Quality Score (GQS) values than guideline-based questions, with median scores of 4 (IQR: 3–5) and 3 (IQR: 2–4), respectively ($p < 0.001$). The proportion of high-quality responses (GQS ≥ 4) was also greater in the FAQ group (76% vs. 42%) (Table 1).

Table 1. Comparison of Global Quality Scores (GQS) Between Guideline-Based and FAQ Sources

Variable	GUIDELINES	FAQ	p-value
Median GQS (IQR)	3 (2–4)	4 (3–5)	<0.001
GQS ≥ 4 , n (%)	21 (42%)	38 (76%)	<0.001
Min–max	1–5	2–5	-

Data are presented as median (interquartile range) or number (percentage). Comparisons were performed using the Mann–Whitney U test for continuous variables and Chi-square test for categorical variables. A p-value <0.05 was considered statistically significant.

Potentially harmful or misleading information was identified in 7% of responses. Such content was significantly more frequent in guideline-based questions than in frequently asked questions (12% vs. 2%, $p = 0.02$) (Table 2).

The most common causes of risk classification were inaccurate fasting recommendations, inappropriate advice regarding anticoagulant management in neuraxial anesthesia, and incomplete explanations of postoperative complications.

Table 2. Distribution of Potentially Harmful or Misleading Information According to Source Type

Variable	GUIDELINES	FAQ	p-value*
Risk Present, n (%)	6 (12%)	1 (2%)	0.02
Risk Absent, n (%)	44 (88%)	49 (98%)	-

Data are presented as number (percentage). Comparisons were performed using Chi-square or Fisher's exact test when appropriate. Risk was defined as the presence of potentially harmful or clinically misleading information. A p-value <0.05 was considered statistically significant.

Quality scores also differed significantly across anesthesia-related content categories (Table 3). The highest proportion of high-quality responses was observed in general anesthesia questions (85%), followed by sedation-related questions (75%). Lower scores were observed in regional anesthesia (60%) and preoperative evaluation topics (50%). Overall comparisons between categories showed a statistically significant difference ($p = 0.004$).

Table 3. Comparison of GQS Scores Across Anesthesia Content Categories

Content Category	n	Median GQS (IQR)	GQS ≥ 4 , n (%)	p-value
General Anesthesia	20	4 (3–5)	17 (85%)	0.004
Regional Anesthesia	20	3 (2–4)	12 (60%)	
Sedation / MAC	20	4 (3–5)	15 (75%)	
Preoperative Evaluation	20	3 (2–4)	10 (50%)	
Postoperative Care	20	4 (3–5)	14 (70%)	

Data are presented as median (interquartile range) or number (percentage). Overall comparisons were performed using the Kruskal–Wallis test. Post-hoc pairwise comparisons were conducted using Dunn–Bonferroni correction when appropriate. *p-value represents overall comparison between categories.

Inter-rater reliability analysis demonstrated good agreement among the three evaluators (ICC = 0.88, 95% CI: 0.81–0.93) (Table 4).

Table 4. Inter-Rater Reliability Analysis for Global Quality Scores

Measure	ICC (2,1)	95% Confidence Interval	Interpretation
Global Quality Score	0.88	0.81–0.93	Good agreement

Intraclass correlation coefficient (ICC) was calculated using a two-way random-effects model with absolute agreement definition. ICC values were interpreted as follows: <0.5 poor, 0.5–0.75 moderate, 0.75–0.9 good, >0.9 excellent reliability.

4. Discussion

This study suggests that ChatGPT provides high-quality responses to anesthesia-related questions commonly asked by patients, with overall favorable Global Quality Score values. These findings are consistent with previous studies in other medical fields, where ChatGPT has demonstrated strong performance in addressing general informational content⁹.

However, a significant decline in response quality was observed for guideline-derived questions⁸. This finding is consistent with reports by Ozgor et al. in uro-oncology and Yeo et al. in hepatology, where AI models struggled to accurately translate complex guideline recommendations into patient-appropriate language^{6,7}. Anesthesia guidelines often involve nuanced risk stratification and patient-specific decision-making, which may explain this limitation¹⁰.

The detection of potentially harmful information in 7% of responses, although relatively low, is clinically relevant⁸. Similar rates have been reported in studies evaluating ChatGPT recommendations in spine surgery and perioperative thromboprophylaxis^{9,11}. Even isolated inaccuracies in anesthesia-related advice—such as fasting recommendations or anticoagulant management—may raise clinically relevant patient safety concerns.

Among the seven responses classified as potentially harmful, inaccurate fasting recommendations, inappropriate advice regarding anticoagulant management in neuraxial anesthesia, and incomplete explanations of postoperative complications were the most common causes of risk classification¹². Such inaccuracies may lead to patient misunderstanding and could potentially compromise perioperative safety if interpreted without professional medical guidance⁸.

Previous research evaluating traditional internet-based anesthesia information using DISCERN and JAMA benchmark criteria consistently reported poor quality and inadequate risk disclosure². Compared with these findings, ChatGPT appears to provide more structured and comprehensible information. Nevertheless, the absence of professional accountability and real-time guideline updates remains a significant limitation¹³.

Overall, ChatGPT may serve as a useful adjunct for general patient education¹⁴; however, it should not replace clinician-guided counseling or evidence-based guidelines. Given the increasing use of artificial intelligence tools by patients seeking perioperative information, clinicians should be aware of both the potential benefits and limitations of AI-generated medical content. Collaboration between AI developers and anesthesiology societies may improve the safety and reliability of such tools in the future.

4.1. Limitations

This study evaluated only a single version of ChatGPT, and responses may vary with future updates. The analysis was limited to Turkish-language queries, which may affect generalizability. Additionally, GQS scoring is inherently subjective, despite the involvement of multiple expert evaluators. Although ChatGPT performed well in general patient-oriented questions, its reduced accuracy in guideline-based and clinically critical topics limits its reliability as an independent source of medical advice.

Furthermore, responses generated by large language models may vary depending on prompt phrasing and session context. As artificial intelligence models are continuously updated, future versions of ChatGPT may yield different responses.

5. Conclusion

ChatGPT achieved higher quality scores for frequently asked anesthesia-related questions than for guideline-based questions (median GQS: 4 vs. 3). Potentially harmful or misleading information was identified in 7% of responses and occurred more frequently in guideline-based questions. Although ChatGPT may serve as a useful adjunct for patient education, clinician oversight remains essential when discussing guideline-dependent or clinically critical topics.

Declarations: The authors declare that there is no conflict of interest and no financial support, grants, or donations were received for this study. No technical or language assistance was received.

References

1. Miniksar OH, Yuksek A. Does The Internet Provide Enough Information About Sepsis for The General Public. *Cukurova Anestezi ve Cerrahi Bilimler Dergisi* 2021;4:173–81. <https://doi.org/10.36516/jocass.2021.86>.
2. Yuksek A, Talih G. Quality of internet-based information on obstetric anesthesia. *Ann Med Res* 2019;26:2904. <https://doi.org/10.5455/annalsmedres.2019.08.493>.
3. Sam Ozdemir M, Ozdemir H. Is YouTube really a source of information on MRI guided prostate biopsy? *Bozok Tip Dergisi* 2024. <https://doi.org/10.16919/bozoktip.1514887>.
4. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med* 2016;375:1216–9. <https://doi.org/10.1056/NEJMp1606181>.
5. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on Medical Questions in the National Medical Licensing Examination in Japan: Evaluation Study. *JMIR Form Res* 2023;7:e48023. <https://doi.org/10.2196/48023>.
6. Ozgor F, Caglar U, Halis A, Cakir H, Aksu UC, Ayranci A, et al. Urological Cancers and ChatGPT: Assessing the Quality of Information and Possible Risks for Patients. *Clin Genitourin Cancer* 2024;22:454-457.e4. <https://doi.org/10.1016/j.clgc.2023.12.017>.
7. Yeo YH, Samaan JS, Ng WH, Ting P-S, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol* 2023;29:721–32. <https://doi.org/10.3350/cmh.2023.0089>.
8. Ortac M, Ergul RB, Yazili HB, Ozervarli MF, Tonyali S, Sarilar O, et al. ChatGPT's competence in responding to urological emergencies. *Ulus Travma Acil Cerrahi Derg* 2025;31:291–5. <https://doi.org/10.14744/tjtes.2024.03377>.
9. Nagesh D, Keating DP, Divakaruni RV, Beutel BG. Feasibility evaluation of large language models in anesthesia-specific post-operative care instructions for total knee arthroplasty. *PEC Innovation* 2026;8:100444. <https://doi.org/10.1016/j.pecinn.2025.100444>.
10. Tulgar S, Aksu C, Selvi O, Sultan P, Dogan AT, Thomas DT, et al. A comparative evaluation of the quality of responses provided by different large language model chatbots to frequently asked questions regarding nerve blocks. *BMC Anesthesiol* 2026;26:98. <https://doi.org/10.1186/s12871-025-03596-9>.
11. Olcar HA, Aydin AG, Uluturk S, Aras DS, Korkmaz M, Korkmaz IN. Evaluation of ChatGPT-4 Artificial Intelligence Responses to Common Problems in Patients after Hip Replacement with Expert Opinion. *Bozok Tip Dergisi* 2023. <https://doi.org/10.16919/bozoktip.1318421>.
12. Balci MF, Akdemir C, Yildirim F. Is Artificial Intelligence-Assisted Pregnancy Counseling Feasible? An Evaluation of the Quality of ChatGPT Responses. *Cukurova Anestezi ve Cerrahi Bilimler Dergisi* 2025;8:336–40. <https://doi.org/10.36516/jocass.1753846>.
13. Alber DA, Yang Z, Alyakin A, Yang E, Rai S, Valliani AA, et al. Medical large language models are vulnerable to data-poisoning attacks. *Nat Med* 2025;31:618–26. <https://doi.org/10.1038/s41591-024-03445-1>.
14. Alanoglu Z, Tulgar S, Kilicaslan A, Can OS. Artificial Intelligence and Large Language Models: Editorial Reflections. *Turk J Anaesthesiol Reanim* 2025;53:280–1. <https://doi.org/10.4274/TJAR.2025.252355>.