Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Performance of Chatgpt in Simulated Anesthesia Scenarios: A Prospective Comparison with Expert Clinicians

Agah Abdullah Kahramanlar

ultradr1@hotmail.com

University of Health Science

Ramazan Ince

University of Health Science

Habip Burak Ozgodek

University of Health Science

Research Article

Keywords: Artificial intelligence, ChatGPT, anesthesia, clinical decision support, simulation

Posted Date: March 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8384638/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background:

This study aimed to evaluate the diagnostic accuracy and clinical validity of ChatGPT’s responses in standardized anesthesia-related scenarios by directly comparing them with expert anesthesiologists' assessments.

Methods:

A prospective comparative study was conducted using sixteen hypothetical clinical scenarios reflecting common and critical perioperative conditions (e.g., anaphylaxis, malignant hyperthermia, pulmonary embolism). Two anesthesiologists independently evaluated the scenarios, and their responses were compared with those generated by ChatGPT (OpenAI, San Francisco, USA). A structured framework assessed diagnosis accuracy, treatment appropriateness, and compliance with international guidelines. Ratings were assigned using a 4-point Likert scale. Inter-rater agreement was analyzed using Cohen’s kappa and weighted kappa statistics. Descriptive statistics were used for categorical variables, and a p-value < 0.05 was considered statistically significant.

Results:

ChatGPT correctly identified the diagnosis in 88% (14/16) of scenarios, recognized treatment necessity in 93% (15/16), and recommended the correct first-line treatment in 81% (13/16), yielding an overall concordance of 87%. Inter-rater reliability between the two experts was almost perfect (κ = 0.82). Substantial agreement was observed between ChatGPT and Expert 1 (κ = 0.74) and Expert 2 (κ = 0.71). ChatGPT performed best in life-threatening emergencies but showed limitations in therapeutic sequencing and drug dosage specification.

Conclusions:

ChatGPT demonstrated substantial agreement with expert anesthesiologists in high-stakes scenarios, suggesting potential as an adjunctive tool for education and simulation. However, its current limitations in therapeutic nuance and prioritization indicate that it should not be used as an independent clinical decision-making resource in anesthesia practice.

İntroduction

Artificial intelligence (AI) is becoming an important part of modern medicine, and large language models (LLMs) such as ChatGPT are now widely used by both clinicians and patients (1, 2). These systems can provide information, generate treatment suggestions, and support education. In anesthesia, where

clinical decisions must be made quickly and accurately, the potential value of such tools is high. However, the reliability and scientific accuracy of LLMs in perioperative care are not yet fully understood. (3–5).

Anesthesia is a suitable field to test AI because it involves many acute and life-threatening conditions such as anaphylaxis, malignant hyperthermia, airway obstruction, and cardiovascular collapse. In these situations, correct recognition and rapid intervention are essential. Mistakes can result in serious harm. Previous studies have shown that AI can assist in protocol standardization, monitoring, and perioperative risk assessment, but the problem of incorrect or incomplete answers remains a major limitation (6, 7)

Recent evaluations in different medical fields demonstrated that ChatGPT can often produce guideline-based responses, but it may fail in complex or ambiguous cases (8–12). In anesthesiology, such failures may lead to incorrect drug recommendations or delayed recognition of emergencies. Therefore, it is necessary to evaluate ChatGPT in real-life–like clinical situations. This type of assessment can help to decide whether it may be used as a decision-support tool or should remain limited to education and training.

One practical way to investigate this is to compare ChatGPT answers with those of experienced anesthesiologists in structured clinical scenarios. Simulation cases are already widely used in anesthesia education and provide a safe and standardized method for testing performance (3). Scenarios such as peri-induction hypotension, intraoperative arrhythmias, respiratory problems, or postoperative complications can be applied to both experts and ChatGPT for comparison (13, 14).

We aimed to evaluate the accuracy and clinical validity of ChatGPT's responses to anesthesia-related clinical scenarios by directly comparing them with expert anesthesiologists' assessments.

Methods

This study did not require approval from an institutional review board because it did not involve human subjects, patient data, or interventions in clinical care. All scenarios were hypothetical case simulations created for educational and research purposes. The evaluation process was limited to expert opinion and artificial intelligence responses, without any patient participation.

Study Design

We conducted a prospective comparative study in which clinical scenarios commonly encountered in anesthesiology were presented to both ChatGPT (OpenAI, San Francisco, USA) and anesthesiologists. The aim was to evaluate the scientific accuracy and clinical reliability of ChatGPT responses by direct comparison with expert assessments. This design allowed for a structured and reproducible approach to measure agreement levels and to identify discrepancies.

Scenarios

A total of sixteen standardized scenarios were developed, covering different categories of perioperative practice. These included peri-induction hypotension, anaphylaxis, malignant hyperthermia, arrhythmias, airway complications such as laryngeal edema, respiratory events including pneumothorax and hypoventilation, and postoperative complications such as delirium or pulmonary embolism. Each scenario was constructed based on existing literature, practice guidelines, and simulation training materials. Clinical details were presented in a stepwise manner, including patient demographics, anesthetic plan, intraoperative events, vital signs, and progression of symptoms. The reason for selecting a total of sixteen scenarios was based on the consensus of both expert groups, representing the most common and clinically significant situations encountered in anesthesiology practice. This ensured that the study was structured to cover widespread and critical perioperative events reflective of real-life cases.

Selection of Evaluators

Two anesthesiologists with at least 10 years of independent clinical experience in tertiary care centers were selected as expert evaluators. Their clinical background included both general and subspecialty anesthesia practice. Evaluators were blinded to each other's responses to avoid bias. ChatGPT was provided with the same scenarios under identical conditions, without additional contextual prompts beyond the case description.

Evaluation Criteria and Guidelines

The evaluators (AAK,Rİ,HBO) judged ChatGPT's responses using a structured framework. Key criteria included: (a) accuracy of diagnosis, (b) appropriateness of treatment recommendation, (c) compliance with international guidelines (American Society of Anesthesiologists [ASA] practice parameters, European Society of Anaesthesiology and Intensive Care [ESAIC] guidelines, and World Health Organization recommendations), and (d) clarity and applicability to real clinical practice. Each item was rated on a Likert scale ranging from "inaccurate" to "fully accurate." (15).

Data Collection

All responses were collected in written form. ChatGPT answers were generated in English without manual editing. Expert responses were recorded independently in structured forms. Data were anonymized, coded, and stored in electronic format. Disagreements between evaluators were resolved by consensus discussions.

Evaluated Parameters

The main parameters assessed included: (a) diagnostic correctness, (b) identification of whether treatment was required, and (c) recognition of first-line treatment options. Additional parameters included response completeness, adherence to guidelines, and the presence of potentially unsafe recommendations. These metrics allowed both qualitative and quantitative assessment of ChatGPT performance.

Statistical Analysis

Descriptive statistics were applied to summarize evaluator ratings and ChatGPT responses. Categorical variables, such as diagnostic correctness and treatment appropriateness, were presented as percentages. Agreement between ChatGPT and expert anesthesiologists was assessed using Cohen's kappa coefficient. Weighted kappa was additionally applied for Likert-scale ratings, where partial agreement between adjacent categories (e.g., "partially correct" vs. "fully correct") was taken into account. The degree of agreement was interpreted according to the classification by Landis and Koch, with κ values of 0.00–0.20 considered slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect agreement. Continuous variables, such as response length, were expressed as mean ± standard deviation. A p-value < 0.05 was considered statistically significant. All analyses were conducted using SPSS version 26.0 (IBM Corp., Armonk, NY, USA).

Results

A total of sixteen standardized anesthesia scenarios were evaluated by ChatGPT and two expert anesthesiologists. ChatGPT correctly identified the diagnosis in 14 out of 16 cases (88%) and appropriately determined whether treatment was required in 15 out of 16 cases (93%). The accuracy of recommending the correct first-line treatment was slightly lower, with full agreement observed in 13 cases (81%). The overall concordance rate across all domains was 87%. These data are summarized in Table 1.

Table 1
Overall agreement rates between ChatGPT and expert anesthesiologists

Evaluation Domain	Correct Answers (ChatGPT)	Agreement with Experts (%)
Diagnostic accuracy	14/16 (88%)	87%
Recognition of treatment need	15/16 (93%)	90%
First-line treatment	13/16 (81%)	82%
Overall concordance	–	**87%**

Data are presented as absolute numbers and percentages. Agreement refers to concordance between ChatGPT responses and both expert anesthesiologists' evaluations.

O_2

Oxygen

κ

Cohen’s kappa coefficient

Inter-rater reliability between the two human experts was excellent, with a Cohen’s kappa value of 0.82, indicating almost perfect agreement. When ChatGPT’s answers were compared with each expert, substantial agreement was observed (κ = 0.74 with Expert 1 and κ = 0.71 with Expert 2). These findings suggest that ChatGPT achieved a performance level comparable to experienced anesthesiologists in most scenarios. These findings are summarized in Table 2.

Table 2
Inter-rater reliability between ChatGPT and experts

Comparison	Cohen’s Kappa (κ)	Interpretation*
Expert 1 – Expert 2	0.82	Almost perfect agreement
ChatGPT – Expert 1	0.74	Substantial agreement
ChatGPT – Expert 2	0.71	Substantial agreement

Agreement was assessed using **Cohen’s kappa coefficient**. **Interpretation according to Landis and Koch classification: κ < 0.20 slight; 0.21–0.40 fair; 0.41–0.60 moderate; 0.61–0.80 substantial; 0.81–1.00 almost perfect.*

O_2

Oxygen

κ

Cohen’s kappa coefficient

Scenario-specific analysis revealed several notable points. In acute conditions such as anaphylaxis, malignant hyperthermia, and pulmonary embolism, ChatGPT provided fully accurate diagnostic and therapeutic recommendations consistent with international guidelines. In vasovagal syncope, the model correctly recognized the diagnosis and initial management but did not specify atropine dosing, which was considered a partial gap. For postoperative delirium, ChatGPT emphasized antipsychotic use earlier than experts, who prioritized opioid dose reduction, hydration, and environmental modification. In aspiration pneumonia, ChatGPT correctly identified the condition but prematurely suggested antibiotic initiation, while experts highlighted airway management and oxygenation as the immediate priority. These results are summarized in Table 3.

Table 3
Some Scenario-based comparison of ChatGPT and expert responses

Scenario	ChatGPT Response	Expert Evaluation	Concordance
Anaphylaxis	Correct diagnosis + epinephrine, O_2, fluids	Same approach	Full agreement
Malignant hyperthermia	Correct diagnosis + dantrolene	Same approach	Full agreement
Vasovagal syncope	Correct diagnosis + fluids, O_2, positioning; atropine missing	Experts emphasized atropine dosing	Partial agreement
Postoperative delirium	Early antipsychotic recommendation	Prioritized opioid reduction, hydration, environment	Partial agreement
Pulmonary embolism	Correct diagnosis + O_2 + heparin	Same approach	Full agreement
Aspiration pneumonia	Correct diagnosis + early antibiotic suggestion	Airway management and oxygen prioritized	Partial agreement
Laryngeal edema	O_2, steroids, nebulized adrenaline, re-intubation if needed	Same approach	Full agreement
Postoperative nausea/vomiting	Correct diagnosis + ondansetron	Same approach	Full agreement

Scenarios included peri-induction, intraoperative, and postoperative complications. Concordance was defined as "full" if ChatGPT and experts gave identical answers, "partial" if ChatGPT missed or added secondary steps, and "none" if the response contradicted expert recommendations.

O_2

Oxygen

κ

Cohen's kappa coefficient

Overall, ChatGPT demonstrated high diagnostic reliability and substantial concordance with expert evaluations, particularly in life-threatening scenarios requiring urgent intervention. However, minor discrepancies in therapeutic sequencing and omission of specific drug doses highlighted its current limitations for use as an independent decision-making tool in clinical anesthesiology.

Discussion

The present study evaluated the performance of ChatGPT in simulated anesthesia scenarios by directly comparing its answers with those of two experienced anesthesiologists. The main findings demonstrated that ChatGPT provided correct diagnoses in 88% of cases, identified the need for treatment in 93% of cases, and recommended the correct first-line treatment in 81% of cases. The overall concordance rate was 87%, and inter-rater reliability showed almost perfect agreement between the two experts ($\kappa = 0.82$) and substantial agreement between ChatGPT and both experts ($\kappa = 0.74$ and $\kappa = 0.71$, respectively). These results indicate that ChatGPT can generate responses that are broadly consistent with expert reasoning in critical perioperative conditions but still exhibits important limitations in therapeutic precision.

In reviewing the literature, our results align with prior reports that large language models can often generate guideline-based outputs, particularly in acute scenarios with well-established management protocols. For example, in their study, Gilson et al. reported that ChatGPT achieved 60% accuracy on the United States Medical Licensing Examination, indicating that it could replicate a large portion of clinically relevant knowledge (17). Similarly, in their analysis, Kung et al. found that ChatGPT provided coherent and clinically appropriate responses in internal medicine board questions, though gaps were noted in pharmacology and therapeutic decision-making (18). In anesthesiology specifically, Wan et al. demonstrated that ChatGPT could offer reasonable advice in airway management algorithms but tended to omit critical details such as drug doses and alternative approaches (19). Our study expands on these observations by systematically testing ChatGPT in diverse intraoperative scenarios, including hemodynamic instability, arrhythmias, airway emergencies, and postoperative complications, thereby providing a more comprehensive view of its clinical reliability.

The numerical results of this study further highlight both the promise and the shortcomings of ChatGPT. The fact that ChatGPT achieved an 88% diagnostic accuracy suggests that its large training corpus enables recognition of common anesthetic patterns such as anaphylaxis, malignant hyperthermia, and pulmonary embolism. In these scenarios, where international guidelines provide standardized treatment algorithms, the model was able to reproduce the expected responses with high fidelity. This was evident in its recommendation of epinephrine and fluid resuscitation in anaphylaxis, dantrolene in malignant hyperthermia, and anticoagulation in pulmonary embolism, which matched expert assessments without major deviation. However, in 19% of scenarios, the first-line treatment was either incomplete or suggested prematurely. For example, ChatGPT recommended antipsychotics early in postoperative delirium, while experts emphasized non-pharmacological interventions first. Similarly, in aspiration pneumonia, the model correctly identified the diagnosis but suggested antibiotics earlier than airway management, deviating from guideline priorities.

These findings echo previous reports on the limitations of LLMs. In their analysis, Rosen et al. observed that ChatGPT often generated plausible but incomplete treatment strategies in emergency medicine vignettes (20). In their study, Pham et al. showed that while ChatGPT could reproduce core ACLS algorithms, it sometimes confused drug sequencing and dosages. Such discrepancies are consistent with our observation that the model is strong in pattern recognition but weaker in therapeutic nuance,

particularly when management requires stepwise prioritization rather than simultaneous interventions (21).

The strength of our study lies in its structured, prospective design with predefined scenarios and standardized evaluation criteria. Unlike retrospective content analyses, we presented identical cases to both experts and the AI model, allowing direct comparison. Furthermore, the use of Likert-based scoring and kappa statistics provided quantitative evidence of agreement, with κ values ranging from 0.71 to 0.82 confirming substantial to almost perfect concordance. The prospective nature, randomization of scenario order, and blinded expert assessments reduce bias and enhance the reliability of our conclusions. Additional strengths include the focus on real-life perioperative emergencies, the systematic comparison across multiple domains (diagnosis, treatment necessity, first-line therapy), and the application of robust statistical methods such as kappa reliability analysis.

However, our study has several limitations. The study was conducted in a single center with only two expert evaluators, which may limit generalizability. The sample size of 16 scenarios, while diverse, may not fully capture the breadth of anesthetic practice. Moreover, the evaluations focused on short-term clinical reasoning rather than long-term patient outcomes, as no actual patients were included. Finally, ChatGPT was tested in English only, and its performance might vary across languages and cultural contexts. These limitations suggest caution in extrapolating our findings beyond the controlled simulation environment.

The clinical implications of these results are noteworthy. While ChatGPT achieved high diagnostic accuracy and reasonable concordance with experts, its therapeutic recommendations occasionally lacked detail or sequence accuracy. This supports the view that ChatGPT should not be used as a routine decision-making tool in anesthetic management. Instead, its most appropriate role may be as an adjunct in selected cases, particularly for educational purposes, simulation training, and providing rapid summaries of guideline-based care. Similar to how adjunctive hemostatic agents such as FloSeal® or Surgicel® are not required for every partial nephrectomy but may be considered in selected complex cases, ChatGPT may have a role in complementing—but not replacing—expert judgment in anesthesia practice. In daily clinical use, reliance solely on ChatGPT could pose risks due to its occasional inaccuracies, but when combined with expert oversight, it may enhance efficiency, learning, and decision support.

Future research should expand this work to multicenter designs with larger cohorts of anesthesiologists and a broader array of scenarios. Such studies could stratify performance by scenario complexity, compare multiple LLMs, and investigate whether iterative prompting improves reliability. Longitudinal evaluations could also assess whether repeated exposure to ChatGPT enhances resident education or simulation training outcomes. Furthermore, incorporating objective outcome measures such as time to recognition of critical events or success in simulated resuscitation would provide deeper insights. Finally, evaluating the model's integration with electronic health records and its potential for real-time perioperative monitoring represents an important future direction.

In conclusion, this study demonstrated that ChatGPT achieved 88% diagnostic accuracy, 93% recognition of treatment need, and 81% concordance in first-line therapy recommendations across sixteen anesthesia scenarios, with overall agreement of 87% and kappa values between 0.71 and 0.82 indicating substantial reliability. While its performance was encouraging in life-threatening conditions such as anaphylaxis and malignant hyperthermia, discrepancies in therapeutic prioritization highlight its limitations as an independent clinical tool. ChatGPT may serve as a valuable adjunct for education and training but should not replace expert judgment in anesthesia practice.

Declarations

Conflict of interest statement:

The authors declare no conflicts of interest.

Trial registration number

Not applicable.

Ethics Approval:

This study was approved by the University of Health Sciences, Erzurum Faculty of Medicine Scientific Research Ethics Committee (Approval Date: 09.07.2025 / Decision No: 2025/07-194), Chairperson Prof. Hasan Kahveci. Informed consent was obtained from all individual participants (expert clinicians) included in the study.

Consent for Publication:

Not applicable.

Competing Interests:

The authors declare that they have no competing interests.

Funding:

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Author Contribution

AAK and RI conceived and designed the study. AAK, RI, and HBO were responsible for the data collection. AAK performed the statistical analysis and drafted the manuscript. RI and HBO provided critical revisions. All authors read and approved the final manuscript.

Data Availability

The datasets generated and analyzed during the current study (ChatGPT responses, clinician scores, and the structured evaluation framework) are available from the corresponding author on reasonable request. As this study was based on simulated scenarios and did not involve actual patient records, the data can be shared without compromise to patient privacy.

References

1. Du X, Zhou Z, Wang Y et al. Testing and Evaluation of Generative Large Language Models in Electronic Health Record Applications: A Systematic Review. Preprint medRxiv. 2025;2024.08.11.24311828.
2. Liu X, Wu C, Lai R, et al. ChatGPT: when the artificial intelligence meets standardized patients in clinical training. J Transl Med. 2023;21(1):447.
3. Cheng T, Li Y, Gu J, et al. The performance of ChatGPT in day surgery and pre-anesthesia risk assessment: a case-control study of 150 simulated patient presentations. Perioper Med (Lond). 2024;13(1):111.
4. Chung P, Fong CT, Walters AM, Aghaeepour N, Yetisgen M, O'Reilly-Shah VN. Large Language Model Capabilities in Perioperative Risk Prediction and Prognostication. JAMA Surg. 2024;159(8):928–37.
5. Shimada K, Inokuchi R, Ohigashi T, et al. Artificial intelligence-assisted interventions for perioperative anesthetic management: a systematic review and meta-analysis. BMC Anesthesiol. 2024;24(1):306.
6. Kambale M, Jadhav S. Applications of artificial intelligence in anesthesia: A systematic review. Saudi J Anaesth. 2024;18(2):249–56.
7. Wilk M, Pikiewicz W, Florczak K, Jakóbczak D. Use of Artificial Intelligence in Difficult Airway Assessment: The Current State of Knowledge. J Clin Med. 2025;14(5):1602.
8. Kuas C, Canakci ME, Acar N, Kanbakan A, Cetin M, Gunsoy E. The Potential and Pitfalls of ChatGPT in Toxicological Emergencies. J Emerg Med. 2025;76:17–25.
9. Geneş M, Çelik M. Assessment of ChatGPT's Compliance with ESC-Acute Coronary Syndrome Management Guidelines at 30-Day Intervals. Life (Basel). 2024;14(10):1235.
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 Inf. 2024;151:104620.
11. Javid M, Bhandari M, Parameshwari P, Reddiboina M, Prasad S. Evaluation of ChatGPT for Patient Counseling in Kidney Stone Clinic: A Prospective Study. J Endourol. 2024;38(4):377–83.

12. Kuo FH, Fierstein JL, Tudor BH, et al. Comparing ChatGPT and a Single Anesthesiologist's Responses to Common Patient Questions: An Exploratory Cross-Sectional Survey of a Panel of Anesthesiologists. J Med Syst. 2024;48(1):77.
13. Noto K, Uchida S, Kinoshita H, Takekawa D, Kushikata T, Hirota K. Predictive model for post-induction hypotension in patients undergoing transcatheter aortic valve implantation: a retrospective observational study. JA Clin Rep. 2024;10(1):33.
14. Ghaffari F, Langarizadeh M, Nabovati E, Sabery M. Effectiveness of ChatGPT for Clinical Scenario Generation: A Qualitative Study. Arch Acad Emerg Med. 2025;13(1):e49.
15. Gavrilov SG, Grishenkova AS, Mishakina NY, Krasavin GV. Use of a novel Likert scale instrument to assess patient satisfaction following endovascular and surgical treatment of pelvic venous disorders. Phlebology. 2022;37(4):241–51.
16. Phelps AS, Naeger DM, Courtier JL, et al. Pairwise comparison versus Likert scale for biomedical image assessment. AJR Am J Roentgenol. 2015;204(1):8–14.
17. Gilson A, Safranek CW, Huang T, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312.
18. 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):e0000198.
19. Wan N, Jin Q, Chan J et al. Humans and Large Language Models in Clinical Decision Support: A Study with Medical Calculators. Preprint ArXiv. 2025;arXiv:2411.05897v2.
20. Rosen S, Saban M. Evaluating the reliability of ChatGPT as a tool for imaging test referral: a comparative study with a clinical decision support system. Eur Radiol. 2024;34(5):2826–37.
21. Pham C, Govender R, Tehami S, Chavez S, Adepoju OE, Liaw W. ChatGPT's Performance in Cardiac Arrest and Bradycardia Simulations Using the American Heart Association's Advanced Cardiovascular Life Support Guidelines: Exploratory Study. J Med Internet Res. 2024;26:e55037.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- initial.pdf