JHWCR JOURNAL OF HEALTH, WELLNESS AND COMMUNITY RESEARCH 3007-0570

Vol. 4 Issue 6, 2026
Open Access | DOI: 10.61919/ea3vsa53

Original Article

Comparative Analysis of Anesthesiologists Judgment and ChatGPT in Predicting Intra Operative and Emergency Related Anesthesia Complications

Kifayat Ullah[1], Abdul Mabood[2], Huzaim Ahmad[3], Niha Fayyaz[1], Hamza Khan[1], Murad Ali[4], Khadija Tun Nisa[1], Mukhtiar Ahmad[5]

[1] Institute of Paramedical Sciences, Khyber Medical University, Peshawar, Pakistan
[2] Khyber Medical University Hospital and Research Center, Peshawar, Pakistan
[3] Faculty of Allied Health Sciences, Superior University, Lahore, Pakistan
[4] King's Institute of Nursing & Allied Health Sciences, Swat, Pakistan
[5] Khyber Medical University, Peshawar, Pakistan

✉ **Correspondence:** Mukhtiar Ahmad. mukhtiar.ipms@kmu.edu.pk

Received: 09 January 2026 | Revised: 15 January 2026 | Accepted: 12 March 2026 | Published: 30 March 2026

HOW TO CITE

Kifayat Ullah et al. Comparative Analysis of Anesthesiologists Judgment and ChatGPT in Predicting Intra Operative and Emergency Related Anesthesia Complications. Journal of Health, Wellness and Community Research. DOI:https://doi.org/10.61919/ea3vsa53

ARTICLE INFORMATION

Author Contributions	Concept: KU, KTN; Design: KU, AM, KTN; Data Collection: AM, NF, HK, MuA; Analysis: HA, MkA; Drafting: KU, HA, NF, HK, MuA, KTN, MkA.
Ethical Approval	The study was approved by research ethical board of Khyber Medical University, Peshawar, Pakistan
Informed Consent	Written informed consent was obtained from all participants
Conflict of Interest	The authors declare no conflict of interest.
Funding	No external funding.
Data Availability	Available from the corresponding author on reasonable request.
AI Declaration	No generative AI was used for conceptualization, analysis, or content generation. AI use was limited to language refinement, and the authors remain fully responsible for the accuracy, originality, validity, and integrity of the manuscript's content.

ABSTRACT

Background: Artificial intelligence and large language models are increasingly being investigated for perioperative decision support, but their agreement with anesthesiologist judgment in predicting anesthesia-related complications remains insufficiently characterized. Objective: To compare anesthesiologist and ChatGPT predictions of intraoperative and emergency-related anesthesia complications using the same clinical information. Methods: This observational comparative study included 140 patients undergoing surgical procedures at Lady Reading Hospital and Khyber Teaching Hospital, Peshawar. Anesthesiologists and ChatGPT independently assessed standardized anonymized clinical information. Inter-method associations were assessed using chi-square analysis, and agreement was quantified using Cohen's kappa. Results: Of 140 participants, 73 (52.1%) were aged 41–70 years, 88 (62.9%) were male, and 124 (88.6%) underwent elective surgery. Agreement was perfect for bradycardia and hypo-/hyperthermia (κ=1.000 each), and remained high for hypotension (κ=0.959), bleeding (κ=0.867), hypertension (κ=0.826), and hypoventilation/hypoxia (κ=0.793). Lower concordance was observed for shivering (κ=0.659) and tachycardia (κ=0.646). All analyzable inter-method associations had $p<0.001$. Conclusion: ChatGPT classifications showed substantial concordance with anesthesiologist judgments for several common anesthesia-related complications. These findings demonstrate inter-method agreement but do not independently establish predictive accuracy against actual clinical outcomes. Keywords: Artificial intelligence; ChatGPT; anesthesiology; intraoperative complications; clinical decision support; perioperative risk assessment; Cohen's kappa.

INTRODUCTION

Anesthesia care requires continuous assessment, rapid clinical judgment, and timely intervention because perioperative physiological disturbances can progress quickly and contribute to serious morbidity or mortality. Although advances in monitoring, pharmacology, airway management, and perioperative protocols have improved anesthetic safety, clinically important adverse events continue to occur in operating-room practice (1). Airway-related complications in particular can result in severe patient harm and illustrate the importance of anticipating deterioration before a critical event becomes established (2). Prediction of perioperative complications therefore

remains a central component of anesthetic practice, requiring integration of patient characteristics, comorbidities, surgical factors, airway assessment, laboratory findings, and physiological information. This process is inherently cognitively demanding, and clinical performance may vary according to workload, experience, competing tasks, and the complexity of the perioperative environment (3).

Artificial intelligence has increasingly been investigated as a means of supporting perioperative risk assessment and clinical decision-making. Machine-learning approaches can process multiple clinical variables and have been explored for prediction, monitoring, and perioperative management across different stages of anesthesia care (4). More specifically, deep-learning models using physiological signals have demonstrated the potential to identify or predict clinically important events such as intraoperative hypotension (5). These developments suggest that computational approaches may complement conventional clinical assessment, particularly when several patient-specific and procedural factors must be considered simultaneously. However, the performance of purpose-built machine-learning models based on structured physiological data cannot automatically be generalized to conversational artificial-intelligence systems that generate responses from natural-language clinical information.

Large language models such as ChatGPT represent a different form of artificial intelligence because they are designed to interpret and generate natural language rather than function solely as conventional statistical prediction algorithms. Recent evidence indicates growing interest in their use within anesthesiology and critical care for clinical reasoning, information synthesis, perioperative assessment, education, and decision support, although the quality and clinical maturity of these applications remain variable (6). Generative AI chatbots have consequently been proposed as potentially useful adjuncts in anesthesia practice, particularly for organizing clinical information and supporting structured assessment rather than independently directing patient care (7). Large language models have also demonstrated substantial medical knowledge and reasoning capabilities on anesthesiology-related examination material, supporting their potential usefulness in knowledge-intensive clinical tasks while not establishing their accuracy in real-world patient management (8).

Evidence directly addressing perioperative patient assessment is beginning to emerge. ChatGPT has been evaluated using simulated day-surgery and pre-anesthesia presentations, demonstrating that language models can process clinical scenarios and generate perioperative risk assessments under controlled conditions (9). Comparative investigations of AI models for preoperative anesthesia planning have similarly suggested potential value in assisting with structured clinical reasoning and planning (10). Nevertheless, performance in simulated cases or structured clinical questions does not necessarily translate into reliable prediction of complications in actual patients. Studies examining AI interpretation of clinical data in real-world contexts have also demonstrated limitations, emphasizing that apparently strong knowledge performance should not be equated with dependable clinical accuracy across all tasks (11).

This distinction is particularly important in anesthesia, where adverse events may evolve rapidly and where incorrect risk classification could influence preparation, monitoring, or intervention. The usefulness of ChatGPT therefore depends not only on whether its predictions resemble those of clinicians but also on whether those predictions correspond to actual patient outcomes. Agreement between an AI system and an anesthesiologist represents concordance between two decision-makers; it does not, by itself, establish that either prediction is correct. In addition, the use of generative AI in healthcare raises concerns regarding reliability, transparency, privacy, accountability, and appropriate professional oversight (12). Current discussions in anesthesiology accordingly emphasize that ChatGPT should be considered a potentially supportive technology rather than an autonomous substitute for specialist clinical judgment (13).

Despite increasing research on artificial intelligence in perioperative medicine, comparatively limited evidence directly compares ChatGPT-generated predictions with anesthesiologists' clinical judgments using the same patient information and subsequently relates those predictions to documented intraoperative outcomes. Existing literature has focused more commonly on conventional machine-learning models, simulated clinical scenarios, examination performance, or preoperative planning rather than patient-level prediction of specific anesthesia-related complications during routine clinical care (4,6,8–10). Determining where ChatGPT agrees with anesthesiologists, where the two approaches differ, and how their predictions correspond to actual clinical events is necessary before meaningful conclusions can be drawn regarding the role of large language models in perioperative decision support.

Therefore, the present study aimed to compare the predictions generated by anesthesiologists and ChatGPT for intraoperative and emergency-related anesthesia complications among patients undergoing surgical procedures. Using the same available clinical information for both approaches, the study sought to assess the level of agreement between anesthesiologist and ChatGPT predictions and to compare these predictions with documented intraoperative outcomes. The primary research question was whether ChatGPT could generate predictions of common anesthesia-related complications that were comparable to anesthesiologists' clinical judgments while maintaining clinically meaningful correspondence with the complications actually observed during perioperative care.

MATERIALS AND METHODS

The study used an observational comparative design to evaluate the agreement and predictive performance of anesthesiologists and ChatGPT in identifying anesthesia-related complications using the same clinical information. The study was conducted in the Departments of Anesthesiology at Lady Reading Hospital (LRH) and Khyber Teaching Hospital–Medical Teaching Institution (KTH-MTI), Peshawar, Khyber Pakhtunkhwa, Pakistan. Both institutions are tertiary-care referral hospitals providing anesthesia services for a broad range of elective and emergency surgical procedures. The overall study activities, including ethical approval, participant recruitment, data collection, data entry, statistical analysis, interpretation, and preparation of the research report, were completed over approximately four to six months.

The target population comprised of patients undergoing elective or emergency surgical procedures requiring anesthesia at the participating hospitals. Patients were considered eligible when adequate perioperative clinical information and complete intraoperative monitoring records were available for prediction and subsequent outcome assessment. Required monitoring information included electrocardiography, blood pressure measured non-invasively or invasively, peripheral oxygen saturation, end-tidal carbon dioxide, respiratory rate, and relevant anesthetic-agent information. Cases were excluded when consent was declined, monitoring records were incomplete or of insufficient quality because of substantial missing information or artifacts, or documented intraoperative outcomes were inadequate for comparison with the predictions. A non-probability convenience sampling approach with consecutive recruitment of eligible patients presenting during the study period was used until the planned sample of 140 participants was achieved. The sample size was determined using a prevalence-based approach based on a 95% confidence level, an expected prevalence of 5%, and a 5% margin of error, with a final target of 140 participants.

Following institutional approval, eligible patients were approached and informed about the purpose and procedures of the study, and verbal or written informed consent was obtained before inclusion. Data were collected using a structured study-specific proforma designed to capture demographic characteristics, clinical history, preoperative assessment findings, relevant laboratory information, American Society of Anesthesiologists physical-status classification, type of surgical procedure, type of anesthesia, perioperative monitoring parameters, anesthetic information, and documented anesthesia-related complications. Demographic and clinical variables included age, sex, height, elective or emergency surgical status, presence of comorbid conditions, laboratory-investigation status, ASA class, and anesthesia type. Analysis was restricted to cases with sufficiently complete clinical information for both prediction approaches and clearly documented intraoperative outcomes.

For each included case, the available clinical information was reviewed by consultant anesthesiologists, who recorded their predictions regarding the occurrence of intraoperative or emergency-related anesthesia complications on the basis of routine clinical assessment. The same anonymized clinical information was separately submitted to ChatGPT using a standardized prompt to generate corresponding predictions. The two prediction approaches therefore received the same available patient information, allowing paired comparison at the individual-patient level.

The complication variables evaluated in the dataset included hypotension, shivering, bradycardia, hypertension, bleeding, tachycardia, hypoventilation/hypoxia, hypo- or hyperthermia, failed block, local anesthetic systemic toxicity, embolism, aspiration, spasm, delayed recovery, aortocaval compression, anxiety, and difficult airway. Predictions generated by the anesthesiologists and ChatGPT were subsequently compared with the anesthesia-chart documentation of complications occurring during the perioperative period, which served as the clinical outcome reference for performance assessment.

Several procedural measures were used to reduce avoidable variation between the two prediction approaches. Uniform participant-selection criteria and a standardized data-collection process were applied across the participating hospitals, and the same anonymized clinical information was used for the anesthesiologist and ChatGPT assessments.

Predictions were generated separately so that the comparison reflected the classifications produced by each approach from a common clinical-information set. Patient-identifying information was excluded from the clinical information submitted for AI-based assessment, and confidentiality was maintained throughout data collection, analysis, and reporting. Cases with substantial missing monitoring information or inadequately documented outcomes were excluded rather than used for outcome-performance calculations.

Data was entered, cleaned, and analyzed using IBM SPSS Statistics version 26.0. Categorical variables were summarized using frequencies and percentages, while continuous variables were summarized using mean and standard deviation or median and interquartile range according to their distribution. For each analyzable anesthesia-related complication, predictions were evaluated at the patient level. The predictive performance of anesthesiologists and ChatGPT relative to documented clinical outcomes was assessed using two-by-two classification tables, from which sensitivity, specificity, positive predictive value, negative predictive value, and overall accuracy were determined where the observed data permitted estimation.

Agreement between anesthesiologist and ChatGPT classifications for the same complication was quantified using Cohen's kappa coefficient. McNemar's test was used for paired binary comparisons when assessing differences between anesthesiologist and ChatGPT classifications for the same complication rather than for comparisons between different clinical complications. Chi-square analysis was used for categorical associations when the assumptions of the test were appropriate. Complications with no variability in one or both classifications were not interpreted using kappa as meaningful agreement estimates. All statistical tests were two-sided, and a p-value <0.05 was considered statistically significant.

Ethical approval was obtained from the Clinical Research Committee of the Institute of Paramedical Sciences, Khyber Medical University, Peshawar, and administrative permission for data collection was obtained from Lady Reading Hospital and Khyber Teaching Hospital. The research was conducted in accordance with institutional ethical requirements and the ethical principles of the Declaration of Helsinki. Participant confidentiality was maintained throughout the study, and clinical information used for ChatGPT-based prediction was anonymized before submission to the AI system.

RESULTS

A total of 140 participants were included in the analysis. The largest age category was 41–70 years, comprising 73 participants (52.1%), followed by 42 (30.0%) aged 16–40 years and 25 (17.9%) aged 1–15 years. Males accounted for 88 participants (62.9%). Elective procedures predominated, accounting for 124 cases (88.6%), whereas 16 procedures (11.4%) were performed as emergencies. Most participants had no documented comorbidity (127/140, 90.7%), had normal laboratory investigations (112/140, 80.0%), and were classified as ASA I (110/140, 78.6%). General anesthesia was used in 87 participants (62.1%) and regional anesthesia in 53 (37.9%) (Table 1).

Table 1. Demographic and Clinical Characteristics of the Study Participants (N = 140)

Variable	Category	n (%)
Age, years	1–15	25 (17.9)
	16–40	42 (30.0)
	41–70	73 (52.1)
Sex	Male	88 (62.9)
	Female	52 (37.1)
Height	2 feet	7 (5.0)
	3 feet	14 (10.0)
	4 feet	6 (4.3)
	5 feet	109 (77.9)
	6 feet	4 (2.9)
Type of surgery	Elective	124 (88.6)
	Emergency	16 (11.4)
Comorbidities	Yes	13 (9.3)

Variable	Category	n (%)
	No	127 (90.7)
Laboratory investigations	Normal	112 (80.0)
	Abnormal	28 (20.0)
ASA physical status	ASA I	110 (78.6)
	ASA II	23 (16.4)
	ASA III	7 (5.0)
Type of anesthesia	General anesthesia	87 (62.1)
	Regional anesthesia	53 (37.9)

Abbreviation: ASA, American Society of Anesthesiologists physical status classification.

The reported chi-square analyses showed associations between anesthesiologist and ChatGPT classifications for each of the eight analyzable complications (all $p<0.001$). The largest chi-square values were observed for bradycardia and hypo-/hyperthermia ($\chi^2=140.000$ for each), followed by hypotension ($\chi^2=128.976$) and bleeding ($\chi^2=105.339$). Derived φ coefficients ranged from 0.659 for shivering to 1.000 for bradycardia and hypo-/hyperthermia, indicating that the statistical associations between the two classification approaches varied in magnitude across complications (Table 2).

Table 2. Association Between Anesthesiologist and ChatGPT Predictions for Anesthesia-Related Complications (N = 140)

Complication	χ^2	df	p-value	φ
Hypotension	128.976	1	<0.001	0.960
Shivering	60.867	1	<0.001	0.659
Bradycardia	140.000	1	<0.001	1.000
Hypertension	97.804	1	<0.001	0.836
Bleeding	105.339	1	<0.001	0.867
Tachycardia	66.866	1	<0.001	0.691
Hypoventilation/Hypoxia	91.961	1	<0.001	0.810
Hypo-/Hyperthermia	140.000	1	<0.001	1.000

χ^2, chi-square statistic; φ, phi coefficient. φ was derived as $\sqrt{(\chi^2/N)}$ from the reported chi-square statistic and N=140.

Cohen's kappa analysis similarly demonstrated varying degrees of concordance between the two prediction approaches. The highest coefficients were observed for bradycardia and hypo-/hyperthermia ($\kappa=1.000$ for each), followed by hypotension ($\kappa=0.959$), bleeding ($\kappa=0.867$), and hypertension ($\kappa=0.826$). Kappa coefficients were 0.793 for hypoventilation/hypoxia, 0.659 for shivering, and 0.646 for tachycardia. All eight estimable kappa coefficients had reported p-values <0.001 (Table 3).

Table 3. Cohen's Kappa Agreement Between Anesthesiologist and ChatGPT Predictions (N = 140)

Complication	Cohen's κ	p-value
Hypotension	0.959	<0.001
Shivering	0.659	<0.001
Bradycardia	1.000	<0.001
Hypertension	0.826	<0.001
Bleeding	0.867	<0.001
Tachycardia	0.646	<0.001
Hypoventilation/Hypoxia	0.793	<0.001
Hypo-/Hyperthermia	1.000	<0.001

κ, Cohen's kappa coefficient. Complications reported as having no variability in one or either classifications were not included because a meaningful kappa estimate could not be interpreted.

Across the analyzable complications, the numerical pattern was consistent between the chi-square effect-size estimates and the kappa coefficients. Bradycardia and hypo-/hyperthermia showed the strongest inter-method concordance, whereas tachycardia and shivering showed comparatively lower agreement. Hypotension, bleeding, hypertension, and hypoventilation/hypoxia remained strongly associated between the two prediction approaches. These analyses quantify concordance between anesthesiologist and ChatGPT classifications; they do not by themselves establish diagnostic or predictive accuracy relative to the actual intraoperative outcome.

DISCUSSION

The present study found substantial concordance between anesthesiologist and ChatGPT classifications for the anesthesia-related complications that could be analyzed. Agreement was strongest for bradycardia and hypo-/hyperthermia, for which perfect kappa coefficients were reported, and remained high for hypotension, bleeding, and hypertension. Comparatively lower, although still substantial, concordance was observed for hypoventilation/hypoxia, shivering, and tachycardia. These findings indicate that ChatGPT frequently classified the evaluated complications in a manner similar to anesthesiologists when both approaches were provided with the same clinical information. Importantly, however, concordance between two prediction approaches is distinct from predictive accuracy against observed clinical outcomes; therefore, the present results principally support similarity of classification rather than independent validation of ChatGPT as a clinical prediction system.

The observed concordance is broadly consistent with emerging evidence suggesting that large language models can reproduce selected elements of anesthesia-related clinical reasoning under structured conditions. Cheng et al. compared ChatGPT with expert assessment using 150 simulated day-surgery presentations and found comparable performance for several aspects of ASA physical-status classification and pre-anesthesia risk assessment, although differences remained in recommendations for patients with greater clinical complexity (9). Similarly, exploratory comparisons of contemporary AI models for preoperative anesthesia planning have demonstrated that language models can generate clinically relevant anesthesia plans while still requiring expert assessment and further validation before clinical implementation (10). Studies using anesthesiology examination material have also shown that large language models can perform strongly on structured knowledge and reasoning tasks (8,14). These findings provide context for the present inter-method agreement, but examination performance and simulated case assessment should not be considered equivalent to prospective prediction of complications in actual perioperative care.

The relatively strong concordance observed for hypotension and other hemodynamic complications may partly reflect the availability of recognizable relationships among patient characteristics, perioperative physiology, anesthetic factors, and common cardiovascular responses. Artificial-intelligence research in perioperative medicine has demonstrated that computational models can integrate multiple clinical variables for risk stratification and intraoperative event prediction (4). Purpose-built deep-learning approaches have also been investigated for predicting intraoperative hypotension from physiological waveforms (5). Such models differ fundamentally from ChatGPT because they are specifically trained or configured to process structured clinical or waveform data. Nevertheless, these studies illustrate that perioperative complications with relatively well-characterized physiological predictors may be more amenable to computational assessment. The present study did not directly investigate the mechanism underlying the observed agreement, and the high concordance should therefore not be interpreted as evidence that ChatGPT uses the same reasoning processes as anesthesiologists.

Agreement was comparatively lower for tachycardia and shivering. These complications may occur in heterogeneous clinical contexts and can be influenced by several concurrent perioperative factors, making prediction dependent on the completeness and specificity of the information presented to the assessor. This observation is compatible with the broader literature showing that the performance of large language models in anesthesiology varies according to the clinical task, information supplied, model characteristics, and method of evaluation (6,7). Evidence from other clinical interpretation tasks has likewise demonstrated that strong general medical-language capabilities do not guarantee specialist-level performance when an AI system is applied to real clinical data (11). Accordingly, variation in agreement across complications in the present study should be considered clinically meaningful rather than assuming that a single overall measure represents ChatGPT performance across all perioperative conditions.

The findings also illustrate an important distinction between agreement, reliability, and clinical validity. Cohen's kappa measures concordance beyond chance but does not establish whether either assessor correctly predicted the subsequent clinical event. Similarly, the significant chi-square findings demonstrate associations between anesthesiologist and ChatGPT classifications but do not independently demonstrate sensitivity, specificity, positive predictive value, negative predictive value, or clinical benefit. This distinction is particularly important for rare complications, for which limited event prevalence can substantially influence agreement statistics and, in the present study, resulted in several outcomes having insufficient variability for meaningful kappa estimation.

Consequently, the current findings should be interpreted as evidence of inter-method concordance rather than evidence that ChatGPT can independently replace established perioperative risk assessment.

The potential use of generative AI in clinical settings also requires attention to reproducibility, reliability, transparency, and data governance. Standardized methods for evaluating the completeness, accuracy, relevance, and evidence basis of health information produced by AI systems have been proposed because apparently plausible responses may not consistently represent clinically reliable information (15). Ethical concerns regarding privacy, accountability, algorithmic bias, transparency, and inappropriate reliance on AI-generated recommendations remain particularly relevant when patient information is involved (12). Within anesthesiology, ChatGPT has therefore been discussed primarily as a potential adjunct to professional practice rather than a substitute for specialist clinical assessment and responsibility (13). The current findings support this cautious interpretation because similarity to anesthesiologist classification alone is insufficient to demonstrate safety or effectiveness in patient management.

The study has several strengths. Both prediction approaches were evaluated using the same available clinical information, allowing patient-level comparison without systematically providing one approach with a richer information set. Data were obtained from two tertiary-care hospitals, and the analysis examined multiple clinically relevant anesthesia-related complications rather than a single outcome. The paired assessment also permitted quantification of inter-method agreement using Cohen's kappa. However, several limitations constrain interpretation. The sample comprised 140 participants and was dominated by elective procedures, which accounted for 88.6% of cases, while only 11.4% involved emergency surgery. In addition, 78.6% of participants were classified as ASA I, indicating predominance of a relatively low-risk population. This clinical spectrum may have increased the proportion of uncomplicated or concordantly negative classifications and limits generalizability to higher-risk perioperative populations. Several uncommon complications showed little or no variability and consequently could not be meaningfully assessed.

A further limitation is that the aggregate results available for the present analysis did not provide the complete outcome-referenced contingency data required to quantify sensitivity, specificity, predictive values, overall accuracy, and confidence intervals separately for ChatGPT and anesthesiologists. The reported statistics therefore establish inter-method concordance but cannot determine whether one approach was more accurate in predicting documented complications. The performance of an LLM may also depend on the precise model version, prompt structure, input completeness, and temporal context, making detailed procedural standardization important for reproducibility (6,15). Finally, the study was conducted at two hospitals within the same geographical setting and did not include external validation in a clinically distinct population. These limitations should be considered when interpreting the apparent strength of agreement.

Despite these limitations, the findings contribute to the developing literature on potential applications of generative AI in anesthesiology by showing that ChatGPT can produce classifications that frequently correspond with anesthesiologist judgments when presented with the same structured clinical information. Large language models have also been investigated as tools for medical education and structured clinical-reasoning support, including simulation-based approaches in which AI-generated feedback enhanced aspects of clinical decision-making training (16). Such applications may ultimately prove useful in education or decision support, but the present study did not evaluate educational effectiveness, changes in clinician behavior, patient outcomes, or safety. Future evaluation of this approach should therefore distinguish clearly between agreement with clinicians, accuracy against observed outcomes, and actual clinical utility.

CONCLUSION

ChatGPT demonstrated substantial to perfect concordance with anesthesiologist classifications for the analyzable anesthesia-related complications, with the strongest agreement observed for bradycardia, hypo-/hyperthermia, hypotension, bleeding, and hypertension and comparatively lower agreement for tachycardia and shivering. These findings indicate that ChatGPT can generate complication classifications similar to those of anesthesiologists when provided with the same clinical information; however, the reported analyses establish inter-method agreement rather than independent predictive accuracy against actual clinical outcomes. ChatGPT should therefore be interpreted as a potentially supportive computational approach rather than a validated substitute for anesthesiologist assessment on the basis of the present data.

REFERENCES

1. Irita K, Tsuzaki K, Sanuki M, Sawa T, Nakatsuka H, Makita K, et al. Recent changes in the incidence of life-threatening events in the operating room: JSA surveys between 2001 and 2005. Masui. 2007;56:1433-46.
2. Cook TM, Woodall N, Frerk C; Fourth National Audit Project. Major complications of airway management in the UK: results of the Fourth National Audit Project of the Royal College of Anaesthetists and the Difficult Airway Society. Part 1: Anaesthesia. Br J Anaesth. 2011;106(5):617-31. doi:10.1093/bja/aer058.
3. Almghairbi DS, Marufu TC, Moppett IK. Anaesthesia workload measurement devices: qualitative systematic review. BMJ Simul Technol Enhanc Learn. 2018;4(3):112-6.
4. Yoon HK, Yang HL, Jung CW, Lee HC. Artificial intelligence in perioperative medicine: a narrative review. Korean J Anesthesiol. 2022;75(3):202-15. doi:10.4097/kja.22157.
5. Jo YY, Jang JH, Kwon JM, Lee HC, Jung CW, Byun S, et al. Predicting intraoperative hypotension using deep learning with waveforms of arterial blood pressure, electroencephalogram, and electrocardiogram: retrospective study. PLoS One. 2022;17(8). doi:10.1371/journal.pone.0272055.
6. Daccache N, Zako J, Morisson L, Laferrière-Langlois P. The applications of ChatGPT and other large language models in anesthesiology and critical care: a systematic review. Can J Anaesth. 2025;72(6):904-22. doi:10.1007/s12630-025-02973-9.
7. Barroso A, Casans R. Application of generative artificial intelligence chatbots in the field of anesthesia. Rev Esp Anestesiol Reanim. 2025;72:501667.
8. Angel MC, Rinehart JB, Cannesson MP, Baldi P. Clinical knowledge and reasoning abilities of AI large language models in anesthesiology: a comparative study on the American Board of Anesthesiology examination. Anesth Analg. 2024;139(2):349-56. doi:10.1213/ANE.0000000000006892.
9. Cheng T, Li Y, Gu J, He Y, He G, Zhou P, 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;13:111. doi:10.1186/s13741-024-00469-6.
10. Wang B, Tian Y, Wang XT. An exploratory comparison of AI models for preoperative anesthesia planning. J Med Syst. 2025;49:104. doi:10.1007/s10916-025-02243-7.
11. Çamkıran V, Tunç H, Achmar B, Ürker TS, Kutlu İ, Torun A. Artificial intelligence (ChatGPT) ready to evaluate ECG in real life? Not yet! Digit Health. 2025;11:20552076251325279. doi:10.1177/20552076251325279.
12. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res. 2023;25. doi:10.2196/48009.
13. Gupta B, Ahluwalia P, Gupta A, Mahaseth R. ChatGPT in anesthesiology practice: a friend or a foe. Saudi J Anaesth. 2024;18(2):150-3. doi:10.4103/sja.sja_336_23.
14. Patel S, Ngo V, Wilhelmi B. Evaluating large language models on American Board of Anesthesiology-style anesthesiology questions: accuracy, domain consistency, and clinical implications. J Cardiothorac Vasc Anesth. 2025;39(9):2511-5. doi:10.1053/j.jvca.2025.05.033.
15. Sallam M, Barakat M, Sallam M. Pilot testing of a tool to standardize the assessment of the quality of health information generated by artificial intelligence-based models. Cureus. 2023;15(11). doi:10.7759/cureus.49373.
16. Brügge E, Ricchizzi S, Arenbeck M, Keller MN, Schur L, Stummer W, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. BMC Med Educ. 2024;24:1391. doi:10.1186/s12909-024-06399-7.