ISSN: 1607-8322, e-ISSN: 2220-5799 Anaesthesia, Pain & Intensive Care
Vol 29(5); August 2025 DOI: 10.35975/apic.v29i5.2873

REVIEW ARTICLE AI IN ANESTHESIOLOGY

Hallucination in artificial intelligence and its implications on anesthesiology practice and patient outcomes

Shoaib Nawaz [1] , Kiran Ahmad [2] , Odai Khamash [3] , Ejaz Khan [4] , Roni Mendonca [5]

Authors affiliations:

1. Shoaib Nawaz, New York Medical College - Metropolitan Hospital, 1901 1st avenue, New York, NY 10029, USA: Email: shoaibnawaz89@hotmail.com; {ORCID:0000-0001-9119-5648}
2. Kiran Ahmad, Hamad Medical Corporation, Doha- Qatar; Email: kiranahmad_59@yahoo.com; {ORCID:0009-0000-7380-5078}
3. Odai Khamash, New York Medical College - Metropolitan Hospital, 1901 1st avenue, New York, NY 10029, USA: Email: okhamash@gmail.com
4. Ejaz Khan, New York Medical College - Metropolitan Hospital, 1901 1st avenue, New York, NY 10029, USA: Email: ekhan4@nychhc.org
5. Roni Mendonca, New York Medical College - Metropolitan Hospital, 1901 1st avenue, New York, NY 10029, USA: Email: roni.mendonca@nychhc.org

Correspondence: Shoaib Nawaz, **Email:** shoaibnawaz89@hotmail.com; {ORCID:0000-0001-9119-5648}

ABSTRACT

Artificial intelligence (AI) is revolutionizing anesthesiology by enhancing patient monitoring, optimizing drug dosing, and predicting adverse intraoperative events. AI-driven models, particularly those utilizing deep learning, are increasingly used for anesthesia depth monitoring, hemodynamic control, and perioperative risk stratification. However, a significant challenge in AI-driven healthcare is AI hallucination (AIH)—a phenomenon where AI generates misleading, incorrect, or fabricated information. In anesthesiology, hallucinations can lead to severe consequences, such as incorrect dosing recommendations, misinterpretation of patient monitoring data, and flawed clinical decision support, all of which pose risks to patient safety. This article explores the concept of AIH, its causes, real-world examples of its impact in healthcare, and its potential consequences for anesthesiology practice. We also discuss mitigation strategies, including improving data quality, implementing clinician-in-the-loop models, and ensuring regulatory oversight. As AI becomes increasingly integrated into anesthetic practice, recognizing and addressing the risks of AIH is crucial for improving patient safety and maintaining the integrity of anesthetic care.

Abbreviations: AI: Artificial intelligence, BIS: Bispectral index, DDA: Data-driven analytics, ABM: Algorithm-based management, XAI: explainable Artificial intelligence

Keywords: Artificial intelligence; BIS; Data-driven analytics; Algorithm-based management; Machine learning; AI in anesthesiology

Citation: Nawaz S, Ahmad K, Khamash O, Khan E, Mendonca R. Hallucination in artificial intelligence and its implications on anesthesiology practice and patient outcomes. Anaesth. pain intensive care 2025;29(5):376-381.
DOI: 10.35975/apic.v29i5.2873

Received: May 09, 2024; **Revised:** October 26, 2024; **Accepted:** January 01, 2025

1. INTRODUCTION

Artificial intelligence has emerged as a powerful tool in modern medicine, particularly in anesthesiology, where it is being used to optimize clinical workflows, improve decision-making, and enhance patient safety. AI applications in anesthesiology include anesthesia depth monitoring, where machine learning models analyze electroencephalography (EEG) and bispectral index (BIS) signals to optimize sedation levels and reduce the risk of intraoperative awareness in addition to detecting abnormal Capnography levels.[1,2] Additionally, AI-driven hemodynamic management systems are being

developed to predict and prevent perioperative hypotension, allowing anesthesiologists to proactively intervene before adverse events occur.[3,4] AI-assisted models are also used to predict difficult airways based on patient characteristics, aiding in the selection of appropriate airway management strategies.[5]] Furthermore, AI plays a role in perioperative pharmacology by assisting in drug dosing optimization based on patient-specific variables, reducing postoperative complications and improving analgesia.

Despite these advancements, AI in healthcare is not infallible. A significant limitation is AIH, where AI systems generate misleading or entirely false outputs that appear plausible but are incorrect. In clinical settings, these hallucinations can result in dangerous errors, such as incorrect drug dosing recommendations, misinterpretation of vital signs, and flawed clinical decision support. In anesthesiology, such errors could lead to severe consequences, including intraoperative hemodynamic instability, inadequate analgesia, or increased morbidity and mortality. This article explores the concept of AIH, its causes, implications in anesthesiology, real-world examples of its effects in healthcare, and strategies to mitigate these risks.

2. AI HALLUCINATION

AIH refers to the phenomenon where artificial intelligence systems generate outputs that are incorrect, misleading, or entirely fabricated, despite appearing plausible. The underlying causes of AIH are complex, and their occurrence can be traced to several key factors related to model development, training, and deployment.

2.1. Causes:

One of the main causes of AIH is overfitting. Overfitting occurs when AI models are excessively trained on a limited or biased dataset, causing the model to perform well on training data but fail to generalize when faced with new, unseen data. In healthcare, this can manifest when models are trained on data from a homogenous patient population or when data quality is insufficient. For example, if a machine learning model is trained on a narrow demographic—such as a predominantly Caucasian population—the model may perform poorly when applied to more diverse populations, leading to errors in predictions and diagnoses.[6] In anesthesiology, this overfitting could result in AIH that lead to incorrect anesthesia depth recommendations or inaccurate risk predictions for patients from underrepresented groups.

Another critical factor contributing to AIH is the quality of input data. AI systems rely on large datasets of clinical information for training, but poor-quality or noisy data can significantly degrade the accuracy of predictions. Inaccurate sensor data, for instance, may cause AI algorithms used for real-time anesthesia monitoring to generate misleading information. If an anesthesia machine misreads the patient's blood pressure due to a faulty sensor, an AI model could misinterpret the data and suggest unnecessary interventions or fail to alert clinicians to critical changes in the patient's condition.[3] Similarly, missing or incomplete patient records, such as missing lab results or history of allergies, can lead to flawed decisions. These data artifacts can amplify the risk of hallucination, as the AI system is forced to make predictions or decisions based on incomplete or erroneous data.

Additionally, AI systems are susceptible to what is known as "algorithmic misinterpretation." This occurs when a model misclassifies complex patterns within healthcare data. For example, AI systems might have difficulty interpreting nuanced physiological changes in patients with complex comorbidities, such as those with both sepsis and heart failure, leading to incorrect or delayed treatment recommendations. Furthermore, when AI models are faced with scenarios outside the training data, they may extrapolate inappropriately, leading to hallucinated predictions that do not reflect real-world conditions. For instance, an AI-based anesthesiology tool might incorrectly predict a low-risk anesthesia depth for a patient with a high-risk cardiac history, leading to under-sedation or other adverse outcomes.[7]

Finally, the absence of real-time feedback loops and continuous learning mechanisms also exacerbates the issue of hallucination in AI systems. Most AI models used in anesthesiology are trained and deployed in a static manner, meaning they do not adapt or learn from real-time clinical feedback. As a result, models may not be able to correct themselves when faced with new, unforeseen clinical situations. In a high-stakes environment like the operating room, this inability to adjust to new information can be catastrophic, as AI recommendations may continue to perpetuate errors or misunderstandings without the intervention of human clinicians.[8]

These causes can be summarized as follows:

- **Overfitting**:

  Over-specialization to training data occurs when an AI model is trained too closely on a specific dataset, limiting its ability to generalize. For example, a model trained on healthy adults may misjudge sedation needs for pediatric or elderly patients, leading to under- or over-sedation.

- **Poor Data Quality:**

  Inaccurate or incomplete data: Such as faulty sensor readings or missing patient records, can lead to erroneous outputs. A malfunctioning anesthesia machine providing noisy blood pressure data could prompt inappropriate interventions.

- **Algorithmic Misinterpretation:**

  AI may misclassify complex patient conditions, such as comorbidities (e.g., sepsis and heart failure), resulting in flawed predictions. For instance, an AI tool might underestimate anesthesia risks for a patient with a high-risk cardiac history.

- **Lack of Adaptive Feedback:**

  Static AI models, common in anesthesiology, lack real-time learning capabilities. Without continuous feedback, these systems may perpetuate errors in dynamic surgical environments.

2.2. Adverse events and implications:

Real-world examples underscore the significant dangers posed by AIH in healthcare, particularly in anesthesiology. One notable example comes from the field of oncology, where IBM Watson for Oncology, an AI-driven clinical decision support system, was found to recommend incorrect treatment protocols for cancer patients. These recommendations were sometimes based on fabricated or incomplete patient data, rather than real-world evidence. In one instance, Watson recommended a highly toxic regimen for a patient with a rare form of cancer, which was subsequently identified as potentially harmful after a human oncologist reviewed the case.[9,10] This case highlighted how AI models trained on incomplete datasets or flawed assumptions can generate hallucinated treatment protocols that pose risks to patients' health.

In anesthesiology, AIH have had severe consequences as well. One case involved a machine learning algorithm used to predict pediatric opioid dosing in the operating room. The AI model, trained on outdated pharmacokinetic data, suggested dosages far higher than clinically recommended for a pediatric patient, which could have resulted in respiratory depression and overdose if not caught by the attending anesthesiologist. This example demonstrates the potential danger of AIH in recommending inappropriate drug dosages, particularly for sensitive patient populations such as children.

Misinterpretations of medical imaging data by AI systems have also resulted in significant patient harm. A deep-learning algorithm developed to diagnose pneumonia in chest X-rays was found to be highly biased. The model failed to distinguish between real lung pathology and incidental findings like metal implants, incorrectly diagnosing metal objects as signs of pneumonia. This type of AIH illustrates how reliance on AI for diagnostic imaging without sufficient validation can lead to catastrophic misdiagnoses.[11] In anesthesiology, similar AI misinterpretations could result in undetected airway complications or overlooked signs of critical intraoperative events, such as pneumothorax.

AI systems used for real-time monitoring of anesthesia depth have also been implicated in generating misleading information that could lead to patient harm. For example, a deep-learning model designed to monitor anesthesia depth failed to accurately track sedation levels for a patient with underlying neurologic conditions. As a result, the patient was under-sedated during surgery, experiencing intraoperative awareness, a potentially traumatic event.[6] While the AI system recommended increasing sedative doses, it was not able to distinguish between different types of sedation responses, exacerbating the issue.

Studies estimate that AI hallucination may affect up to 10-15% of clinical decision support outputs, underscoring the need for robust safeguards.[7] Additionally, model drift, where AI model performance degrades over time due to changes in clinical practice or patient populations, poses a long-term risk in anesthesiology, potentially leading to persistent hallucinated outputs if not addressed.

3. MANAGING AIH

To address the challenges posed by AIH in anesthesiology, it is essential to implement strategies that ensure the reliability and safety of AI-driven systems. One key strategy is to improve the quality and diversity of data used to train AI models. This can be achieved by ensuring that datasets reflect a wide range of patient populations, including those from different age groups, ethnicities, and comorbid conditions. In anesthesiology, this would help ensure that AI systems are capable of making accurate predictions for all patient types, reducing the risk of biases that could lead to hallucinated outputs. High-quality datasets with well-curated data are essential for minimizing errors and improving predictive accuracy.[12]

Additionally, AI models should undergo rigorous validation using external datasets before being deployed in clinical settings. External validation helps to identify

Table 1: A summary of the applications, risks and mitigation strategies to prevent AI hallucination

Application	AI Technologies	Benefits	Risks of Hallucination	Mitigation Strategies
Anesthesia Depth Monitoring	Deep neural networks for EEG/BIS analysis, signal processing algorithms	Optimizes sedation, reduces intraoperative awareness	Misinterpretation of EEG/BIS or capnography signals due to neurological conditions or data noise; model drift over time	Diverse datasets including neurological conditions, clinician oversight, regular model retraining, uncertainty quantification
Hemodynamic Management	Machine learning models (e.g., random forests, LSTM networks) for vital sign prediction	Predicts hypotension, guides timely interventions	Errors from noisy sensor data (e.g., pulse oximetry); data imbalance leading to biased predictions	Data quality checks, real-time feedback loops, cross-validation with diverse patient data, automated sensor calibration
Risk Prediction	Decision trees, ensemble models for airway and risk assessment	Identifies difficult airways, predicts perioperative risks	Misclassification of patient characteristics due to anatomical anomalies or insufficient training data	External validation with diverse datasets, explainable AI for transparency, clinician review of predictions
Decision Support	Reinforcement learning, regression models for drug dosing	Optimizes drug dosing, reduces postoperative complications	Inappropriate dosing recommendations due to outdated pharmacokinetic data or overfitting to specific populations	Clinician-in-the-loop validation, regulatory oversight, periodic model updates, integration of patient-specific variables

potential weaknesses in the model's performance and ensure that it can be safely applied to diverse patient populations.[13] This validation process should be continuous, incorporating feedback from real-world clinical outcomes to improve the model's reliability and adaptability over time.

Clinician-in-the-loop approaches, where human clinicians maintain oversight and intervene when necessary, are essential to prevent AIH from leading to patient harm. In anesthesiology, AI should not replace the clinician's judgment, but rather support it by providing additional data-driven insights. Anesthesiologists should be trained to interpret AI recommendations with a critical eye, using their clinical expertise to validate or override AI-generated outputs as needed.[7,13]

The development of explainable AI (XAI) is another crucial strategy for minimizing AIH in clinical practice. XAI models allow clinicians to understand the rationale behind AI predictions, increasing transparency and trust in AI-driven systems. In anesthesiology, XAI could help anesthesiologists assess how an AI model reached a particular recommendation, enabling them to detect any potential errors or inconsistencies in the model's logic before acting on its suggestions.[14]

Strengthening feedback loops is also vital. AI models used in anesthesiology should be designed to adapt and improve continuously based on real-time feedback from clinicians and patient outcomes. This allows the AI system to refine its predictions and improve its accuracy over time. Feedback loops enable AI models to correct themselves and learn from past mistakes, reducing the risk of hallucination and ensuring that the system remains relevant and effective in the dynamic environment of the operating room.

Finally, regulatory oversight is critical to ensure the safety of AI technologies in healthcare. Regulatory bodies should establish clear guidelines for the development, testing, and deployment of AI systems in anesthesiology, with an emphasis on patient safety.[14,15]] These guidelines should address issues such as model validation, data quality, and continuous monitoring of AI performance in clinical settings.

By addressing these strategies, the risks of AIH in anesthesiology can be mitigated, enhancing patient safety and ensuring that AI-driven tools support

clinicians in delivering the best possible care. A summary of the mitigation strategies to prevent artificial intelligence hallucination is provided in Table 1.

To maximize AI's potential in anesthesiology, future efforts should focus on developing adaptive AI systems that learn from real-time clinical feedback, reducing hallucination risks. Integrating multimodal data (e.g., EEG, vital signs, capnography, and patient history) can enhance model accuracy. Collaboration between AI developers, anesthesiologists, and regulators will ensure safe, transparent, and effective AI tools. Research into hybrid models combining AI with human expertise could further mitigate risks, enabling precision anesthesia care.

4. CONCLUSION

AI is rapidly transforming anesthesiology by improving patient monitoring, optimizing drug administration, and predicting intraoperative complications. However, AI hallucination poses a significant challenge that could compromise patient safety if not properly addressed. The risks of erroneous AI-generated recommendations, misinterpretation of patient data, and automation bias highlight the importance of robust validation, clinician oversight, and continuous monitoring. By implementing strategies such as improving data quality, developing explainable AI models, and maintaining regulatory oversight, AI hallucination can be minimized, paving the way for safer and more effective anesthesia care.

5. Conflict of interests

All authors declare that there was no conflict of interest.

6. Funding

The study utilized the hospital resources only, and no external or industry funding was involved.

7. Authors' contribution

SN: Literature review, manuscript writing

KA, OK: manuscript writing

EK: Literature review

RM: Supervision and references review

8. REFERENCES

1. Connor CW. Artificial intelligence and machine learning in anesthesiology. *Anesthesiology.* 2019;131(6):1346–1359. PMCID: PMC6778496 DOI: 10.1097/ALN.0000000000002694

2. Spijkerboer FL, Overdyk FJ, Dahan A. A machine learning algorithm for detecting abnormal patterns in continuous capnography and pulse oximetry monitoring. J Clin Monit Comput. 2024 Aug;38(4):915-925. 024-01155-0. Epub 2024 Apr 15. PMID: 38619716; PMCID: PMC11297897 DOI: 10.1007/s10877-024-01155-0

3. Sidiropoulou T, Tsoumpa M, Griva P, Galarioti V, Matsota P. Prediction and Prevention of Intraoperative Hypotension with the Hypotension Prediction Index: A Narrative Review. J Clin Med. 2022 Sep 22;11(19):5551. PMID: 36233419 PMCID: PMC9571689 DOI: 10.3390/jcm11195551

4. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. JAMA. 2020;323(11):1052–1060. PMID: 32065827 PMCID: PMC7078808 DOI: 10.1001/jama.2020.0592

5. Matava C, Pankiv E, Ahumada L, Weingarten B, Simpao A. Artificial intelligence, machine learning and the pediatric airway. Paediatr Anaesth. 2020 Mar;30(3):264-268. DOI: 10.1111/pan.13792. Epub 2020 Jan 2 PMID: 31845543 DOI: 10.1111/pan.13792

6. Singhal M, Gupta L, Hirani K. A Comprehensive Analysis and Review of Artificial Intelligence in Anaesthesia. Cureus. 2023 Sep 11;15(9):e45038. PMID: 37829964 PMCID: PMC10566398 DOI: 10.7759/cureus.45038

7. Ocampo Osorio F, Alzate-Ricaurte S, Mejia Vallecilla TE, Cruz-Suarez GA. The anesthesiologist's guide to critically assessing machine learning research: a narrative review. BMC Anesthesiol. 2024 Dec 18;24(1):452. PMID: 39695968 PMCID: PMC11654216 DOI: 10.1186/s12871-024-02840-y

8. Choudhury A, Asan O. Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review. JMIR Med Inform. 2020 Jul 24;8(7):e18599. PMID: 32706688 PMCID: PMC7414411 DOI: 10.2196/18599

9. Liu C, Liu X, Wu F, Xie M, Feng Y, Hu C. Using Artificial Intelligence (Watson for Oncology) for Treatment Recommendations Amongst Chinese Patients with Lung Cancer: Feasibility Study. J Med Internet Res. 2018 Sep 25;20(9):e11087. PMID: 30257820 PMCID: PMC6231834 DOI: 10.2196/11087

10. Istasy P, Lee WS, Iansavichene A, Upshur R, Gyawali B, Burkell J, et al. The Impact of Artificial Intelligence on Health Equity in Oncology: Scoping Review. J Med Internet Res. 2022 Nov 1;24(11):e39748. PMID: 36005841 PMCID: PMC9667381 DOI: 10.2196/39748

11. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 2018 Nov 6;15(11):e1002683. DOI: 10.1371/journal.pmed.1002683. PMID: 30399157; PMCID: PMC6219764

12. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019 Apr 20;393(10181):1577-1579. PMID: 31007185 DOI: 10.1016/S0140-6736(19)30037-6

13. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, Kohane IS, Saria S. The Clinician and Dataset Shift in Artificial Intelligence. N Engl J Med. 2021 Jul 15;385(3):283-286. PMID: 34260843 PMCID: PMC8665481 DOI: 10.1056/NEJMc2104626

14. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial Intelligence in Anesthesiology: Current Techniques, Clinical Applications, and Limitations. Anesthesiology. 2020 Feb;132(2):379-394. PMID: 31939856 PMCID: PMC7643051 DOI: 10.1097/ALN.0000000000002960

15. Amann J, Blasimme A, Vayena E, Frey D, Madai VI; Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. 2020 Nov 30;20(1):310. PMID: 33256715 PMCID: PMC7706019 DOI: 10.1186/s12911-020-01332-6

