REVIEW ARTICLE

CONTENTS

Artificial Intelligence for predicting adverse surgical outcomes: Challenges, limitations and implications for clinical translation – a narrative review

Zain Ahmed[1], Tanvi Prem Kumar[1], Alina Stachyra[2]
[1]MEDICAL UNIVERSITY OF LUBLIN, LUBLIN, POLAND
[2]DEPARTMENT OF REHABILITATION, MEDICAL UNIVERSITY OF LUBLIN, LUBLIN, POLAND

ABSTRACT

The rise in the number of surgeries per year has led to the development of many artificial intelligence models for predicting surgical complications. Despite their ever-growing use in healthcare, artificial intelligence is not up to the mark yet. We need to search and critically overcome the hurdles preventing their safe and reliable use in surgical care. This narrative review aims to find and analyze the main limitations and challenges of artificial intelligence in predicting surgical outcomes. Across the reviewed literature, key limitations were identified in four domains: data-related, methodological limitations, performance and generalizability, and barriers to clinical implementation. Common issues included missing and imbalanced datasets, small sample sizes, retrospective single-center designs, high risk of bias, and inadequate external validation. Although several studies reported high predictive performance, these findings were often derived from non-representative datasets and lacked prospective validation. Additional concerns included limited interpretability, ethical and privacy risks, workflow integration difficulties, and potential amplification of healthcare disparities. Despite their potential, AI models for surgical outcome prediction remain constrained by multiple challenges. Substantial improvements in data quality, transparency, fairness, and robust multicenter prospective validation are required before AI can be safely and reliably integrated into routine surgical decision-making.

KEY WORDS: bias, transparency, postoperative complications, risk assessment

Wiad Lek. 2026;79(3):605-610. doi: 10.36740/WLek/218217 DOI

INTRODUCTION

According to the World Health Organization (WHO) more than 234 million major surgical procedures are performed worldwide each year, highlighting the vast scale of surgical care [1]. Despite advances in perioperative management, 10 to 25 percent of surgical patients experience major postoperative complications leading to higher mortality, prolonged hospitalization, increased need for intensive care, and greater healthcare costs [2, 3]. Importantly, nearly half of these adverse events are related to surgical care itself and are considered preventable. These complications typically arise from the the combination of multiple variables, and it is humanly impossible to understand and predict all their linked effects on the surgical outcome [3].

In this context, artificial intelligence (AI) has gained increasing attention in surgery. AI in general can be understood as computer systems that perform tasks such as prediction, classification, and decision-making. Machine learning (ML), a key branch of AI, uses algorithms to pick up patterns from a set of data, while deep learning can use multilayered neural networks to find complex associations within the data [4-6]. Predicting surgical complications using AI models can now be achieved because of the large amount of digital data available from health care centers [5]. Hence, ML has an upper hand over traditional prediction methods that use statistics [6].

Despite considerable enthusiasm surrounding AI in healthcare, it has several drawbacks that have prevented it from being utilized on a larger scale [7, 8]. There is a pressing need not only to develop validated predictive models but also to ensure that AI outputs are reliable, interpretable, transparent, ethical and clinically meaningful to support safe surgical decision-making.

AIM

This narrative review aims to find and analyze the main limitations and challenges of artificial intelligence in predicting surgical outcomes.

MATERIALS AND METHODS

SEARCH STRATEGY

A literature search was conducted using PubMed for articles published within the last ten years. The following Medical Subject Headings (MeSH) terms were applied: "Artificial Intelligence", "Deep Learning", "Surgical Procedures, Operative", "Risk Assessment", "Treatment Outcome". The following keywords were additionally searched in titles and abstracts: limitations, challenges, drawbacks, barriers, bias, transparency, "black box".

Inclusion criteria:

Studies were included if they:

1. Discussed the use of AI in surgery, and
2. Focused on predicting surgical outcomes, and
3. Addressed limitations of AI in surgical practice.

Exclusion criteria:

1. Non-English articles
2. Conference abstracts without full text

No ethical approval was required as this study is based on published literature.

REVIEW

DATA-RELATED LIMITATIONS

Across the included studies, substantial limitations related to data quality, representativeness, and structure were consistently reported. Several authors highlighted the problem of missing data [2, 7, 9, 10]. Moglia et al.(2021) reported that most studies did not describe how missing data were handled, representing a major potential source of bias. Likewise, two studies emphasized that substantial missing data and exclusion of patients with incomplete records distorted training datasets[9,10]. One article noted that in their study outcomes were often derived from administrative codes rather than manual chart review, raising concerns that automated data may not fully reflect clinical reality [2].

Population imbalance and lack of representativeness were also recurring issues. Sargiotis et al. (2024) described how reliance on datasets like United Network for Organ Sharing (UNOS) registries led to overrepresentation of White and male patients introducing demographic bias into predictive models [10]. Zander et al.(2025) similarly argued that inadequate numbers of patients within certain demographic groups limited the ability to build fair prediction models [11]. In pediatric cardiac surgery, Florquin et al.(2024) highlighted severe class imbalance as a major stumbling-block to predicting complications, moreover rare but critical complications were difficult for algorithms to learn [12].

Finally, beyond issues of missingness and representativeness, several studies highlighted structural limitations in how surgical data are generated, organized, and shared. The inherently heterogeneous nature of surgical datasets was described as a barrier for the AI models to interpret and process it [13]. Moglia et al. (2021) emphasized that characteristically different data require costly and labor-intensive anonymization, curation, and standardization further restricting dataset availability [7]. Interoperability barriers between institutions further impede data pooling, as differences in electronic health record systems, privacy regulations, and security concerns limit multicenter collaboration [14].

METHODOLOGICAL LIMITATIONS

Many authors emphasized that single-centre and retrospective designs limited clinical applicability because the data were prone to selection bias and the models were developed without prospective validation in real clinical workflows [15-18].

Small sample size further constrained model reliability in several surgical contexts. Göktürk et al.(2025) acknowledged that limited sample size weakened statistical power despite the use of synthetic minority oversampling technique (SMOTE), while Golubovic et al.(2025) and Takkavatakarn et al.(2023) noted that models were built on small, procedure-specific cohorts. Moglia et al.(2021) also attributed limited robustness of models to consistently small datasets across studies [8, 14, 16].

Risk of bias in model development was widespread. A review found that 29 of 31 models were at high risk of bias as per Prediction model Risk Of Bias Assessment Tool (PROBAST), largely due to inadequate sample size, overfitting, excessive predictors relative to events, and lack of external validation [9]. Nayebirad et al.(2025) also reported very high PROBAST bias across percutaneous coronary intervention (PCI) prediction studies, with overfitting being a central concern [19]. Sargiotis et al. (2024) further documented bias arising from exclusion of certain patient groups, use of non-standardized registries, and poor calibration reporting [10].

External validation practices were generally weak. Groot et al.(2021) found that only 10 of 59 orthopedic ML models had any external validation and none were prospective [20]. Moglia et al.(2021) similarly noted that no robot-assisted surgery model had been tested on external datasets, preventing conclusions about robustness or required training sample size [7]. Göktürk et al.(2025), Bektaş et al.(2022), and Golubovic et al.(2025) all stressed the need for large, multicenter, prospective validation before clinical deployment

[8,16,17]. Yu et al.(2025) reported that only 2 of 10 PCI studies conducted external validation despite high reported accuracy [18].

Finally, methodological inconsistency was evident. Moglia et al.(2021) noted that studies failed to define appropriate performance thresholds or agree on which metrics should be prioritized to deem AI models safe to use clinically, while Takkavatakarn et al.(2023) highlighted the lack of head-to-head comparison between models across populations [7, 14].

PERFORMANCE AND GENERALISABILITY

Xue et al (2021). achieved high Area Under the Receiver Operating Characteristic Curve (AUROC) for multiple complications, indicating great model performance, but acknowledged that single-center data and incomplete variable sets limited transferability to other hospitals [2]. Göktürk et al. (2025) described their model only as proof-of-concept lacking external validation [8].

A few reviews questioned whether superior performance translated into meaningful clinical benefit. Nayebirad et al.(2025) found that although ML models had numerically higher c-statistics than traditional logistic regression, differences were not statistically significant [19]. Sargiotis et al. (2024) highlighted that outcomes beyond one year post-transplant were undermined by loss to follow-up and missing data, weakening reliability for chronic risk prediction [10].

Florquin et al.(2024), Göktürk et al.(2025), Salah et al.(2025) all stated that the generalizability was significantly reduced by their single-center and retrospective designs [8,12,15]. Xue et al (2021) goes on to mention that absence of certain clinical values can potentially hinder predictive accuracy as well as generalizability [2]. Ethnic and geographic specificity further constrained generalizability. Dong et al. (2025) demonstrated strong performance for post surgical gastrointestinal bleeding in predominantly Chinese cohorts but explicitly stated that global applicability required prospective, multicenter validation [6]. Similarly, Yu et al.(2025) reported that most PCI studies were conducted in Asian registries [18].

CLINICAL IMPLEMENTATION CHALLENGES

Interpretability ("black box") was a dominant barrier across studies for clinical implementation. Although Dong et al. (2025), Göktürk et al.(2025)zha, and Salah et al.(2025) employed explainable Ai tools, all acknowledged that they do not fully resolve transparency problems or provide clear, actionable guidance when models conflict with clinical judgment [6,8,15]. Nayebirad et al.(2025) further added that this opacity may make ML unpopular despite their better performance [19]. Harris & Matthews (2024) emphasized that narrow, task-specific algorithms lack holistic clinical reasoning and will possibly struggle with atypical cases [5].

Workflow integration posed additional challenges. Balch et al. (2021) described persistent difficulties embedding AI tools into electronic health records, citing usability issues, cost, and clinician mistrust [4]. Moglia et al.(2021) likewise warned that the abundant use of technical explanations discouraged adoption by surgeons and other healthcare professionals [7]. Bedford et al.(2024) noted the absence of clear guidelines for perioperative AI [21].

Ethical, legal, and privacy risks were repeatedly raised across the literature. Dong et al. (2025) identified the difficulty of securing consent for massive datasets and the persistent danger of re-identifying people from supposedly anonymous data [6]. Moglia et al.(2021) expanded on these concerns, noting broader risks involving cybersecurity, liability for AI-related harm, and the need for updated professional credentialing and certification [7]. D'Oria et al.(2024) emphasized that these technologies could threaten patient autonomy while potentially making existing healthcare inequalities even worse [13].

Fairness and equity concerns were noted. Lucas et al.(2024) demonstrated clear racial disparities in colorectal cancer readmission models, including higher false negatives for "Other" race and higher false positives for Black patients [22]. Bedford et al.(2024) and Sargiotis et al. (2024) both warned that biased data could propagate inequities in risk assessment and treatment allocation [10, 21].

DISCUSSION

Many studies demonstrate promising predictive performance; unfortunately, reality reveals persistent and interconnected problems related to data quality, methodology, generalisability, and clinical implementation that currently restrict the safe and reliable integration of AI into perioperative care. These shortcomings appear to be systemic spanning the entire lifecycle of AI development.

The most fundamental building block of a prediction model is high quality and quantity data [7]. A common phrase used is "Garbage in, Garbage out", highlighting that AI is only as robust as the data they ingest. Consistent findings of missing data, class imbalance, small sample size and demographic skew

give way to bias and noise. Missing data emerged as a pervasive problem, with several reviews showing that incomplete records and inconsistent reporting distort model training [7, 9, 10]. As highlighted by Xue et al (2021)., many models are trained on proxies of clinical reality rather than true clinical events due to the use of administrative codes [2]. This raises concerns that some AI tools may fail to learn patterns in meaningful physiological or surgical risk factors. Demographic and class imbalance further threaten fairness and validity. The algorithms often reflect the biases of available datasets rather than the diversity of real world patients. Additionally, structural barriers to data sharing including heterogeneity of formats, privacy concerns, and interoperability limitations prevent the creation of truly representative and multicenter datasets.

A striking finding of this review is the dominance of retrospective, single-center study designs. This suggests that much of the current evidence base for surgical AI reflects model performance in historical datasets rather than showcasing clinical effectiveness. The widespread risk of bias identified is particularly concerning. Models with too many complexities, combined with minimal to no external validation, create a high likelihood of overfitting which means models capture noise rather than true associations. Therefore, they perform well with training data-sets but fail in new clinical environments. The scarcity of robust prospective external validation represents a major gap between research and practice. Without testing models across different hospitals, populations, and workflows, it remains unclear whether AI tools are truly generalizable.

These factors taken along with the lack of transparency in understanding these tools demotivates professionals in using them in their practice. This 'black-box' nature breeds mistrust and confusion as to how reliable the decision making of the model is, especially in comparison to the years of experience and critical thinking of a physician. Even if the prediction tools were transparent and accurate, the legal and ethical implications can not be overlooked. The concerns raised by Dong et al. (2025) regarding potential re-identification of anonymized patient data highlights a deeper tension between the drive for large-scale data sharing and the duty to protect patient privacy [6]. As predictive models become increasingly dependent on massive, multi-institutional datasets, the traditional frameworks of informed consent and confidentiality may no longer be sufficient, necessitating new regulatory and ethical standards for data in surgical research and practice. Moglia et al. (2021)'s emphasis on cybersecurity risks and medico-legal liability further illustrates that integrating AI into operative environments introduces vulnerabilities [7]. If an AI system contributes to a harmful clinical decision, it remains unclear whether responsibility lies with the surgeon, the institution, or the technology developers. This ambiguity could discourage clinical adoption and requires clearer legal frameworks before AI can be safely embedded into surgical workflows.

CONCLUSIONS

Artificial intelligence for predicting adverse surgical outcomes shows clear potential but remains premature for routine clinical use. At present, AI functions more as a research tool than a dependable clinical decision aid. Moving forward, meaningful progress will depend less on developing ever more complex algorithms and more on improving data quality, conducting large prospective multicenter validations, and ensuring transparency, fairness, and accountability. If these challenges are addressed, AI could evolve into a valuable adjunct that reliably supports clinical judgment in surgical care.

REFERENCES

1. Weiser TG, Regenbogen SE, Thompson KD, Haynes AB, Lipsitz SR, Berry WR, et al. An estimation of the global volume of surgery: a modelling strategy based on available data. Lancet. 2008 Jul;372(9633):139-44. doi: 10.1016/S0140-6736(08)60878-8. DOI
2. Xue B, Li D, Lu C, King CR, Wildes T, Avidan MS, et al. Use of Machine Learning to Develop and Evaluate Models Using Preoperative and Intraoperative Data to Identify Risks of Postoperative Complications. JAMA Network Open. 2021 Mar 30;4(3):e212240. doi: 10.1001/jamanetworkopen.2021.2240. DOI
3. Bronnert R, Besch G, Hild O, Lihoreau T, Chaussy Y, Ferreira D. Performance of artificial intelligence models for predicting intraoperative complications during surgery in real time: a systematic review and meta-analysis protocol. BMJ Open. 2025 Oct;15(10):e106204. doi: 10.1136/bmjopen-2025-111663. DOI
4. Balch JA, Delitto D, Tighe PJ, Zarrinpar A, Efron PA, Rashidi P, et al. Machine Learning Applications in Solid Organ Transplantation and Related Complications. Front Immunol. 2021 Sep 16;12(1664-3224). doi: 10.3389/fimmu.2021.739728. eCollection 2021. DOI
5. Harris J, Matthews J. Artificial Intelligence: Predicting Perioperative Problems. Br J Hosp Med (Lond). 2024 Aug 30;85(8):1-4. doi: 10.12968/hmed.2024.0262. DOI

6. Dong J, Jin Z, Li C, Yang J, Jiang Y, Li Z, et al. Machine Learning Models With Prognostic Implications for Predicting Gastrointestinal Bleeding After Coronary Artery Bypass Grafting and Guiding Personalized Medicine: Multicenter Cohort Study. J Med Internet Res. 2025 Mar 6;27:e68509. doi: 10.2196/68509. DOI
7. Moglia A, Georgiou K, Georgiou E, Satava RM, Cuschieri A. A systematic review on artificial intelligence in robot-assisted surgery. Int J Surg. 2021 Nov;95:106151. doi: 10.1016/j.ijsu.2021.106151. DOI
8. Göktürk Y, Başarslan SK, Göktürk Ş, Kocaman H, Yıldırım H. Prediction of postoperative haemorrhage after cerebral tumour surgery using machine learning algorithms. BMC Med Inform Decis Mak. 2025 Oct 23;25(1):392. doi: 10.1186/s12911-025-03245-8. DOI
9. Zhang H, Jiang L, Zheng J, Li C. Supervised machine learning-based bias risk of prognostic models for total knee or hip arthroplasty patients: A systematic review. Medicine. 2025 Oct 17;104(42):e45230–0. doi: 10.1097/MD.0000000000045230. DOI
10. Sargiotis GC, Sergentanis TN, Elpida Pavi, Kostas Athanasakis. Predictive Performance of Artificial intelligence Models on Heart and Lung Posttransplant Health Outcomes: A Systematic Review. Exp Clin Transplant. 2024 Nov;22(11):823-833. doi: 10.6002/ect.2024.0207. DOI
11. Zander T, Kendall MA, Wolansky RL, Grimsley EA, Parikh R, Sujka J, et al. Fairness of machine learning readmission predictions following open ventral hernia repair. Surg Endosc. 2025 Aug;39(8):5035-5045. doi: 10.1007/s00464-025-11927-7. DOI
12. Florquin R, Florquin R, Schmartz D, Dony P, Briganti G. Pediatric cardiac surgery: machine learning models for postoperative complication prediction. J Anesth. 2024 Dec;38(6):747-755. doi: 10.1007/s00540-024-03377-7. DOI
13. D'Oria M, Raffort J, Condino S, Cutolo F, Bertagna G, Raffaella Berchiolli, et al. Computational surgery in the management of patients with abdominal aortic aneurysms: Opportunities, challenges, and future directions. Semin Vasc Surg. 2024 Sep;37(3):298-305. doi: 10.1053/j.semvascsurg.2024.07.005. DOI
14. Takkavatakarn K, Hofer IS. Artificial Intelligence and Machine Learning in Perioperative Acute Kidney Injury. Adv Kidney Dis Health. 2023 Jan;30(1):53-60. doi: 10.1053/j.akdh.2022.10.001. DOI
15. Salah M, Al-Ghashmi M, Baker A, Kamkoum H, Alhabash S, Alnawasra H, et al. Interpretable machine learning prediction of extracorporeal shock wave lithotripsy outcomes for urinary stones: a retrospective cohort study. Arch Ital Urol Androl. 2025 Dec 24;97(4):14333. doi: 10.4081/aiua.2025.14333. DOI
16. Golubovic M, Peric V, Stosic M, Stojiljkovic V, Zivic S, Kamenov A, et al. Predicting Major Adverse Cardiovascular Events After Cardiac Surgery Using Combined Clinical, Laboratory, and Echocardiographic Parameters: A Machine Learning Approach. Medicina. 2025 Jul 23;61(8):1323. doi: 10.3390/medicina61081323. DOI
17. Mustafa Bektaş, Beata, Jaime Costa Pereira, Burchell GL, Donald. Artificial Intelligence in Bariatric Surgery: Current Status and Future Perspectives. Obes Surg. 2022 Aug;32(8):2772-2783. doi: 10.1007/s11695-022-06146-1. DOI
18. Yu MY, Yoo HY, Han GI, Kim EJ, Son YJ. Comparing the Performance of Machine Learning Models and Conventional Risk Scores for Predicting Major Adverse Cardiovascular Cerebrovascular Events After Percutaneous Coronary Intervention in Patients With Acute Myocardial Infarction: Systematic Review and Meta-Analysis. J Med Internet Res. 2025 Jul 18;27(1438-8871):e76215–5. doi: 10.2196/76215. DOI
19. Nayebirad S, Hassanzadeh A, Vahdani AM, Mohamadi A, Forghani S, Shafiee A, et al. Comparison of machine learning models with conventional statistical methods for prediction of percutaneous coronary intervention outcomes: a systematic review and meta-analysis. BMC Cardiovasc Disord. 2025 Apr 23;25(1):310. doi: 10.1186/s12872-025-04746-0. DOI
20. Groot OQ, Bindels BJJ, Ogink PT, Kapoor ND, Twining PK, Collins AK, et al. Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. Acta Orthop. 2021 Apr 18;92(4):385-93. doi: 10.1080/17453674.2021.1910448. DOI
21. Bedford JP, Redfern OC, O'Brien B, Watkinson PJ. Perioperative risk scores: prediction, pitfalls, and progress. Curr Opin Anaesthesiol. 2025 Feb 1;38(1):30-36. doi: 10.1097/ACO.0000000000001445. DOI
22. Lucas MM, Schootman M, Laryea JA, Orcutt ST, Li C, Ying J, et al. Bias in Prediction Models to Identify Patients With Colorectal Cancer at High Risk for Readmission After Resection JCO Clin Cancer Inform. 2024 Nov:8:e2300194. doi: 10.1200/CCI.23.00194.

CONFLICT OF INTEREST

The Authors declare no conflict of interest

CORRESPONDING AUTHOR

Zain Ahmed
Medical University of Lublin,
Lublin, Poland
e-mail: zainahmed0506@gmail.com

ORCID AND CONTRIBUTIONSHIP
Zain Ahmed: 0009-0002-6799-8365 A B D E F
Tanvi Prem Kumar: 0009-0000-7989-3363 B D E F
Alina Stachyra: 0009-0008-8633-5245 E F

A – Work concept and design, B – Data collection and analysis, C – Responsibility for statistical analysis, D – Writing the article, E – Critical review, F – Final approval of the article

RECEIVED: 04.12.2025
ACCEPTED: 20.02.2026