ANESTHESIOLOGY

Openventio PUBLISHERS

Review Article

Machine Learning Applications for Prediction of Postoperative Complications

Pieter Lehmann[1*], Dimitris Brenner[2] and Alessandro Dvorak[3]

[1]*Department of Anesthesiology and Pain Management, Jutland University of Clinical Medicine, Denmark*
[2]*Department of Critical Care Medicine, Aegean Institute of Biomedical Research, Greece*
[3]*Division of Anesthesia and Intensive Care, Aegean Institute of Biomedical Research, Greece*

ABSTRACT

Machine learning has moved from a methodological curiosity to an active area of perioperative research, offering a way to exploit the dense, high-dimensional data generated in modern surgical care rather than relying on the handful of variables used by conventional risk scores.

This review surveys the current state of machine learning applications for predicting postoperative complications, covering general morbidity, mortality, acute kidney injury, and delirium as the four categories with the most mature published evidence.

Across these categories, models trained on automatically curated electronic health record data, continuous intraoperative vital sign streams, and even unstructured clinical text have generally outperformed conventional logistic regression-based scores on measures of discrimination.

The strongest work exploits data that traditional scores cannot use at all, such as intraoperative hemodynamic waveforms for kidney injury prediction and free-text preoperative documentation for mortality risk, rather than simply re-fitting familiar preoperative variables with more complex algorithms.

We conclude that external validation across institutions, consistent outcome definitions, and prospective evidence that model-guided decisions actually reduce complications remain the principal barriers separating this literature from routine clinical adoption.

Keywords: *machine learning; postoperative complications; risk prediction; acute kidney injury; postoperative delirium*

Citation: Lehmann P, Brenner D and Dvorak A (2023) Machine Learning Applications for Prediction of Postoperative Complications. Anesthesiology. 3(1):01-05.

DOI: 10.17140/2023.vol3.i01.pp01-05

Corresponding Author: *Pieter Lehmann*

E-mail Address: *p.lehmann@jucm.dk*

Article History: *Received 15 June 2023 | Accepted 10 July 2023 | Published 10 August 2023*

Copyright: *© 2023 Pieter Lehmann, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (**https://creativecommons.org/licenses/by/4.0/**).*

INTRODUCTION

Anesthesiologists and surgeons have used risk scores for decades to sort patients into rough tiers of surgical risk, but tools like the ASA physical status classification or the Revised Cardiac Risk Index were built from a handful of variables chosen for their availability and clinical plausibility rather than their predictive power. The modern perioperative electronic health record contains vastly more information than any of these

scores was designed to use: continuous vital sign waveforms, laboratory trajectories, medication timing, free-text notes, and increasingly granular intraoperative telemetry from anesthesia workstations. Machine learning offers a way to actually exploit that density of data, learning nonlinear interactions among dozens or hundreds of variables instead of forcing a clinician to hand-pick a short list of predictors. Over the past several years this has moved from a methodological curiosity into a genuine area of clinical research activity, with models now published for general postoperative morbidity, mortality, acute kidney injury, and delirium, among other endpoints [1].

This review surveys where that work currently stands. We are not trying to catalog every published algorithm — the field is moving too quickly and the individual models too heterogeneous in cohort, features, and validation quality for that to be useful. Instead we describe the major complication categories where machine learning has been applied, the data sources and modeling approaches that have proven most productive, and the recurring methodological problems that limit how much of this literature is ready for clinical deployment.

GENERAL POSTOPERATIVE MORBIDITY PREDICTION

The earliest and still most common target for perioperative machine learning is a composite outcome — "any complication," a Clavien-Dindo grade above some threshold, or a bundle of specific adverse events — predicted from data available before or during surgery. Corey and colleagues demonstrated that models built from automatically curated EHR data, without manual chart abstraction, could identify high-risk surgical patients with performance rivaling models built on painstakingly curated registries, which mattered practically because it meant risk stratification could scale beyond the handful of institutions willing to fund dedicated data abstraction [2]. That basic idea — that routinely collected data, mined by a sufficiently capable algorithm, can substitute for expensive manual curation — has been a recurring theme since.

Digestive and abdominal surgery has attracted particular attention because postoperative complications there are common, morbid, and heterogeneous, which makes them both clinically important and analytically tractable for models with many candidate predictors. A recent scoping review catalogued the growing number of models developed specifically for digestive surgery and found substantial variability in outcome definitions, feature sets, and reporting quality, even as the underlying performance metrics were often respectable [1]. A parallel systematic review focused on major abdominal surgery reached a similar conclusion: artificial intelligence models frequently outperformed traditional logistic regression-based scores in discrimination, but few had been externally validated in populations different from their derivation cohort, and fewer still had been tested prospectively against the standard of care [3]. Colorectal surgery specifically has seen enough model development now to support its own systematic review and meta-analysis, which found that machine learning approaches modestly but consistently outperformed conventional risk models for complications such as anastomotic leak and surgical site infection, while cautioning that pooled estimates were limited by inconsistent outcome definitions across studies [4].

Orthopedic surgery offers a somewhat different picture, in part because many orthopedic complications (periprosthetic joint infection, venous thromboembolism, dislocation) have well-characterized risk factors that were already incorporated into earlier scoring systems. A recent review of AI applications in this space described models achieving good discrimination for complications after arthroplasty and spine surgery, but noted that clinical uptake has lagged model development — surgeons are understandably reluctant to change practice based on an algorithm whose internal logic they cannot inspect, and few of these tools have been integrated into intraoperative decision support in a way that changes what actually happens in the operating room [5].

Oncologic surgery has become a fertile area for interpretable modeling in particular, since cancer patients undergoing major operations often carry enough comorbidity and physiologic derangement that risk stratification genuinely changes preoperative counseling and postoperative disposition planning. An explainable machine learning model developed specifically for inpatients with cancer undergoing major operations used SHAP-based feature attribution to let clinicians see which preoperative variables were driving an individual patient's risk estimate, addressing one of the most common objections to black-box prediction — that a number without an explanation is of limited use to a surgeon planning an operation or a patient deciding whether to proceed with one [6]. Complementing purely retrospective validation, a JAMA Network Open study distinguished itself by deploying a machine learning algorithm on a mobile platform and reporting actual performance metrics from that deployment, illustrating both that this technology can move from a research dataset into a usable clinical tool and how much additional validation work that transition requires [7].

MORTALITY PREDICTION

Postoperative mortality is a comparatively rare event in elective surgery, which makes it a harder prediction target statistically even though it is the outcome clinicians care most about avoiding. Several groups have nonetheless built models with useful discrimination for this endpoint by leveraging large multicenter datasets that provide enough positive cases to train on. A model developed and validated

across multiple centers for preoperative prediction of postoperative mortality showed that performance held up reasonably well when the model was applied to populations it had not been trained on, which is one of the more reassuring findings in this literature given how often external validation exposes substantial performance degradation [8]. Similar work restricted to abdominal surgery specifically found that machine learning models incorporating routine preoperative laboratory and vital sign data could stratify mortality risk with discrimination exceeding conventional scores, though absolute event rates remained low enough that calibration — whether predicted probabilities matched observed frequencies — required careful attention [9].

A methodologically distinctive line of work has incorporated natural language processing of clinical notes alongside structured data. A deep neural network model that combined structured EHR variables with NLP-derived features from preoperative documentation found that unstructured text carried predictive signal not captured in structured fields alone — a finding with intuitive appeal, since much of what an anesthesiologist actually assesses during a preoperative evaluation (functional status, frailty, the patient's overall clinical gestalt) is recorded in free text rather than discrete fields [10]. Age-specific modeling has also emerged as its own subfield, with machine learning approaches to postoperative mortality in older adults undergoing emergency surgery showing that risk factors and their relative importance differ meaningfully from elective, younger populations, reinforcing that a single mortality model trained on a general surgical population may not generalize well to the subgroups where the stakes are highest [11].

An influential thread of this work has tried to predict several serious complications from one shared feature set rather than building separate models for each outcome. A deep learning model trained to predict postoperative mortality, acute kidney injury, and reintubation from a single set of preoperative and intraoperative variables demonstrated that a common architecture could achieve good discrimination across genuinely distinct physiologic endpoints, which is practically appealing because it means a single model pipeline embedded in an anesthesia information system could surface multiple risk estimates simultaneously rather than requiring separate infrastructure for each outcome [12].

ACUTE KIDNEY INJURY

Postoperative acute kidney injury has become something of a proving ground for intraoperative, waveform-level machine learning, because its pathophysiology is tightly linked to hemodynamic events (hypotension, low cardiac output, nephrotoxin exposure) that unfold on a minute-to-minute timescale during anesthesia — exactly the kind of continuous, high-frequency data that static preoperative risk scores cannot use. A deep-learning algorithm built on real-time intraoperative vital sign signals from major non-cardiac surgery demonstrated that models updating continuously during the case could flag rising AKI risk before it would be apparent from postoperative creatinine trends alone, raising the possibility of genuinely actionable intraoperative alerts rather than a retrospective risk label [13]. A related dynamic modeling approach similarly used serial postoperative data to update AKI risk severity over time rather than producing a single static prediction at the end of surgery, an approach that better matches how kidney injury actually evolves clinically over the days following an operation [14].

Cardiac surgery, where AKI is both common and closely tied to specific intraoperative exposures such as cardiopulmonary bypass time, has generated its own dedicated modeling literature; a machine learning model built specifically for AKI after off-pump coronary artery bypass grafting found that intraoperative hemodynamic variables added meaningful predictive value beyond preoperative renal function alone, consistent with the broader finding across this subfield that intraoperative data, not just baseline patient characteristics, is where much of the AKI risk signal actually resides [15]. Taken alongside the multi-outcome architecture from Fritz and colleagues, this body of work suggests that AKI prediction specifically benefits from continuous intraoperative monitoring in a way that some other complications, driven more by fixed preoperative risk factors, may not [12].

POSTOPERATIVE DELIRIUM

Delirium prediction has arguably produced the most methodologically mature subset of this literature, partly because delirium is common enough in older surgical populations to generate adequately powered datasets and partly because its diagnosis, while still requiring clinical assessment, is reasonably standardized through instruments like the Confusion Assessment Method. Cardiac surgery populations, where delirium incidence is particularly high, have been a natural starting point: a study developing delirium prediction models specifically in patients undergoing cardiovascular surgery compared several machine learning algorithms and found gradient-boosted tree methods generally outperformed simpler logistic regression baselines, though the margin of improvement was modest enough to raise the usual question of whether the added complexity is worth the loss of interpretability [16]. Work restricted to preoperative electronic health record data — deliberately excluding intraoperative variables — showed that useful delirium risk stratification is achievable before a patient ever enters the operating room, which matters practically because it allows preoperative counseling and prehabilitation planning rather than only intraoperative or postoperative intervention [17].

Non-cardiac surgery has followed a similar trajectory, with a machine learning-based prediction model for postoperative delirium in that broader population finding comparable discrimination to what had been reported in cardiac surgery cohorts, suggesting the underlying preoperative risk factors — age, cognitive impairment, frailty, polypharmacy — generalize reasonably well across surgical specialties even though the absolute incidence of delirium varies considerably between them [18]. A subsequent comparative study evaluating multiple machine learning approaches head-to-head for delirium after non-cardiac surgery reinforced that finding while highlighting meaningful variability in performance depending on which specific algorithm and feature set was used, a reminder that "machine learning" is not a single method but a family of quite different modeling choices that do not all perform equally on the same problem [19]. Most recently, an updated prediction model incorporated a broader set of geriatric and functional variables that earlier models had often omitted, reflecting the field's gradual recognition that delirium risk is driven as much by baseline cognitive and functional reserve as by the surgical insult itself [20]. A systematic review and meta-analysis pooling this now-substantial body of delirium prediction studies concluded that machine learning models offer a real, if incremental, improvement over conventional delirium risk tools, while flagging the same external validation gaps seen throughout this literature more broadly [21].

METHODOLOGICAL LIMITATIONS AND BARRIERS TO CLINICAL ADOPTION

Several problems recur across essentially every complication category surveyed above, and they are worth naming explicitly because they explain why so few of these models have moved beyond publication into routine clinical use. First, external validation remains the exception rather than the rule. A model that discriminates well in the institution where it was trained frequently performs worse, sometimes substantially so, when applied to a different hospital's patient mix, documentation practices, and baseline complication rates — the multicenter mortality validation work is notable precisely because it is uncommon for this literature to test that question at all [8]. Second, outcome definitions are inconsistent enough across studies that pooling results, as the colorectal and delirium meta-analyses both had to do, requires accepting real heterogeneity in what "complication" actually meant from one dataset to the next [4,21]. Third, interpretability remains a genuine barrier rather than a purely academic concern; clinicians are reasonably hesitant to act on a risk score they cannot interrogate, which is exactly the gap that explainable modeling approaches using techniques like SHAP attribution are trying to close, with only partial success so far [6].

A further limitation, less often discussed, is that most of these models have been validated against retrospective outcomes rather than tested in a way that changes clinical decisions and then measures whether that change improved outcomes. The JAMA Network Open mobile deployment is one of the few examples in this literature that moved a model into an actual point-of-care tool and reported on its use there, and even that study measured performance rather than a downstream change in complication rates attributable to the tool's use [7]. Orthopedic AI applications show the same pattern — models with respectable retrospective discrimination that have not yet been shown to change surgical decision-making in a way that measurably reduces complications [5]. Until more of this literature closes that gap between predicting an outcome and demonstrably improving it, machine learning risk models will remain, for most complications, a promising research tool rather than a standard part of perioperative care.

CONCLUSION

Machine learning has demonstrably improved the technical performance of postoperative complication prediction across general morbidity, mortality, acute kidney injury, and delirium, generally outperforming conventional logistic regression-based risk scores on measures of discrimination when tested on held-out data from the same population that generated the training set. The strongest and most clinically promising work has come from applications that exploit data conventional scores cannot use at all — continuous intraoperative vital sign streams for kidney injury, free-text clinical documentation for mortality risk, and broad automatically curated EHR feature sets for general morbidity — rather than simply re-fitting a familiar list of preoperative variables with a fancier algorithm. What the field still lacks, almost uniformly, is external validation across institutions, consistent outcome definitions that would allow honest comparison between models, and prospective evidence that deploying these tools actually changes clinical decisions in a way that reduces complications rather than merely predicting them more accurately after the fact. Closing that gap, more than developing yet another algorithm with marginally better discrimination on a single-center retrospective cohort, is what will determine whether machine learning becomes a routine part of perioperative risk assessment or remains a large and still-growing research literature that clinicians read about more often than they actually use.

DECLARATIONS

Acknowledgements

The authors thank their respective departments for institutional support during the preparation of this review.

Funding

This review received no external funding.

Conflict of Interest

The authors declare no conflicts of interest.

Author Contributions

Pieter Lehmann conceived the review and drafted the initial manuscript. Dimitris Brenner performed the literature search and synthesized the sections on mortality and acute kidney injury prediction. Alessandro Dvorak synthesized the sections on general morbidity and delirium prediction and critically revised the manuscript. All authors reviewed and approved the final version.

Ethics Statement

This is a literature review that did not involve new human or animal data collection and did not require ethics committee approval.

Data Availability

No new data were generated for this review. All sources discussed are cited and publicly available through their respective publishers.

REFERENCES

1. Machine learning to predict postoperative complications after digestive surgery: a scoping review (2023) British Journal of Surgery, 110(12):1646.

2. Corey KM, et al. (2018) Development and Validation of Machine Learning Models to Identify High-Risk Surgical Patients Using Automatically Curated Electronic Health Record Data. Annals of Surgery.

3. The prediction of surgical complications using artificial intelligence in patients undergoing major abdominal surgery: A systematic review (2021) Surgery.

4. Systematic review and meta-analysis of the role of machine learning in predicting postoperative complications following colorectal surgery: how far has machine learning come? (2025) PMC12626580.

5. Artificial Intelligence for Predicting Postoperative Complications in Orthopedics: A Review of Clinical Applications, Challenges, and Future Directions (2025) PMC12840821.

6. Explainable Machine Learning Model to Preoperatively Predict Postoperative Complications in Inpatients With Cancer Undergoing Major Operations (2024) JCO Clinical Cancer Informatics, DOI:10.1200/CCI.23.00247.

7. Performance of a Machine Learning Algorithm Using Electronic Health Record Data to Predict Postoperative Complications and Report on a Mobile Platform (2020) JAMA Network Open.

8. Multi-center validation of machine learning model for preoperative prediction of postoperative mortality (2022) PMC9276734.

9. Machine learning-based prediction of postoperative mortality risk after abdominal surgery (2025) PMID:40291872.

10. Predicting Postoperative Mortality With Deep Neural Networks and Natural Language Processing: Model Development and Validation (2022) PMC9131148.

11. Machine Learning Prediction of Postoperative Mortality in Older Emergency Surgery Patients (2026) Journal of Surgical Research.

12. Fritz BA, et al. (2019) Deep-learning model for predicting postoperative mortality, acute kidney injury, and reintubation using a single feature set. Anesthesiology, PMC7170922.

13. A deep-learning algorithm using real-time collected intraoperative vital sign signals for predicting acute kidney injury after major non-cardiac surgeries: a modelling study (2025) PLOS Medicine, DOI:10.1371/journal.pmed.1004566.

14. A deep learning-based dynamic model for predicting acute kidney injury risk severity in postoperative patients (2023) PMC10683578.

15. Prediction of postoperative acute kidney injury in patients undergoing off-pump coronary artery bypass grafting: A machine learning model (2025) PMC13448026.

16. Development of postoperative delirium prediction models in patients undergoing cardiovascular surgery using machine learning algorithms (2023) Scientific Reports, DOI:10.1038/s41598-023-48418-5.

17. Postoperative delirium prediction using machine learning models and preoperative electronic health record data (2021) BMC Anesthesiology, DOI:10.1186/s12871-021-01543-y.

18. Machine learning-based prediction model for postoperative delirium in non-cardiac surgery (2023) BMC Psychiatry, DOI:10.1186/s12888-023-04768-y.

19. Machine learning models for predicting postoperative delirium after noncardiac surgery: A comparative study (2026) PMID:42219241.

20. Development and Validation of a Prediction Model for Postoperative Delirium (2025) PMC13264316.

21. Machine Learning-Based prediction models for postoperative delirium: a systematic review and meta-analysis (2025) PMC12506401.