

Machine Learning in Perioperative Medicine: A Comparative Review of Predictive, Causal, and Foundation Model Approaches in Surgical Data Science

Enoch Chi Ngai Lim^{a, d} , Chi Eung Danform Lim^{a, b, c} 

Abstract

This review aimed to clarify the methodological landscape of machine learning (ML) in perioperative medicine by comparing three major paradigms: supervised predictive models, causal inference frameworks, and foundation models. It addressed the central research question of which ML approach is most appropriate for specific perioperative clinical scenarios and under what data conditions. A narrative review was conducted using peer-reviewed literature from PubMed, Scopus, and Web of Science (2015–2026). Studies were included if they reported quantitative ML performance in surgical or anesthesia contexts and addressed at least one of the three paradigms. Evidence was synthesized to compare methodological characteristics, performance, interpretability, and clinical applicability. Supervised models (e.g., XGBoost, Random Forest) dominate current practice, demonstrating strong predictive performance but lacking causal interpretability. Causal ML approaches, including meta-learners and doubly robust estimators, effectively capture treatment heterogeneity but rely on strong assumptions and complex validation. The TabPFN foundation model performs well on small datasets without tuning but shows limitations in calibration and regression tasks. Across all paradigms, external validation is limited, and interpretability and generalizability remain key challenges. No single ML paradigm is universally optimal for perioperative applications. Method selection should align with the clinical question and data structure. Future progress depends on multicenter validation, integration of paradigms, and development of regulatory-compliant, interpretable clinical decision-support systems.

Keywords: Machine learning; Perioperative medicine; Causal inference; TabPFN; Surgery; Data science; Treatment heterogeneity

Introduction

More than 300 million surgical interventions are performed worldwide annually [1]. Data-driven technology that manages the complexity and scale of current surgical populations is essential for perioperative risk management, anesthesia safety, and postoperative outcomes. Machine learning (ML) has emerged as the primary technology in this area. The broad range of utilization, documented in numerous systematic reviews, crosses several perioperative domains, including intensive care unit (ICU) admission and postoperative death, organ-specific complications, and anesthesia drug interactions [2, 3]. Despite its growth, the field of ML remains disjointed. Each ML technique has a different inferential perspective: some are outcome- or treatment-effect predictors, and others perform zero-shot inference on small tabular datasets. The lack of structured comparison makes it difficult for researchers and clinicians to identify what is the best technique for a specific clinical question. This review directly investigates that problem.

Three distinct ML paradigms exist in perioperative literature. The most established approach is supervised classification, in which ensemble methods, such as XGBoost and Random Forest, are used to predict binary or categorical outcomes from labeled training data [2–4]. The second is causal ML, which integrates ML with potential outcomes paradigms to derive individual treatment effects from observational data [5]. The third is TabPFN (Tabular Prior-data Fitted Network), a transformer-based foundation model pre-trained on artificial tabular datasets and therefore does not require task-specific training to perform classification [6]. Each paradigm has its own objectives and is suited to different data contexts.

This review evaluates the published evidence on the efficiency and interpretability of the three paradigms, data requirements, and clinical accuracy. It aims to address three questions: what is the best ML paradigm for a specific perioperative clinical query? What data situations justify the use of each approach? What limitations do existing methods have, and what should be the focus of future investigations?

Manuscript submitted May 1, 2026, accepted May 13, 2026
Published online June 10, 2026

^aTranslational Research Department, Specialist Medical Services Group, Earwood NSW 2206, Australia

^bNICM Health Research Institute, Western Sydney University, Westmead NSW 2145, Australia

^cData Science Institute, University of Technology Sydney, Ultimo NSW 2007, Australia

^dCorresponding Author: Enoch Chi Ngai Lim, Translational Research Department, Specialist Medical Services Group, Earwood NSW 2206, Australia. Email: enoch@smsg.au

doi: <https://doi.org/10.14740/aicm27>
AI in Clinical Medicine
2819-7437 (online)

Table 1. Comparative Summary of ML Paradigms in Perioperative Research

Dimension	Traditional supervised ML	Causal ML/meta-learners	TabPFN
Primary purpose	Outcome prediction	Causal effect estimation	Outcome prediction
Sample size requirement	Large preferred	Moderate to large	Small to large
Hyperparameter tuning	Extensive	Moderate	None required
Causal inference capacity	No	Yes	No
Missing data handling	Requires imputation	Requires imputation	Native handling
Interpretability tools	SHAP, EBM	CDT, SHAP	Moderate (SHAP)
Calibration	Variable	Good (doubly robust)	Task-dependent
External validation evidence	Moderate (multicenter studies emerging)	Limited	Early stage
Regulatory readiness (SaMD)	Moderate	Moderate	Early stage

AIPW: augmented inverse probability weighting; CDT: causal distillation tree; EBM: Explainable Boosting Machine; ML: machine learning; SaMD: Software as a Medical Device; SHAP: Shapley Additive explanations; TabPFN: Tabular Prior-data Fitted Network.

Methodology

This narrative review aims to integrate the existing theoretical, methodological, and empirical evidence to evaluate different ML approaches in the perioperative domain. The narrative approach was selected because the area encompasses numerous methodological approaches, such as predictive modeling, causal inference, and the research of foundational models, which do not correlate in a straightforward way to a particular systematic review. A systematic literature search was conducted in PubMed, Scopus, and Web of Science, covering the years 2015 to 2026. The search was performed using the terms “perioperative,” “surgical,” “machine learning,” “causal inference,” “TabPFN,” “tabular foundation model,” and “ICU prediction,” combined by Boolean operators.

Inclusion criteria were as follows: (1) publication of quantitative ML model performance in a surgical or anesthesia context; (2) peer-reviewed studies published in English; and (3) studies addressing at least one of the three paradigms being compared. Studies that were not peer-reviewed and were in preprint form were excluded, unless they provided a unique methodological contribution that had been cited multiple times. Studies classified as grey literature and conference abstracts which did not provide full methods were also excluded.

ML Methods: A Taxonomy With Clinical Illustrations

Supervised classification models

Supervised classification is the most commonly used ML approach in perioperative medicine. Two systematic reviews assessing its extent found that ensemble techniques, especially XGBoost, Random Forest, and gradient boosting, dominate the literature, while logistic regression remains the gold standard [2, 3]. The targets were highly varied in the clinical realm, including ICU admissions, 30-day mortality, postoperative

acute kidney injury, respiratory failure, and surgical site infections. One of the most recent meta-analyses of ML models focused on ICU mortality, covering 123 models from 40 studies and 317,028 patients, and reported a pooled area under the curve (AUC) of 0.83, with age, heart rate, and respiratory rate as the most commonly identified predictors [7].

The training and validation of models involved various databases. These include public repositories such as MIMIC-IV and the eICU Collaborative Research Database, as well as the American College of Surgeons National Surgical Quality Improvement Program (NSQIP), and other national databases, all of which have aided the development of large-scale models. Although the risks of generalizability are well documented, datasets from single-center electronic health records are still widely used. A systematic review conducted using the Prediction Model Risk Of Bias Assessment Tool (PROBAST) revealed that most perioperative ML studies lack external validation, and many use only internal splits of single-center data [3]. Key databases utilized in the wider field are summarized in Table 1. Multi-task deep learning models that have been trained on cohorts of over 50,000 patients on multiple postoperative complications have shown that in addition to preoperative variables, the integration of intraoperative physiological time series records and variables improves prognostic capabilities, compared to the static models [8].

TabPFN is unique in the supervised learning paradigm. It is a transformer-based foundation model trained on millions of synthetic tabular datasets [6]. It has no task-specific training or hyperparameter tuning. Instead, it relies on in-context learning. It has been benchmarked against tree-based models and shown to perform comparably to other models on datasets with sample sizes less than 10,000 observations. Performance, though, has been shown to decline in the original architecture [6, 9]. Its severe class imbalance calibration remains a concern and is often a defining characteristic of surgical outcome datasets. TabPFN is also advantageous for its speed. For typical tabular classification tasks, the model performs inference in under 1 s. This speed makes it ideal for preliminary exploratory modeling in a surgical research setting where datasets are scarce [6].

Causal inference and meta-learner frameworks

Because supervised classification relies on estimating statistical correlations, it has limitations in its applicability. A patient subgroup with a predicted higher risk of mortality may not receive a particular intervention, as the prediction and the decision to treat are distinct inferential questions. Causal ML frameworks address this challenge by estimating counterfactual outcomes: what would have happened to an individual patient under a different treatment while keeping all other conditions the same [5]? This is particularly important for the field of perioperative medicine as treatment strategies can vary significantly across patients and institutions. The same can be said for the use of vasopressors, anesthesia, intravenous fluids, and surgical techniques.

Applications of the framework using the T-Learner, X-Learner, and the doubly robust augmented inverse probability-weighted (AIPW) estimator in clinical datasets have served as early examples of meta-learner algorithms [5]. These techniques assess the likelihood of treatment due to covariates, then model outcome variance between treatment groups to individualize treatment effects. Given the combined structural flexibility of random forests and local average treatment effect estimation, causal forests have also been used in surgery to target subgroups of patients who respond differently to the same treatment [5]. The treatment-effect subgroups identified by these methods have not been detectable using standard prediction algorithms in studies of the effects of vasopressors on renal function and the dose-response relationships of anesthetic drugs [10, 11]. The AIPW method's doubly robust property allows the use of outcome and propensity score models that are less than perfect, making it highly applicable to confounded, high-dimensional datasets common in perioperative studies [5].

These methods depend on the assumptions of unconfoundedness, positivity, and the stable unit treatment value assumption (SUTVA). Although none of these assumptions can be definitively substantiated with observational data, analysts have used propensity score overlap and sensitivity analyses to assess their validity. The inverse of the observational study paradigm's constraints is present in surgical observational studies, such as residual confounding from unmeasured intraoperative variables (e.g., anesthetic depth) and variability in fluid hemodynamic balance, which are fundamental limitations of these studies.

Interpretability tools

The integration of clinical ML is a complex process that considers factors beyond predictive accuracy. For ML models to be incorporated into clinical practice, they must be interpretable and their outputs trustworthy. Clinicians' understanding of model outputs is a critical factor in the implementation process. In clinical ML, the most frequently used explainability tool is the Shapley additive explanations (SHAP) model, which accounts for 46.5% of explainable AI applications in the healthcare sector [12]. SHAP values are derived from cooperative game theory and can be used to assign both global feature importance val-

ues and individual patient prediction ascriptions. SHAP values quantify the individual contribution of each feature to a particular prediction [12]. However, SHAP explanations may become unstable in highly correlated feature spaces and can be computationally intensive for complex models, highlighting an important trade-off between interpretability and scalability in clinical ML applications. In perioperative care, the SHAP value method has been used in multiple studies [8, 10, 11] to identify predictive factors for surgical outcomes, including preoperative renal function, the American Society of Anesthesiologists (ASA) physical status, and certain metabolic parameters. In randomized clinical decision support studies, researchers observed a marked positive response among participants to tools that incorporated structured clinical reasoning and SHAP model outputs. This reinforces the idea that model outputs must be interpretable and easily translatable into clinical practice.

Causal distillation trees (CDTs) enable the extension of interpretability to causal models. Specifically, CDTs simplify complex causal forests into causal models that can be structured as easily understood decision trees or causal trees, making decisions transparent and interpretable. Through these methods, individual-level causal results can be communicated to clinicians without advanced knowledge of causal ML. The Explainable Boosting Machine (EBM) also offers additive global interpretability, affecting the contributions of individual features and their pairwise interactions. This model also has the benefit of being easily understood without losing important predictive measures [11].

Using TabPFN in health economic modeling

An innovative application of TabPFN is emerging in the field of surgical health economics [13]. Traditional budget impact models and cost-effectiveness analyses use a fixed, population-based approach to costs and outcomes that does not factor in variability among patients or evolving patterns of clinical practice. Due to TabPFN's approach to missing data, its ability to quantify uncertainty and posterior probabilities, and its overall empirical performance on small datasets, it is well-suited for use in patient-level economic modeling and dynamic budget impact scenarios, and in surgical settings for real-time resource allocation [13]. These applications are still in the conceptual phase and require further empirical research. Nevertheless, it emphasizes the potential of tabular foundation models to reach beyond classification and to expand into areas where impactful decision-making processes are driven by uncertainty.

Method Comparison

While supervised classification, as a form of ML, is the most developed in terms of perioperative applications, most of its success is realized within single-center contexts. Regarding large, curated datasets, their success is well-documented, and they form the basis for most clinical risk and decision-support models [2, 3]. While causal ML frameworks may provide ac-

tionable insights on treatment heterogeneity that purely predictive models do not, they require larger sets and more strategic design [5]. For example, in treatment individualization, it is more relevant that two patient subsets exhibit opposing renal responses to the same vasopressor, rather than predicting only acute kidney injury (AKI) [10]. Compared to traditional tuning pipelines, TabPFN is more advantageous for unconventional data-scarce environments [6, 9]. However, its classification-first structure, poor calibration, and insufficient task selection can make it more challenging to determine appropriate tasks for its use.

The end-user application of all three approaches remains questionable. As noted in the systematic review [3], most perioperative ML studies rely on single-center datasets, resulting in a lack of external validation. Emerging research on multicenter validation studies has shown that multitask tree-based models for predicting AKI, respiratory failure, and in-hospital mortality have maintained predictive ability across diverse geographic and temporal cohorts. Extending this principle to causal and foundation model paradigms is an immediate research priority. Table 1 compares the three paradigms across key methodological dimensions relevant to perioperative research.

Methodological Strengths and Limitations

Of the three paradigms, causal ML frameworks have the most robust capacity for clinical inference. The doubly robust property of AIPW estimators mitigates the effects of model misspecification, whether in the propensity score model or the outcome model [5]. This unique feature makes causal ML particularly beneficial for the complex, often confounding observational datasets endemic to perioperative research. The individual-level treatment effect heterogeneity that causal ML can capture provides the basis for precision medicine in a way that solely population-level associations cannot.

Considerable strengths of TabPFN are the computational speed and ease of use. The synthetic pre-training feature eliminates the need for large-scale, task-specific data procurement, and the privacy features protect against the disclosure of patient data during model training [6]. However, the classification-first design is a limitation for regression, which is critical in cost modeling and in assessing outcomes measured on a continuum in surgical health economics. Many surgical datasets have low-incidence outcomes, and thus, poor calibration limits the model's direct clinical use.

There are several limitations shared by all three paradigms. External validation on independent, multi-center datasets is rare, and the applicability of single-center perioperative datasets to a large, multi-center surgical population is very limited [2, 3]. Inadequately addressed in the literature are treatment variables, which are usually restricted to binary definitions that ignore important clinical components, such as dose, timing, and infusion rate. Class imbalance in outcome variables, such as in-hospital mortality (which is less than 2% in large surgical databases), is a common problem and negatively affects recall across all three paradigms. Pro-

spective validation and studies that investigate real-world applications are lacking.

Emerging Trends and Opportunities

With new publicly available perioperative datasets, ML researchers can now explore new, previously unavailable opportunities. Single-center datasets can now be surpassed by large registries, critical care repositories, and multi-institutional datasets. This extension of datasets provides new opportunities for perioperative forecast models but also requires greater methodology and standardized reporting improvements. The new tabular data foundation models are also expanding, with updated TabPFN models designed for regression and expanded practical sample-size limits to include increasingly continuous perioperative outcomes [6]. However, these regression-capable implementations remain largely developmental and have not yet undergone substantial clinical validation in perioperative settings. New multi-modal models are designed for perioperative prediction by incorporating tabular clinical data, intraoperative physiological waveforms, images, and even free-text clinical notes [4]. These models are designed to include relevant temporal contextual data that older models lack.

Increasing interest is being directed to federated learning and its potential to address the generalizability issue in single-center surgical datasets. By collaboratively training models across multiple sites while keeping patient data confidential, federated learning creates generalizable models while still protecting participant data [14]. Since 2020, there has been an exponential increase in publications of federated learning in healthcare, with 612 publications identified by 2023 [14]. Federated learning allows each institution to maintain data confidentiality while still yielding robust discriminatory models comparable to those of centralized models.

There is progress in establishing more complex, robust regulatory frameworks. The US Food and Drug Administration (FDA) Software as a Medical Device Action Plan has concrete demands for transparency, reproducibility, and post-marketing surveillance of adaptive ML algorithms [15]. Equivalent guidance from the Australian Therapeutic Goods Administration emphasizes explainability and human oversight [16]. These frameworks will shape the pathway for deploying the ML "paradigms" discussed above, especially regarding the foundation model and causal paradigm, for which the evidence is still in the early stages.

Table 2 illustrates the breadth of publicly available perioperative databases now accessible to surgical ML researchers. The field is no longer limited to single-center datasets: large registries, critical care repositories, and multi-institutional platforms now support model development at a scale that was previously unavailable. This expansion creates new opportunities for all three ML paradigms but also raises new demands for methodological rigor and standardized reporting.

Future Research Directions

This review has highlighted several priorities. The most immedi-

Table 2. Representative Perioperative Databases Used in Surgical ML Research

Database	Source/country	Scale	Data types	Typical ML use cases
MIMIC-IV	Beth Israel Deaconess Medical Center, USA	About 300,000 ICU admissions	Vitals, labs, medications, notes	ICU mortality, sepsis prediction, AKI
NSQIP	American College of Surgeons, USA	About 1 million cases/year	Preoperative risk factors, 30-day outcomes	Surgical complication risk scoring
VitalDB	Seoul National University Hospital, Korea	6,388 cases	High-resolution intraoperative signals, drugs	Vasopressor effects, intraoperative monitoring
MOVER	UC Irvine Medical Center, USA	67,134 cases	Demographics, labs, vitals, medications	Anesthetic dose-response, ICU risk
eICU CRD	Philips Healthcare, multi-site USA	About 200,000 ICU stays	Clinical assessments, vitals, labs	Critical care prediction, mortality
National registries (e.g., STS, DSSR)	Multi-institutional, various countries	Millions of records	Procedural data, outcomes, comorbidities	Cardiac surgery outcomes, risk scoring

ML: machine learning; AKI: acute kidney injury; eICU CRD: eICU Collaborative Research Database; ICU: intensive care unit; MIMIC-IV: Medical Information Mart for Intensive Care IV; MOVER: Medical Informatics Operating Room Vitals and Events Repository; NSQIP: National Surgical Quality Improvement Program; STS: Society of Thoracic Surgeons; DSSR: Danish Society for Patient Safety Registry.

ate need across all paradigms is multicenter external validation. Within the PROCAST framework, analyses of perioperative ML studies have consistently identified this as the most “attributable” risk of bias and the most significant impediment to the uptake of these studies in practice [3]. For causal analyses, especially regarding the use of vasopressors, anesthetics, and fluid therapy, prospective interventional studies to assess the effects of ML perioperative guidance are recommended. This will allow the field to develop concrete evidence, rather than the contradictory evidence a hypothesis may present. The establishment of standardized benchmarks in surgical ML will enable “fair” methodological evaluation across “competing” research groups. Using TabPFN as a nuisance function estimator for doubly robust causal learners in causal inference frameworks is desirable from a methodological standpoint and may allow the integration of what are typically seen as opposing paradigms. This approach has not yet been fully developed, but it has the potential to be a robust integration of what is typically seen as opposing horizons.

Conclusions

Although ML methods now play an increasingly important role in perioperative research, the current landscape of the three major paradigms remains fragmented, with each serving different inferential goals and data structures. With large surgical datasets, supervised classification models can provide reliable, characterized performance but are unable to address questions of causality. Causal meta-learner frameworks, which are directly applicable to precision perioperative care, can reveal individualized treatment-effect heterogeneity but are challenging and sensitive to unobserved confounding. For small and medium-sized surgical datasets, TabPFN is an easy and resource-light alternative to traditional pipeline tuning, but its clinical utility is limited in terms of calibration and regression functionality. There is no single approach that can be fully applied to all perioperative contexts. The

next frontier for this domain is the integration of these paradigms into federated, regulatory-compliant clinical systems, supported by multicenter validation and prospective clinical studies. Ultimately, the real test is not model accuracy alone, but whether these tools meaningfully improve patient care at the bedside.

Acknowledgments

None to declare.

Financial Disclosure

None to declare.

Conflict of Interest

None of the authors declared financial and non-financial relationships, activities, or any conflicts of interest regarding this manuscript.

Informed Consent

Non applicable.

Author Contributions

ECNL: conceptualization, data curation, methodology, formal analysis, investigation, project administration, resources, software, writing – original draft, writing – review and editing. CEDL: conceptualization, resources, visualization, software, funding, validation, supervision, writing – review and editing.

Data Availability

The authors declare that data supporting the findings of this study are available within the article.

Artificial Intelligence Use Declaration

The authors used Grammarly (writing assistance software) for language editing and proofreading. No content was generated by artificial intelligence, and the authors take full responsibility for the accuracy and integrity of the manuscript.

References

1. Weiser TG, Regenbogen SE, Thompson KD, Haynes AB, Lipsitz SR, Berry WR, Gawande AA. An estimation of the global volume of surgery: a modelling strategy based on available data. *Lancet*. 2008;372(9633):139-144. [doi pubmed](#)
2. Bellini V, Valente M, Bertorelli G, Pifferi B, Craca M, Mordonini M, Lombardo G, et al. Machine learning in perioperative medicine: a systematic review. *J Anesth Analg Crit Care*. 2022;2(1):2. [doi pubmed](#)
3. Arina P, Kaczorek MR, Hofmaenner DA, Pisciotto W, Refinetti P, Singer M, Mazomenos EB, et al. Prediction of complications and prognostication in perioperative medicine: a systematic review and PROBAST assessment of machine learning tools. *Anesthesiology*. 2024;140(1):85-101. [doi pubmed](#)
4. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347-1358. [doi pubmed](#)
5. Feuerriegel S, Frauen D, Melnychuk V, Schweisthal J, Hess K, Curth A, Bauer S, et al. Causal machine learning for predicting treatment outcomes. *Nat Med*. 2024;30(4):958-968. [doi pubmed](#)
6. Hollmann N, Muller S, Purucker L, Krishnakumar A, Korfer M, Hoo SB, Schirrmeyer RT, et al. Accurate predictions on small data with a tabular foundation model. *Nature*. 2025;637(8045):319-326. [doi pubmed](#)
7. Sun H, Kang M, Zhang H, Jia J, Wang Q. Machine learning for predicting mortality in intensive care unit patients: a prognostic performance systematic review and meta-analysis. *Nurs Crit Care*. 2025;30(6):e70206. [doi pubmed](#)
8. Shickel B, Loftus TJ, Ruppert M, Upchurch GR, Jr., Ozrazgat-Baslanti T, Rashidi P, Bihorac A. Dynamic predictions of postoperative complications from explainable, uncertainty-aware, and multi-task deep neural networks. *Sci Rep*. 2023;13(1):1224. [doi pubmed](#)
9. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? *Adv Neural Inf Process Syst*. 2022;35:507-520. [doi](#)
10. Cui J, Lim ECN, Wu X, Lim CED. Phenylephrine's heterogeneous renal effects in non-cardiac surgery: a causal machine learning study. *Cureus J Comput Sci*. 2025;2:es44389-025-00018-2. [doi](#)
11. Cui J, Li W, Lim ECN, Wu X, Lim CED. Stratified causal inference for intensive care unit risk prediction: informatics-based modeling of anesthetic drug combinations. *JMIR Form Res*. 2026;10:e80294. [doi pubmed](#)
12. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell*. 2020;2(1):56-67. [doi pubmed](#)
13. Lim ECN, Lim CED. Redefining surgical health economics: the potential of TabPFN for real-time precision modelling. *J Comput Commun*. 2025;13:30-40. [doi](#)
14. Teo ZL, Jin L, Li S, Miao D, Zhang X, Ng WY, Tan TF, et al. Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture. *Cell Rep Med*. 2024;5(2):101419. [doi pubmed](#)
15. U.S. Food and Drug Administration. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. 2021. Available from: <https://www.fda.gov/media/145022/download>.
16. Therapeutic Goods Administration. Understanding how we regulate software-based medical devices. 2026. Available from: <https://www.tga.gov.au/resources/guidance/understanding-how-we-regulate-software-based-medical-devices>.