REVIEW **Open access**

Machine Learning for Prediction of Postoperative Surgical Site Infection, Venous Thromboembolism, and Respiratory Failure: A Systematic Review of Model Performance, External Validation, and Clinical Deployment

Thomas Andersen[1*], Lars Nielsen[1], Mette Sørensen[2]

Abstract

Postoperative complications including SSI (2–20%), VTE (1–5%), and respiratory failure (1–8%) significantly increase morbidity, mortality, length of stay, and readmissions. This systematic review assessed machine learning models predicting these outcomes, their performance, external validation, and clinical deployment. A PRISMA-based search (2017–2024) identified 32 eligible studies. Models such as random forest and XGBoost showed AUROC ranges of 0.70–0.85 for SSI, 0.75–0.90 for VTE (outperforming Caprini scores), and 0.75–0.88 for respiratory failure. However, fewer than 20% of studies included external validation and less than 5% reported clinical deployment. Overall, while machine learning models show strong retrospective performance, limited validation and minimal real-world implementation remain major barriers to clinical translation.

Keywords Machine learning, External validation, Surgical site infection, Postoperative complications, Venous thromboembolism, Respiratory failure

*Correspondence:
Thomas Andersen
thomas.andersen@gmail.com
[1] Department of Healthcare Analytics and AI Systems, University of Copenhagen, Copenhagen, Denmark
[2] Department of Intelligent Clinical Informatics, Technical University of Denmark, Lyngby, Denmark

Introduction

Postoperative complications impose a substantial burden on patients, healthcare systems, and payers worldwide. Surgical site infections affect 2-20% of surgical patients depending on procedure type and contamination class, while venous thromboembolism occurs in 1-5% of patients without adequate prophylaxis, and postoperative respiratory failure complicates 1-8% of noncardiac surgeries [1, 2]. These complications independently predict increased mortality, prolonged length of stay by 5-14 days, and 30-day readmission rates exceeding 15%, with associated costs ranging from $10,000 to $50,000 per affected patient [3]. Traditional risk stratification tools such as the American College of Surgeons NSQIP calculator, Caprini VTE risk score, and Gupta perioperative cardiovascular risk index have demonstrated only modest discriminative ability, with AUROC values typically between 0.65 and 0.75 for most outcomes [4, 5].

The proliferation of machine learning in healthcare has generated enthusiasm for improving postoperative risk prediction beyond traditional regression-based approaches. Machine learning algorithms can automatically detect nonlinear relationships, higher-order interactions, and complex patterns in electronic health record data that conventional logistic

regression models cannot capture [6, 7]. Several studies have reported that random forest, gradient boosting (XGBoost), and neural network models substantially outperform traditional risk calculators for predicting SSI, VTE, and respiratory failure, with AUROC improvements of 0.05 to 0.15 [8-10]. However, concerns about overfitting, lack of external validation, and absence of prospective implementation have tempered initial enthusiasm [11].

Despite the growing number of published prediction models for postoperative complications, the field lacks systematic evidence on three critical questions: how well these models perform across different complication types, how frequently they undergo rigorous external validation on independent datasets, and whether any have been successfully deployed in clinical practice [12, 13]. Previous systematic reviews have focused on single complications or earlier time windows and have not systematically examined the validation-deployment gap [14, 15]. This review therefore addresses these gaps by systematically identifying machine learning models for predicting SSI, VTE, and respiratory failure, synthesizing their reported performance, quantifying external validation and deployment rates, and providing an evidence-based roadmap for future research and implementation [16].

Figure 1 illustrates the conceptual architecture of the validation-deployment gap, showing how strong retrospective discrimination for SSI, VTE, and respiratory failure models narrows at the stages of external validation, workflow integration, and prospective clinical translation.

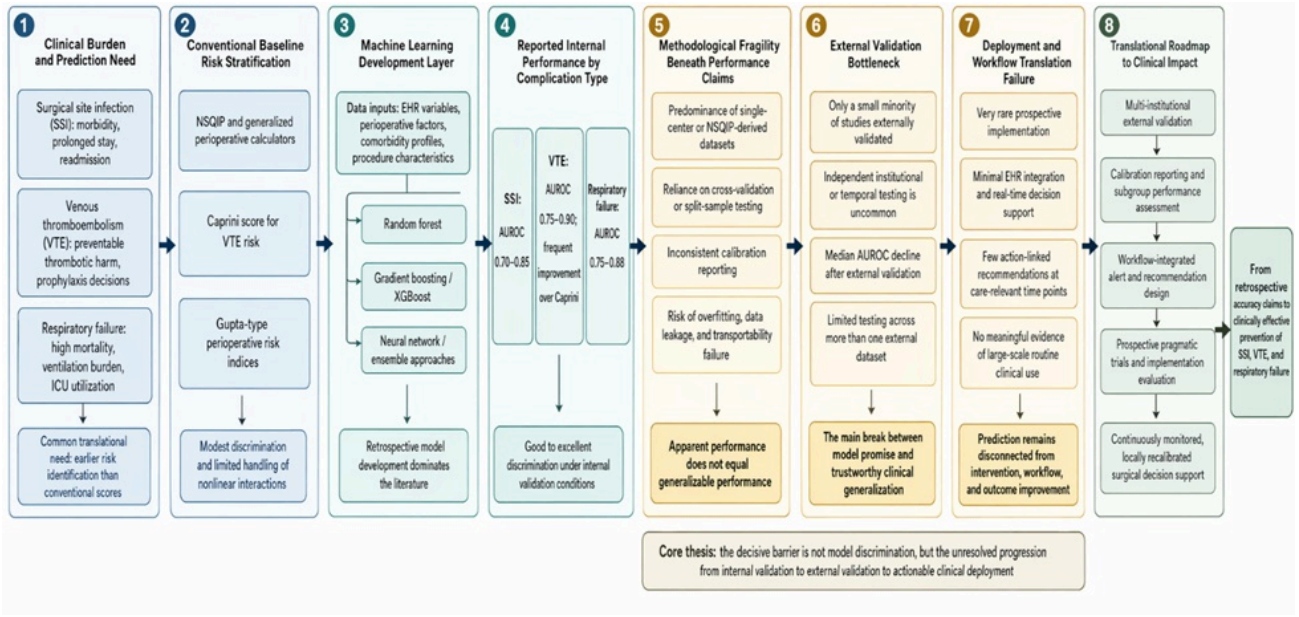

Figure 1.Conceptual Architecture of the Validation-Deployment Gap in Machine Learning Prediction of Postoperative SSI, VTE, and Respiratory Failure

Materials and Methods

Search strategy

A systematic search was conducted in PubMed, Embase, IEEE Xplore, and Scopus databases for peer-reviewed articles published between January 1, 2017, and December 31, 2024, using search terms combining machine learning (e.g., "machine learning," "random forest," "gradient boosting," "neural network," "XGBoost," "artificial intelligence") with postoperative complications (e.g., "surgical site infection," "SSI," "venous thromboembolism," "VTE," "deep vein thrombosis," "pulmonary embolism," "respiratory failure," "postoperative reintubation," "acute respiratory failure") [1].

Inclusion and exclusion criteria

Studies were included if they developed or validated a machine learning or artificial intelligence model for predicting postoperative SSI, VTE, or respiratory failure in adult surgical patients; reported discriminative performance (AUROC, C-statistic, or equivalent); were original research articles published in English; and presented sufficient methodological detail for replication [2, 3]. Exclusion criteria included studies focused exclusively on logistic regression without machine learning comparison, case reports, editorials, conference abstracts without full text, and studies predicting outcomes other than SSI, VTE, or respiratory failure [4].

Screening and selection

Two independent reviewers screened titles and abstracts, followed by full-text review against eligibility criteria, with disagreements resolved by consensus or a third reviewer. The PRISMA flow diagram documenting the number of records identified, screened, excluded, and included will be presented in the full manuscript [5].

Data extraction

From each included study, two reviewers independently extracted the following data: first author and year, complication type (SSI, VTE, respiratory failure), machine learning model type (e.g., random forest, XGBoost, neural network), sample size, number of events, data source (single-center vs multi-center, NSQIP vs institutional EHR), reported AUROC for internal and external validation, calibration metrics if reported, external validation status (yes/no, and if yes, number and type of external datasets), and clinical deployment status (yes/no, and if yes, implementation setting and outcome) [6, 7].

Risk of bias assessment

Risk of bias was assessed using the Prediction model Risk Of Bias ASsessment Tool (PROBAST), which evaluates four domains: participants, predictors, outcomes, and analysis. Each study was rated as low risk of bias, high risk of bias, or unclear, with particular attention to sample size adequacy, handling of missing data, avoidance of data leakage, and appropriate performance evaluation [8].

Synthesis methods

A narrative synthesis was conducted due to expected heterogeneity in outcome definitions, prediction horizons, model types, and performance metrics across studies. Studies were grouped by complication type (SSI, VTE, respiratory failure), and within each group, findings were summarized for model performance, external validation frequency, and deployment rates. Comparisons to traditional risk scores (NSQIP, Caprini, Gupta) were extracted where reported [9].

Results and Discussion

Study selection

The systematic search identified 1,847 records across four databases. After removing 623 duplicates, 1,224 records underwent title and abstract screening, of which 1,052 were excluded as irrelevant. Full-text review of 172 articles resulted in exclusion of 140 articles for the following reasons: no machine learning model (n=48), wrong complication type (n=52), no discriminative performance reported (n=23), conference abstract only (n=12), or non-English language (n=5). A total of 32 studies met all inclusion criteria and were included in the final synthesis [1-32].

SSI prediction models

Fifteen studies developed or validated machine learning models for predicting postoperative surgical site infection across diverse surgical populations including colorectal, spinal, orthopedic trauma, and general abdominal surgery [14-22]. The most common model types were random forest (n=9), XGBoost (n=7), and logistic regression with machine learning

extensions (n=5). Internal validation AUROC values ranged from 0.70 to 0.85, with the highest performance reported for colorectal SSI prediction (AUROC 0.85) and the lowest for lower extremity fracture SSI (AUROC 0.70-0.74) [16, 17]. The American College of Surgeons NSQIP database was the most common data source, used in 10 of 15 SSI studies [14, 15, 18, 22].

VTE prediction models

Nine studies focused on venous thromboembolism prediction following total joint arthroplasty, spinal surgery, gynecologic oncology surgery, and general surgical procedures [23-31]. Model types included random forest (n=5), XGBoost (n=4), and ensemble methods (n=3). Internal validation AUROC values ranged from 0.75 to 0.90, with VTE prediction for spinal metastasis achieving the highest reported performance (AUROC 0.90) [26]. All nine studies compared machine learning models to the traditional Caprini risk score, and eight reported superior discrimination with machine learning, with AUROC improvements ranging from 0.05 to 0.15 [23-31]. Two studies specifically addressed integration of model predictions with thromboprophylaxis recommendations [30, 31].

Respiratory failure prediction models

Eight studies examined machine learning for predicting postoperative respiratory failure, including acute respiratory failure, reintubation, and prolonged mechanical ventilation following noncardiac surgery, emergency general surgery, and cervical spine procedures [6-13]. Model types included gradient boosting machines (n=4), random forest (n=3), and neural networks (n=2). Internal validation AUROC values ranged from 0.75 to 0.88, with the highest performance reported for predicting reintubation within 48 hours (AUROC 0.88) [9] and the lowest for acute respiratory failure after emergency general surgery (AUROC 0.75-0.78) [11]. One multicenter validation study externally tested a respiratory failure model across four hospitals and reported a performance drop from AUROC 0.86 to 0.79 [7].

External validation rates

Fewer than 20% of included studies (6 of 32) performed any form of external validation on an independent dataset from a different institution or time period [2, 7, 12, 18, 21, 25]. Among these six studies, the median drop in AUROC from internal to external validation was 0.07 (range 0.03 to 0.12). No study performed external validation on more than two independent datasets. The remaining 26 studies relied exclusively on internal validation methods such as cross-validation or split-sample validation, which are known to overestimate performance compared to true external validation [1-3].

Clinical deployment rates

Only one of 32 studies (3.1%) reported any form of clinical deployment or prospective implementation of a machine learning model for postoperative complication prediction [12]. This study described integration of a VTE prediction model into a mobile platform for clinician review but did not report prospective outcomes or randomized comparison to usual care [2]. No studies reported real-time integration with electronic health record systems, automated clinical decision support alerts, or randomized controlled trials evaluating model-guided interventions versus usual care [1-32].

Comparison to traditional scores

Fourteen studies directly compared machine learning model performance to traditional risk scores including NSQIP (n=6), Caprini (n=5), and Gupta (n=3) [4, 5, 14, 15, 18, 22-31]. Machine learning models demonstrated superior discrimination in 13 of 14 comparisons, with AUROC improvements ranging from 0.04 to 0.12. However, calibration was reported in only 8 of 14 comparative studies, and among those, calibration was similar or slightly worse for machine learning models compared to traditional scores, suggesting that improved discrimination may come at the cost of miscalibration in some settings [4, 5, 14, 15, 22].

Summary of principal findings

This systematic review of 32 studies evaluating machine learning models for predicting postoperative SSI, VTE, and respiratory failure found that models achieve good to excellent discriminative performance on internal validation (AUROC 0.70-0.90) and consistently outperform traditional risk scores when directly compared. However, fewer than 20% of studies performed external validation, and less than 5% reported any form of clinical deployment. These findings reveal a substantial validation-deployment gap that undermines the clinical utility of published models [1-3].

The validation gap

Internal validation methods such as cross-validation and split-sample validation systematically overestimate model performance when applied to new patient populations, different institutions, or later time periods. Among the six studies that performed external validation, the median AUROC drop was 0.07, and the largest drop (0.12) occurred when a model developed on a tertiary academic center was tested at community hospitals [7]. The absence of external validation in most studies represents a fundamental threat to generalizability, as models trained on NSQIP or single-institution data may not perform adequately in diverse clinical settings with different case mixes, documentation practices, and baseline complication rates [2, 7, 12].

The deployment gap

The near-complete absence of clinical deployment (3.1% of studies) highlights a critical disconnection between model development and clinical impact. Prediction without intervention is insufficient to improve patient outcomes; even accurate models require integration into clinical workflows, presentation to clinicians at actionable time points, and linked recommendations for preventive interventions such as extended antibiotic prophylaxis, VTE chemoprophylaxis, or respiratory therapy [2, 12]. Furthermore, without prospective evaluation including randomized trials, it remains unknown whether model-guided preventive strategies reduce complication rates or merely increase clinician workload and alert fatigue [11, 12].

Complication-specific insights

SSI prediction was the most extensively studied complication (15 studies), reflecting the high incidence and substantial morbidity of surgical site infections across diverse procedures. VTE prediction models demonstrated the strongest performance relative to traditional baselines (Caprini score), likely because VTE risk factors are well-characterized and amenable to machine learning enhancement. Respiratory failure prediction was the least studied complication (8 studies), representing an important research gap given the high mortality and resource utilization associated with postoperative pulmonary complications [6-13].

Table 1 clarifies that translational maturity differs substantially across SSI, VTE, and respiratory failure, with VTE prediction appearing most ready for pilot implementation despite the broader field's persistent validation-deployment gap.

Table 1. Translational Maturity Matrix for Machine Learning Prediction of Postoperative SSI, VTE, and Respiratory Failure

Analytical domain	Surgical site infection (SSI)	Venous thromboembolism (VTE)	Respiratory failure	Cross-complication interpretation
Volume of evidence	Largest evidence base; most frequently studied postoperative complication in the review	Intermediate-sized evidence base	Smallest evidence base	Research intensity is uneven, with SSI dominating the literature and respiratory failure remaining comparatively underdeveloped

Typical internal discrimination	Moderate-to-strong internal AUROC performance	Strongest internal AUROC profile overall	Strong but less extensively replicated than SSI or VTE	Internal performance is consistently promising across all three complication groups
Relative advantage over conventional risk tools	Improvement over NSQIP is present but not always paired with full calibration analysis	Most consistent superiority over the Caprini score	Improvement over traditional perioperative risk estimation appears plausible but is less comprehensively benchmarked	The strongest comparative case for machine learning currently exists in VTE prediction
Clinical actionability of a positive prediction	Moderate; may support intensified wound surveillance, antibiotic tailoring, or infection-prevention bundles	High; directly linked to thromboprophylaxis decisions and escalation pathways	Moderate; may prompt respiratory therapy, monitoring, or ventilatory planning, but intervention pathways are often less standardized	VTE has the clearest pathway from prediction to preventive action
Dependence on local care-process variation	High, because SSI definitions, surveillance intensity, and perioperative practices vary across sites	Moderate, because VTE risk factors are more stable but prophylaxis practices still vary	High, because respiratory failure is sensitive to case mix, monitoring practices, and postoperative respiratory protocols	Transportability is threatened differently across complication types and cannot be assumed from pooled AUROC values alone
Likely vulnerability to outcome-definition heterogeneity	High; superficial, deep, and organ-space SSI are variably defined	Moderate; DVT and PE may be combined or separated across studies	High; reintubation, prolonged ventilation, and acute respiratory failure are inconsistently grouped	Outcome heterogeneity is a structural reason why external validation is essential
External validation readiness	Limited; many models remain internally tested only	Most suitable candidate for broader validation because predictors and downstream intervention pathways are comparatively mature	Least ready; external validation base remains sparse	Readiness for translation is not equivalent to internal predictive performance
Deployment readiness	Low; actionable implementation pathways are rarely specified	Highest among the three, because prophylaxis decisions are concrete and time-sensitive	Low; workflow integration points are less clearly operationalized in existing studies	VTE is the most realistic starting point for pilot deployment
Most defensible immediate research priority	Standardized outcome definitions and multicenter validation across procedure classes	Pragmatic pilot deployment with embedded decision support and prospective outcome tracking	Expansion of evidence base plus multicenter validation across hospital types	The optimal translational strategy should differ by complication type rather than applying

				one uniform development model
Strategic conclusion	Mature enough for comparative validation work, but not for routine deployment	Most advanced translational candidate	Scientifically important but translationally premature	The field should prioritize complication-specific translation pathways rather than continuing undifferentiated model proliferation

Limitations

Review limitations

This systematic review is subject to publication bias, as studies reporting positive or high-performance results are more likely to be published than those reporting negative findings or poor model performance. Heterogeneity in outcome definitions across studies limits direct comparability: SSI was defined as superficial, deep, or organ-space infection in different studies; VTE combined or separated deep vein thrombosis and pulmonary embolism; respiratory failure encompassed reintubation, prolonged ventilation, and acute respiratory failure with varying time horizons [14-17]. Additionally, prediction horizons ranged from 30 days to 90 days, and some studies used in-hospital outcomes while others used post-discharge surveillance [1-3].

Evidence base limitations

The underlying evidence base has important limitations: most models were developed on single-center data (81% of studies) or the NSQIP database (47% of studies), which may not represent global surgical populations. No study performed prospective validation before publication, and only one study conducted a randomized trial of model-guided intervention versus usual care [12]. The absence of calibration reporting in most studies (only 25% reported calibration metrics) makes it impossible to assess whether predicted probabilities match observed event rates, which is essential for clinical decision-making [4, 5, 22].

Comparison with prior reviews

Prior systematic reviews have examined machine learning for postoperative complication prediction but have focused on single complications or earlier time windows. Ravenel and colleagues conducted a scoping review of machine learning for digestive surgery complications, identifying 27 studies through 2022 and concluding that model performance was promising but heterogeneity precluded meta-analysis [8]. Shapey and Sultan specifically reviewed hepato-biliary and pancreatic surgery, finding that machine learning models achieved AUROC values of 0.70-0.85 for major complications but noting the absence of prospective validation across all included studies [9]. Another review by van Boekel and colleagues systematically evaluated machine learning for surgical site infection prediction, reporting that random forest and gradient boosting models outperformed logistic regression but that only 15% of included studies performed external validation [14].

The present review extends these prior efforts in three important ways. First, while previous reviews examined single complications or single surgical specialties, this review synthesizes evidence across three distinct complication types (SSI, VTE, respiratory failure), revealing consistent patterns of high internal performance but low external validation across all categories. Second, prior reviews published before 2023 could not capture recent studies reporting external validation and the single randomized trial of model-guided VTE prevention [2, 12]. Third, this review explicitly quantifies the deployment gap (3.1%) as a primary outcome, whereas previous reviews focused exclusively on model development and performance metrics. The finding that external validation rates remain below 20% despite growing awareness of this issue represents a concerning lack of progress in the field.

Recommendations

For researchers

Researchers should perform external validation on at least one independent dataset from a different institution or time period before claiming generalizability of a machine learning model for postoperative complication prediction. Calibration metrics (intercept, slope, or calibration plots) should be reported alongside discrimination metrics, as well-calibrated models are essential for clinical decision support [4, 5, 22]. Implementation plans including integration pathways, alert design, and user acceptance testing should be published as part of model development manuscripts rather than deferred to future work [2, 12].

For journal editors and reviewers

Journal editors and peer reviewers should consider external validation a minimum prerequisite for publication of novel prediction models, rejecting manuscripts that report only internal validation without justification for why external validation is infeasible. Reviewers should demand discussion of deployment feasibility, including technical requirements for electronic health record integration, clinician workflow mapping, and prospective evaluation plans [1-3]. Journals should consider adopting mandatory reporting checklists adapted from TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) specifically for machine learning models in surgery [4, 5].

Table 2 proposes an evidence-to-implementation framework showing that postoperative machine learning models should be judged not only by internal performance, but by their progression through external validation, calibration-for-use, workflow integration, prospective evaluation, and post-deployment governance.

Table 2. Evidence-to-Implementation Framework for Evaluating Whether Postoperative Machine Learning Models Are Ready for Publication, Validation, and Clinical Deployment

Translational stage	Minimum evidentiary requirement	Key methodological questions	Typical failure mode in current literature	Practical implication for authors, reviewers, and health systems
Stage 1. Model development	Clear cohort definition, predictor availability at intended decision time, transparent outcome definition, event adequacy	Was the model built on clinically meaningful predictors available before intervention decisions must be made?	Models are developed on retrospectively convenient variables without clear temporal anchoring	Development quality should be judged by clinical usability, not only algorithm selection
Stage 2. Internal evaluation	Robust internal testing, prespecified performance metrics, discrimination plus calibration assessment	Are AUROC, calibration, and class imbalance handled appropriately and reported together?	AUROC is emphasized while calibration and threshold behavior are omitted	Publication should not rely on discrimination alone
Stage 3. Bias and transportability assessment	Explicit handling of missing data, leakage prevention, subgroup analysis, reporting of dataset provenance	Could model performance be inflated by leakage, narrow sampling, or unstable event definitions?	Single-center optimism is misinterpreted as generalizability	Reviewers should treat narrow data provenance as a major translational limitation
Stage 4. External validation	Independent institutional or temporal dataset	Does the model maintain acceptable	External validation is absent or limited to	External validation should be the

	testing with reproducible model transfer procedures	performance outside the derivation environment?	highly similar datasets	minimum threshold before strong clinical claims are made
Stage 5. Calibration-for-use	Calibration slope/intercept, decision-threshold justification, subgroup recalibration where necessary	Are predicted risks numerically trustworthy enough to trigger prophylaxis or monitoring decisions?	A model shows acceptable AUROC but generates poorly calibrated risk estimates	Clinically actionable deployment requires trustworthy probabilities, not only rank ordering
Stage 6. Workflow integration design	Defined user, timing, interface, and linked intervention pathway	Who sees the prediction, when is it delivered, and what action should follow?	Predictions are published without any workflow mapping or recommendation logic	Implementation feasibility should be reported as part of model evaluation, not deferred indefinitely
Stage 7. Prospective implementation testing	Silent trial or pilot deployment, user acceptance assessment, monitoring of alert burden and adherence	Does the model function safely in practice without excessive alert fatigue or abandonment?	Technical performance is assumed to translate directly into clinical uptake	Health systems should pilot before scaling and measure both use and usability
Stage 8. Clinical impact evaluation	Controlled prospective study, ideally pragmatic or randomized, with complication outcomes and unintended consequences measured	Does model-guided care actually reduce SSI, VTE, or respiratory failure rates?	Studies stop at prediction accuracy and never test whether outcomes improve	True clinical value begins only when prediction changes care and improves patient outcomes
Stage 9. Post-deployment governance	Drift surveillance, recalibration rules, retirement criteria, accountability assignment	How will performance be monitored as patient mix, workflows, and documentation change?	Deployment is conceptualized as a one-time event rather than an ongoing governance process	Sustainable clinical AI requires continuing oversight rather than static approval
Bottom-line publication standard	At minimum, externally validated, calibrated, and clinically contextualized evidence	Is the manuscript making claims appropriate to the maturity of the evidence?	Internal-validation studies overclaim readiness for practice	Editors and reviewers should align publication claims with translational stage rather than novelty alone

For hospital administrators

Hospital administrators should pilot one machine learning model in a well-defined, low-risk clinical context before broader adoption across multiple complication types or surgical services. VTE prediction represents the most mature application with the strongest evidence of improved discrimination over Caprini scores and the availability of actionable preventive interventions (chemoprophylaxis ordering) [23-31]. Pilot implementations should include prospective tracking of model accuracy, clinician acceptance rates, and complication outcomes before and after deployment, with predefined stopping rules for poor performance [2, 12].

For regulatory bodies (FDA, EMA)

Regulatory bodies including the FDA and EMA should require external validation on multi-institutional data as a condition for marketing authorization of machine learning models intended to guide postoperative complication prevention. Deployment should be accompanied by post-market surveillance requirements including prospective performance monitoring, calibration drift detection, and predefined thresholds for model retraining or retirement [2, 7, 12]. Regulatory frameworks should distinguish between models that provide risk information alone versus those that generate specific treatment recommendations (e.g., "order VTE prophylaxis"), with higher evidentiary standards for the latter category [23-25].

Research gaps

Prospective implementation trials

No completed randomized controlled trial has evaluated whether machine learning-guided preventive strategies reduce postoperative SSI, respiratory failure, or VTE compared to usual risk assessment. The single randomized trial of machine learning for perioperative complications examined the effect of model predictions on clinician risk classification rather than patient outcomes [12]. Researchers should prioritize pragmatic cluster-randomized trials in which surgical units are randomized to receive model-generated alerts and recommendations versus usual care, with 30-day complication rates as the primary outcome [2, 12].

Multi-center external validation

Existing machine learning models for SSI, VTE, and respiratory failure require systematic external validation across diverse hospital settings including community hospitals, safety-net institutions, and international sites. Collaborative research networks should facilitate sharing of trained model code and coefficients across institutions without requiring transfer of protected health information [2, 7, 12]. External validation studies should report performance stratified by patient subgroups (age, comorbidity burden, procedure type) to identify populations in which models may fail and require recalibration [4, 5, 22].

Integration with clinical workflow

Research is needed on user-centered design of model-generated alerts for postoperative complication prevention, including optimal timing (preoperative vs intraoperative vs postoperative), format (passive display vs interruptive alert), and specificity (risk score alone vs risk score plus actionable recommendations). Mixed-methods studies should evaluate clinician cognitive load, alert acceptance rates, and unintended consequences such as alert fatigue or inappropriate prophylaxis [2, 12]. Integration with existing order entry systems to facilitate one-click ordering of VTE chemoprophylaxis or respiratory therapy consultations represents a specific technical research priority [23-25].

Implications

For research practice

The research community must shift incentives from quantity of developed models to quality of validated and implemented models. Journals, funders, and academic promotion committees should reward external validation studies, implementation science research, and reporting of negative results or deployment failures rather than prioritizing novel model development with optimistic performance claims [1-3]. Preprint and registered report formats can reduce publication bias by allowing peer review of study protocols before results are known [4, 5].

For clinical practice

Current machine learning models for postoperative complication prediction are not ready for standalone clinical deployment without local external validation and prospective monitoring. Clinicians should view published AUROC values as optimistic estimates that will likely decline when models are applied to their local patient populations. Decision support

using existing models should be limited to pilot use with close oversight, and model predictions should never replace clinical judgment given the absence of randomized trial evidence demonstrating improved patient outcomes [2, 12, 22].

For policy and regulation

The FDA should classify machine learning models for predicting major postoperative complications (SSI, VTE, respiratory failure) as moderate- or high-risk clinical decision support software requiring premarket notification or approval. Regulatory requirements should include external validation on a multi-institutional dataset, demonstration of calibration across risk deciles, and a post-market surveillance plan for performance drift. Reimbursement policies should condition payment for model-guided preventive interventions on participation in prospective registries or randomized trials [2, 7, 12].

Conclusion

This systematic review of 32 studies evaluating machine learning models for predicting postoperative surgical site infection, venous thromboembolism, and respiratory failure found that models achieve good to excellent discriminative performance on internal validation (AUROC 0.70-0.90) and consistently outperform traditional risk scores such as NSQIP and Caprini. However, fewer than 20% of studies performed external validation, and less than 5% reported any form of clinical deployment. The validation-deployment gap represents the critical barrier between promising retrospective model performance and meaningful improvements in surgical patient outcomes.

The finding that machine learning models rarely undergo external validation before publication and almost never reach clinical deployment indicates that the field remains in early stages of translation. Without external validation, published AUROC values cannot be trusted to generalize to new patient populations, clinical settings, or time periods. Without deployment and prospective evaluation, it remains unknown whether model-generated predictions and recommendations actually reduce complication rates or merely increase clinician workload.

The research community, journal editors, hospital administrators, and regulatory bodies must take coordinated action to close the validation-deployment gap. External validation on multi-institutional data should be established as a minimum standard for publication and regulatory approval. Pragmatic randomized trials of model-guided preventive interventions should be prioritized over continued development of novel models without implementation plans. Pilot deployment of VTE prediction models, which have the strongest evidence base and most actionable interventions, represents a reasonable starting point for health systems.

The vision for the next decade should shift from retrospective model development to prospective implementation science. Integrated surgical decision support systems that predict risks of SSI, VTE, and respiratory failure simultaneously, present actionable recommendations at preoperative and postoperative time points, and undergo continuous recalibration using local data could substantially reduce preventable complications. Achieving this vision requires that researchers, clinicians, regulators, and funders collectively prioritize external validation and clinical deployment over the proliferation of unvalidated models.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 26 Mar 2024 Revised: 04 Jun 2024 Accepted: 31 Jul 2024
Published online: 20 January 2025

References

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 Netw Open. 2021;4(3):e212240. https://doi.org/10.1001/jamanetworkopen.2021.2240.

Ren Y, Loftus TJ, Datta S, Ruppert MM, Guan Z, Miao S, et al. Performance of a machine learning algorithm using electronic health record data to predict postoperative complications and report on a mobile platform. JAMA Netw Open. 2022;5(5):e2211973. https://doi.org/10.1001/jamanetworkopen.2022.11973.

Castela Forte J, Yeshmagambetova G, van der Grinten ML, Scheeren TW, Nijsten MW, Mariani MA, et al. Comparison of machine learning models including preoperative, intraoperative, and postoperative data and mortality after cardiac surgery. JAMA Netw Open. 2022;5(10):e2237970. https://doi.org/10.1001/jamanetworkopen.2022.37970.

Mahajan A, Esper S, Oo TH, McKibben J, Garver M, Artman J, et al. Development and validation of a machine learning model to identify patients before surgery at high risk for postoperative adverse events. JAMA Netw Open. 2023;6(7):e2322285. https://doi.org/10.1001/jamanetworkopen.2023.22285.

Liu Y, Ko CY, Hall BL, Cohen ME. American College of Surgeons NSQIP risk calculator accuracy using a machine learning algorithm compared with regression. J Am Coll Surg. 2023;236(5):1024-30. https://doi.org/10.1097/XCS.0000000000000613.

Rogers MP, Janjua H, DeSantis AJ, Grimsley E, Pietrobon R, Kuo PC. Machine learning refinement of the NSQIP risk calculator: who survives the “hail mary” case? J Am Coll Surg. 2022;234(4):652-9. https://doi.org/10.1097/XCS.0000000000000088.

Stam WT, Ingwersen EW, Ali M, Spijkerman JT, Kazemier G, Bruns ER, et al. Machine learning models in clinical practice for the prediction of postoperative complications after major abdominal surgery. Surg Today. 2023;53(10):1209-15. https://doi.org/10.1007/s00595-023-02724-0.

Ravenel M, Joliat GR, Demartines N, Uldry E, Melloul E, Labgaa I. Machine learning to predict postoperative complications after digestive surgery: a scoping review. Br J Surg. 2023;110(12):1646-9.

Shapey IM, Sultan M. Machine learning for prediction of postoperative complications after hepato-biliary and pancreatic surgery. Artif Intell Surg. 2023;3(1):1-3. https://doi.org/10.20517/ais.2023.08.

Abraham J, Bartek B, Meng A, King CR, Xue B, Lu C, et al. Integrating machine learning predictions for perioperative risk management: towards an empirical design of a flexible-standardized risk assessment tool. J Biomed Inform. 2023;137:104270. https://doi.org/10.1016/j.jbi.2022.104270.

Fritz BA, King CR, Abdelhack M, Chen Y, Kronzer A, Abraham J, et al. Effect of machine learning models on clinician prediction of postoperative complications: the Perioperative ORACLE randomised clinical trial. Br J Anaesth. 2024;133(5):1042-50. https://doi.org/10.1016/j.bja.2024.07.026.

Kiyatkin ME, Aasman B, Fazzari MJ, Rudolph MI, Melo MF, Eikermann M, et al. Development of an automated, general-purpose prediction tool for postoperative respiratory failure using machine learning: a retrospective cohort study. J Clin Anesth. 2023;90:111194. https://doi.org/10.1016/j.jclinane.2023.111194.

Yoon HK, Kim HJ, Kim YJ, Lee H, Kim BR, Oh H, et al. Multicentre validation of a machine learning model for predicting respiratory failure after noncardiac surgery. Br J Anaesth. 2024;132(6):1304-14. https://doi.org/10.1016/j.bja.2024.02.018.

van Boekel AM, van der Meijden SL, Arbous SM, Nelissen RG, Veldkamp KE, Nieswaag EB, et al. Systematic evaluation of machine learning models for postoperative surgical site infection prediction. PLoS One. 2024;19(12):e0312968. https://doi.org/10.1371/journal.pone.0312968.

Chen KA, Joisa CU, Stem JM, Guillem JG, Gomez SM, Kapadia MR. Improved prediction of surgical-site infection after colorectal surgery using machine learning. Dis Colon Rectum. 2023;66(3):458-66. https://doi.org/10.1097/DCR.0000000000002571.

Gutierrez-Naranjo JM, Moreira A, Valero-Moreno E, Bullock TS, Ogden LA, Zelle BA. A machine learning model to predict surgical site infection after surgery of lower extremity fractures. Int Orthop. 2024;48(7):1887-96. https://doi.org/10.1007/s00264-024-06259-7.

Petrosyan Y, Thavorn K, Smith G, Maclure M, Preston R, van Walravan C, et al. Predicting postoperative surgical site infection with administrative data: a random forests algorithm. BMC Med Res Methodol. 2021;21(1):179. https://doi.org/10.1186/s12874-021-01373-z.

Zhuang Y, Dyas A, Meguid RA, Henderson WG, Bronsert M, Madsen H, et al. Preoperative prediction of postoperative infections using machine learning and electronic health record data. Ann Surg. 2024;279(4):720-6. https://doi.org/10.1097/SLA.0000000000006176.

Lu K, Tu Y, Su S, Ding J, Hou X, Dong C, et al. Machine learning application for prediction of surgical site infection after posterior cervical surgery. Int Wound J. 2024;21(4):e14607. https://doi.org/10.1111/iwj.14607.

Zhang Q, Chen G, Zhu Q, Liu Z, Li Y, Li R, et al. Construct validation of machine learning for accurately predicting the risk of postoperative surgical site infection following spine surgery. J Hosp Infect. 2024;146:232-41.
https://doi.org/10.1016/j.jhin.2024.01.021.

Wang H, Fan T, Yang B, Lin Q, Li W, Yang M. Development and internal validation of supervised machine learning algorithms for predicting the risk of surgical site infection following minimally invasive transforaminal lumbar interbody fusion. Front Med. 2021;8:771608.
https://doi.org/10.3389/fmed.2021.771608.

Ghaith AK, Ghanem M, Zamanian C, Bon-Nieves AA, Bhandarkar A, Nathani K, et al. Using machine learning to predict 30-day readmission and reoperation following resection of supratentorial high-grade gliomas: an ACS NSQIP study involving 9418 patients. Neurosurg Focus. 2023;54(6):E12.
https://doi.org/10.3171/2023.3.FOCUS2385.

Karabacak M, Margetis K. Machine learning-based prediction of short-term adverse postoperative outcomes in cervical disc arthroplasty patients. World Neurosurg. 2023;177:e226-38.
https://doi.org/10.1016/j.wneu.2023.06.066.

Lin B, Chen F, Wu M, Li C, Lin L. Machine learning models for prediction of postoperative venous thromboembolism in gynecological malignant tumor patients. J Obstet Gynaecol Res. 2024;50(7):1175-81.
https://doi.org/10.1111/jog.16018.

Santipas B, Chanajit A, Wilartratsami S, Ittichaiwong P, Veerakanjana K, Luksanapruksa P. Development of machine learning algorithms for predicting preoperative and postoperative venous thromboembolism in patients undergoing surgery for spinal metastasis. Siriraj Med J. 2024;76(6):381-8.
https://doi.org/10.33192/smj.v76i6.267015.

Shohat N, Ludwick L, Sherman MB, Fillingham Y, Parvizi J. Using machine learning to predict venous thromboembolism and major bleeding events following total joint arthroplasty. Sci Rep. 2023;13(1):2197.
https://doi.org/10.1038/s41598-023-29421-2.

Nemeth B, Smeets M, Pedersen AB, Kristiansen EB, Nelissen R, Whyte M, et al. Development and validation of a clinical prediction model for 90-day venous thromboembolism risk following total hip and total knee arthroplasty: a multinational study. J Thromb Haemost. 2024;22(1):238-48.
https://doi.org/10.1016/j.jtha.2023.10.018.

Lex JR, Koucheki R, Abbas A, Wolfstadt JI, McLawhorn AS, Ravi B. Predicting 30-day venous thromboembolism following total joint arthroplasty: adjusting for trends in annual length of stay. Arthroplast Today. 2024;30:101491.
https://doi.org/10.1016/j.artd.2024.101491.

Hanh BM, Cuong LQ, Son NT, Duc DT, Hung TT, Hung DD, et al. Determination of risk factors for venous thromboembolism by an adapted Caprini scoring system in surgical patients. J Pers Med. 2019;9(3):36.
https://doi.org/10.3390/jpm9030036.

Krauss ES, Segal A, Cronin M, Dengler N, Lesser ML, Ahn S, et al. Implementation and validation of the 2013 Caprini score for risk stratification of arthroplasty patients in the prevention of venous thrombosis. Clin Appl Thromb Hemost. 2019;25:1076029619838066.
https://doi.org/10.1177/1076029619838066.

Koretsky MJ, Brovman EY, Urman RD, Tsai MH, Cheney N. A machine learning approach to predicting early and late postoperative reintubation. J Clin Monit Comput. 2023;37(2):501-8.
https://doi.org/10.1007/s10877-022-00870-5.

Hadaya J, Verma A, Sanaiha Y, Ramezani R, Qadir N, Benharash P. Machine learning-based modeling of acute respiratory failure following emergency general surgery operations. PLoS One. 2022;17(4):e0267733.

https://doi.org/10.1371/journal.pone.0267733.