ANESTHESIOLOGY

Openventio PUBLISHERS

Review Article

Predictive Analytics for Intensive Care Admission Following Surgery

Nikos Saez-Bermudez[1*], Jakub Eberhardt[2] and Marco Larsen[3]

[1]Department of Critical Care Medicine, Shannon University Hospital, Ireland
[2]Department of Perioperative Medicine, Danube University of Applied Medicine, Hungary
[3]Department of Critical Care Medicine, Danube University of Applied Medicine, Hungary

ABSTRACT

Deciding which surgical patients require postoperative intensive care remains a judgment call made with incomplete information, and legacy scoring systems such as APACHE II, POSSUM, and the Surgical Apgar Score were derived from modest cohorts using a small set of fixed-weight variables.

This review examines the growing body of machine learning and deep learning approaches developed to predict postoperative ICU admission, with particular attention to the harder and more clinically consequential problem of unplanned rather than planned transfer.

Tree-based ensembles trained on combined preoperative and intraoperative variables consistently outperform classical logistic regression and legacy risk indices, with disease-specific models in oncologic, spinal, and cardiac surgery populations showing further gains from tailored feature sets, while deep learning and multimodal architectures are beginning to integrate time-series and narrative EHR data for real-time deterioration prediction.

Despite this progress, nearly every model reviewed remains single-center, retrospective, and externally unvalidated, and calibration, class imbalance, and clinician trust in model interpretability remain unresolved barriers to routine deployment.

We conclude that prospective, multicenter validation paired with implementation studies measuring the effect of model-guided triage on actual bed utilization and outcomes is the necessary next step before these tools can move from promising retrospective analyses to routine perioperative practice.

Keywords: *intensive care unit admission; machine learning; postoperative risk prediction; perioperative medicine; clinical decision support*

Citation: Saez-Bermudez N, Eberhardt J and Larsen M (2025) Predictive Analytics for Intensive Care Admission Following Surgery. Anesthesiology. 5(1):01-05.

DOI: 10.17140/2025.vol5.i01.pp01-05

Corresponding Author: *Nikos Saez-Bermudez*

E-mail Address: *n.bermudez@suh.ie*

Article History: *Received 01 March 2025 | Accepted 21 March 2025 | Published 13 April 2025*

Copyright: *© 2025 Nikos Saez-Bermudez, 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

Deciding who needs a postoperative intensive care bed is one of the more consequential judgment calls in perioperative medicine, and it is made, more often

than not, on incomplete information. Surgeons and anesthesiologists weigh comorbidity burden, the magnitude of the operation, intraoperative course, and a gut sense of how a patient tolerated the day, then book a bed or send the patient to the ward. Get it wrong in one direction and a fragile patient decompensates on a general floor without the nursing ratios or monitoring to catch it early; get it wrong in the other and a scarce, expensive ICU bed is occupied by someone who never needed it, at the expense of someone who does. Traditional risk scores such as APACHE II, POSSUM, and the Surgical Apgar Score were built for exactly this kind of stratification, but they were derived decades ago from modest cohorts, rely on a handful of hand-picked variables, and were validated mostly for mortality rather than for the specific, more immediate question of ICU bed need [1]. Over the past several years, the volume of granular perioperative data captured in electronic health records — vital sign trends, laboratory trajectories, free-text operative notes, intraoperative monitor waveforms — has made it possible to ask whether machine learning can do this triage better, earlier, and with less reliance on a clinician's intuition alone.

This review surveys where that effort currently stands: the modeling approaches being used to predict postoperative ICU admission, what the evidence says about their discriminative performance relative to legacy risk scores, the practical and ethical obstacles to putting them into routine use, and where the field still has real work to do before these tools change how beds get allocated.

FROM RISK SCORES TO LEARNED MODELS

The impetus for machine learning in this space is not that clinicians do a poor job of risk assessment in aggregate, but that fixed-weight scoring systems are blunt instruments. POSSUM and its variants sum a small set of physiological and operative variables into a linear score, which is easy to compute at the bedside but assumes the relationship between each variable and outcome is the same shape for every patient and every operation — an assumption that rarely survives contact with a heterogeneous surgical population. A comparison of APACHE II, POSSUM, and the Surgical Apgar Score in patients undergoing emergency laparotomy illustrates the point well: each score performs adequately in aggregate but the three disagree substantially on individual patients, and none was designed specifically to flag a patient headed for the ICU rather than one headed for death or a major complication more broadly [1]. Machine learning models — gradient-boosted trees, random forests, and neural networks chief among them — relax the linearity assumption and can absorb far more input variables than a bedside score ever could, which is why they have become the default architecture for groups trying to build ICU-admission predictors from scratch.

Early efforts in noncardiac elective surgery demonstrated the basic feasibility of this approach. A prediction model built from preoperative and intraoperative variables to flag patients likely to need more than 24 hours of ICU care after elective noncardiac surgery found that ensemble tree-based algorithms outperformed logistic regression, with the gap widening as the number of candidate predictors grew [2,3]. That pattern — modest but consistent gains for tree ensembles over classical regression, and larger gains once intraoperative data is added to preoperative data — recurs across most of the subsequent literature and is probably the single most reproducible finding in the field so far.

MODEL PERFORMANCE AND THE UNPLANNED-ADMISSION PROBLEM

Planned ICU admissions, arranged before the patient ever goes to the operating room, are the easier prediction problem: the decision has effectively already been made by the surgical team, and a model's job is mostly to formalize existing practice. The harder and more clinically useful target is the unplanned admission — the patient who was slated for the ward, deteriorated in the recovery room or on the floor, and needed emergent ICU transfer. A recently developed model aimed squarely at this problem, built from a major non-cardiac surgery cohort, incorporated intraoperative hemodynamic and ventilatory data alongside preoperative comorbidity variables and achieved discrimination meaningfully better than existing preoperative risk indices for predicting unplanned rather than planned ICU transfer [4,5,6]. The distinction matters practically: a hospital that can identify, before the patient leaves the operating room, which "routine" ward-bound patients carry an elevated risk of unplanned deterioration can intervene — closer floor monitoring, an intermediate-care bed, or a proactive rapid-response consult — rather than reacting after the fact.

Disease-specific models have followed a similar trajectory. In oncology populations, where baseline hematologic derangement and treatment-related organ dysfunction complicate risk assessment, a model incorporating routine hematologic parameters alongside clinical characteristics improved prediction of postoperative ICU need over generic surgical risk indices, suggesting that population-specific feature sets add real value beyond a one-size-fits-all model [7]. Spinal metastasis surgery, a population with unusually high perioperative complication rates and limited applicability of standard scoring systems, has likewise seen an interpretable model developed specifically for preoperative ICU risk stratification, with the authors emphasizing feature transparency — being able to say why the model flagged a given patient — as central to clinician acceptance [8]. Cardiac surgery, which has always run its own dedicated risk-scoring tradition (EuroSCORE,

STS-PROM), has generated a parallel body of machine learning work aimed at predicting in-hospital mortality and adverse events in post-cardiotomy ICU patients, generally reporting modest discrimination gains over legacy scores once large multi-center critical-care databases are used for training, alongside efforts to address the class imbalance inherent to relatively rare adverse outcomes through targeted feature engineering and hyperparameter optimization [9,10,11,12].

An earlier, frequently cited demonstration that machine learning could outperform standard preoperative risk indices for the combined endpoint of postsurgical mortality and ICU admission used a large multicenter cohort and found that gradient boosting substantially improved calibration and discrimination over the American College of Surgeons' NSQIP calculator, particularly for the sickest decile of patients — precisely the group where the cost of misclassification, in either direction, is highest [13].

DEEP LEARNING, MULTIMODAL DATA, AND REAL-TIME DETERIORATION PREDICTION

Where tree-based ensembles dominate models built on structured preoperative variables, deep learning has found its niche in problems that require integrating heterogeneous, high-dimensional, or sequential data — exactly the kind of data an ICU or perioperative unit generates continuously. An interpretable deep learning framework developed for predicting postoperative infections in ICU patients combined structured EHR variables with attention-based architectures to both predict risk and highlight the clinical features driving each prediction, addressing one of the persistent complaints about neural-network-based clinical tools: that they function as black boxes clinicians are asked to trust without being shown their reasoning [14,15,16]. A related and more ambitious line of work is multimodal deterioration prediction, which fuses time-series vital-sign and laboratory trajectories with unstructured clinical notes processed through natural language models, aiming to catch the kind of early deterioration signal that a nurse's narrative note sometimes captures well before it shows up in a formal vital-sign trend [17]. Deep learning models trained on longitudinal inpatient EHR sequences for broader clinical outcome prediction — not limited to ICU admission specifically — have shown that recurrent and transformer-style architectures can extract useful temporal signal from the messy, irregularly sampled data that characterizes real hospital records, a capability that maps directly onto the problem of watching a postsurgical patient's trajectory evolve hour by hour [18].

Emergency department triage systems offer a useful analogy and a cautionary one. A deep-learning-based ED triage system built on electronic medical records demonstrated that model-assisted triage could outperform standard acuity scales in retrospective validation, but the same study underscored how much model performance depends on the idiosyncrasies of the training institution's documentation habits and patient mix — a caution that applies with equal force to postoperative ICU-admission models, most of which remain single-center, retrospective, and unvalidated outside the health system that built them [19].

BARRIERS TO CLINICAL ADOPTION

The gap between a well-performing retrospective model and a deployed clinical tool remains wide, and a recent narrative review of ICU admission risk prediction and artificial intelligence lays out the recurring obstacles clearly: most models are trained and validated on a single institution's data, few have been prospectively tested against actual clinician decision-making in real time, and fewer still have demonstrated that acting on model output changes patient outcomes rather than simply reproducing an accurate but practically inert prediction [20]. Calibration — whether a model's predicted probability of 20% actually corresponds to roughly a one-in-five event rate in the population it is applied to — is repeatedly reported as worse than discrimination, and calibration drifts further when a model trained at one hospital is applied at another with a different case mix, staffing model, or threshold for calling something an ICU-level event in the first place. Class imbalance compounds the problem: unplanned ICU admission and major postoperative deterioration are, thankfully, uncommon events relative to the total surgical volume, and models trained without deliberate attention to this imbalance tend to either under-flag genuine at-risk patients or generate enough false positives that clinicians learn to ignore the alert altogether, an alarm-fatigue failure mode well documented in other clinical decision-support contexts.

There is also a harder, less technical problem sitting underneath all of this: an ICU bed is a scarce resource governed by staffing, physical capacity, and institutional politics as much as by clinical need, so even a perfectly calibrated model cannot by itself solve a bed-allocation problem that is partly logistical. A model that correctly identifies twice as many at-risk patients as current practice is only useful if the hospital has somewhere to put them, and several of the papers reviewed here are candid that their tools are meant to support triage conversations rather than replace the judgment of the surgical and critical care teams making the final call [4,8].

CONCLUSION

The evidence to date supports a fairly specific, bounded claim: machine learning models, particularly tree ensembles trained on combined preoperative and intraoperative data, can outperform legacy risk scores like APACHE II and POSSUM at discriminating which surgical patients will need postoperative ICU-level

care, and this advantage is most pronounced for the unplanned admissions that current practice handles worst. What the evidence does not yet support is routine clinical deployment. Almost every model reviewed here was developed and validated retrospectively at a single center, few have undergone external validation, fewer still have been tested prospectively against real-time clinician decisions, and none has demonstrated in a randomized or quasi-experimental design that acting on its output improves patient outcomes or resource utilization rather than simply producing an accurate number. Interpretability work — attention-based architectures, feature-importance reporting — has made real progress toward addressing clinician distrust of black-box predictions, and disease-specific models in oncology, spinal surgery, and cardiac surgery suggest that tailoring feature sets to a population outperforms generic one-size-fits-all scores. The next phase of work that would actually move this field toward the bedside is prospective, multicenter validation paired with implementation studies that measure whether earlier identification of at-risk patients changes triage decisions, bed utilization, and outcomes — not just whether the receiver operating curve looks good in a retrospective cohort.

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 NSB conceived the review and drafted the manuscript. JE performed the literature search and synthesized the machine learning methodology sections. ML critically revised the manuscript for clinical accuracy and approved the final version. All authors read and approved the final manuscript.

Ethics Statement This is a literature review that did not involve new human or animal data collection; no ethics approval was required.

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

REFERENCES

1. Comparison of Acute Physiology and Chronic Health Evaluation (APACHE) II, Physiological and Operative Severity Score for the Enumeration of Mortality and Morbidity (POSSUM), and Surgical Apgar Score in Patients Undergoing Emergency Laparotomy (2025) PMC13283001.

2. Lan L, Chen F, Luo J, Li M, Hao X, Hu Y, Yin J, Zhu T, Zhou X (2022) Prediction of intensive care unit admission (>24h) after surgery in elective noncardiac surgical patients using machine learning algorithms. Digital Health, DOI:10.1177/20552076221110543.

3. Prediction of intensive care unit admission (>24h) after surgery in elective noncardiac surgical patients using machine learning algorithms (2022) PMC9326842.

4. Development of a machine learning-derived model to predict unplanned ICU admissions after major non-cardiac surgery (2025) BMC Anesthesiology, DOI:10.1186/s12871-025-03195-8.

5. Development of a machine learning-derived model to predict unplanned ICU admissions after major non-cardiac surgery (2025) PMID:40676538.

6. Development of a machine learning-derived model to predict unplanned ICU admissions after major non-cardiac surgery (2025) PMC12273467.

7. A Machine Learning Model to Predict Post-Operative Intensive Care Unit Admission in Patients with Cancer Based on Clinical Characteristics and Hematologic Parameters Data (2025) PMID:42074702.

8. Enhancing preoperative risk stratification: an interpretable machine learning model for ICU admission in spinal metastasis surgery (2025) PMID:41376391.

9. Machine learning-based prediction of in-hospital mortality for post cardiovascular surgery patients admitting to intensive care unit: a retrospective observational cohort study based on a large multi-center critical care database (2022) PMID:36126435.

10. Interpretable machine learning to predict postoperative adverse outcomes in cardiac surgery (2026) PMC13419318.

11. Enhancing machine learning performance in cardiac surgery ICU: Hyperparameter optimization with metaheuristic algorithm (2025) PMC11809875.

12. Optimizing In-Hospital Mortality Prediction After Cardiac Surgery: A Machine Learning Approach Using Feature Engineering for Imbalanced Data (2026) PMC13512290.

13. Corey KM, et al. (2018) Utilizing Machine Learning Methods for Preoperative Prediction of Postsurgical Mortality and Intensive Care Unit Admission. Annals of Surgery, DOI:10.1097/SLA.0000000000003297.

14. An interpretable deep learning framework for predictive modeling of postoperative infections in ICU patients (2026) PLOS ONE, DOI:10.1371/journal.pone.0346896.

15. An interpretable deep learning framework for

predictive modeling of postoperative infections in ICU patients (2026) PMC13065004.

16. An interpretable deep learning framework for predictive modeling of postoperative infections in ICU patients (2026) PMID:41955272.
17. Multimodal Deep Learning for Early Prediction of Patient Deterioration in the ICU: Integrating Time-Series EHR Data with Clinical Notes (2026) arXiv:2603.14719.
18. Deep learning model for clinical outcome prediction using longitudinal inpatient electronic health records (2025) JAMIA Open.
19. A Novel Deep Learning-Based System for Triage in the Emergency Department Using Electronic Medical Records: Retrospective Cohort Study (2022) PMC8749584.
20. Risk Prediction of ICU Admission and Artificial Intelligence: A Narrative Review (2024) HealthManagement.org.