RESEARCH PAPER

AI-Driven Strategies for Optimizing Postoperative Care Efficiency and Reducing Complications in Patients with COVID-19

Original Research Article

Trivedi Darshna Mukundray[1*], **Mohit Bhadla**[2]

[1]*Research Scholar, Department of Computer Science and Engineering, Gandhinagar Institute of Technology, Gandhinagar University, Gujarat, India – 382721*

[2]*Head, Department of Computer Science and Engineering, Gandhinagar University, Gujarat, India*

***Corresponding author:** trivedidarshna@gmail.com

ABSTRACT

Surgery in patients with perioperative SARS-CoV-2 infection carries high rates of pulmonary complications and death, while pandemic pressure on high-dependency unit (HDU) and intensive care beds has made efficient postoperative care critical. We evaluated an integrated artificial intelligence (AI) framework for this population. A synthetic cohort of 10,000 surgical patients was generated from published COVIDSurg risk data. Three components were evaluated: (A) preoperative prediction of major 30-day complications using logistic regression, random forest and XGBoost, benchmarked against a four-variable clinical score, with calibration, SHAP explainability and fairness analysis; (B) a machine-learning (ML) early-warning model trained on simulated 4-hourly vital signs and compared with NEWS2; and (C) Monte Carlo simulation of an AI-guided care pathway versus standard care. XGBoost achieved AUROC 0.840 (95% CI 0.821–0.858) and calibration slope 1.14, versus 0.739 for the clinical score. ML early warning detected 97.2% of deteriorating patients at 3.0 false alerts per 100 patient-days, versus 78.5% for NEWS2 $\geq$ 7 at 3.4 (median lead time 10.6 h). Per 1,000 patients, the AI-guided pathway reduced simulated major complications by 18.0%, deaths by 30.3%, total bed-days by 9.0% and HDU bed-days by 30.3%, robustly across sensitivity scenarios. Sensitivity at a single global threshold was lower in patients under 70 years (71%) than in older patients (92%). Linking AI risk prediction, personalised early warning and risk-stratified pathway decisions could improve both the safety and efficiency of postoperative care; as these findings rest on simulated data, external validation and prospective evaluation are required before clinical use.

Keywords: Artificial intelligence, Machine learning, COVID-19, Postoperative complications, Early warning score, Clinical decision support, Perioperative care, Hospital resource optimisation

How to cite this article: Mukundray TD, Bhadla M., AI-Driven Strategies for Optimizing Postoperative Care Efficiency and Reducing Complications in Patients with COVID-19. Int J Drug Deliv Technol. 2026;16(90s): 1152-1165; DOI: 10.25258/ijddt.16.90s.114

1. INTRODUCTION

Surgery in a patient with SARS-CoV-2 infection carries a markedly higher risk of postoperative complications and death. In the first international cohort of 1,128 surgical patients with perioperative SARS-CoV-2 infection, the COVIDSurg Collaborative reported a 30-day mortality of 23.8% and postoperative pulmonary complications in 51.2% of patients; those with pulmonary complications accounted for 82.6% of all deaths [1]. A subsequent study of 140,231 patients in 116 countries found that surgery within six weeks of a SARS-CoV-2 diagnosis was associated with roughly a fourfold increase in the adjusted odds of 30-day mortality (OR 3.6–4.1), while the risk returned close to baseline after seven weeks [2]. Although vaccination, new variants and better perioperative care have since lowered absolute risk, patients with recent or active infection remain a high-risk surgical group. They also continue to present for emergency surgery, cancer surgery and procedures that cannot be postponed.

The pandemic also exposed how thin postoperative capacity is. Elective surgery was cancelled or deferred on a vast scale, creating backlogs that health systems are still working through [3]. Beds in high-dependency and intensive care units (HDU/ICU) were diverted to COVID-19 care, nursing staff were redeployed and infection-control measures slowed patient flow. In this setting, postoperative efficiency, meaning the right level of monitoring for the right patient for the right length of time, becomes as important as clinical outcome. Over-triage of low-risk patients to HDU consumes scarce beds, while under-triage of high-risk patients delays the recognition of deterioration.

Artificial intelligence (AI) and machine learning (ML) offer tools for both problems. Preoperative risk

models can identify patients most likely to develop complications and so guide decisions on timing of surgery, level of postoperative care and targeted prevention [4, 5]. Continuous or frequent postoperative monitoring combined with ML-based early-warning scores can detect deterioration earlier than threshold-based track-and-trigger systems [6, 7]. Simulation and optimisation methods can then translate individual risk predictions into ward-level decisions about bed allocation and discharge [8].

Research gap. Most published COVID-19 AI work targeted diagnosis, imaging or the prognosis of medical (non-surgical) inpatients [9, 10]. Surgical studies have mostly developed single preoperative risk scores [4, 5] and evaluated them by discrimination alone. Few studies link three components: preoperative risk stratification, postoperative early warning and the operational effect of acting on predictions (complications avoided, length of stay and bed-days). Systematic reviews have also found that many COVID-19 prediction models carry a high risk of bias, with poor calibration reporting and limited attention to fairness and explainability [9].

Objectives. This study aims to:

1. Develop and compare ML models (logistic regression, random forest and gradient boosting) for predicting major 30-day postoperative complications in patients with perioperative SARS-CoV-2 infection, reporting discrimination, calibration and explainability.
2. Evaluate an ML-based postoperative early-warning model against the National Early Warning Score 2 (NEWS2) for detecting deterioration, measuring sensitivity, alert burden and lead time.
3. Estimate, through simulation of care pathways, the effect of an AI-guided, risk-stratified postoperative pathway on complications, length of stay and HDU bed use.

Contributions. The paper provides (i) an integrated AI framework that spans the perioperative pathway from pre-surgical triage to discharge; (ii) a reproducible evaluation on a synthetic cohort whose risk structure follows published international COVID-19 surgical data; (iii) an operational analysis that converts model outputs into outcomes managers can use (beds, days and complications); and (iv) a structured discussion of deployment, governance and bias. Because the cohort is synthetic, the results show the framework's potential and set out the evaluation design. They are not clinical effect estimates and must be confirmed on real patient data before clinical use. Section 2 reviews related work, Section 3 describes the methodology, Section 4 reports results, Section 5 discusses them and Section 6 concludes.

2. LITERATURE REVIEW

2.1 Surgical outcomes with perioperative SARS-CoV-2 infection

The COVIDSurg Collaborative produced the main evidence base on surgery during the pandemic. Its first cohort study [1] identified male sex (adjusted OR 1.75), age ≥ 70 years (OR 2.30), ASA grade 3–5 (OR 2.35), malignant diagnosis (OR 1.55), emergency surgery (OR 1.67) and major surgery (OR 1.52) as independent predictors of 30-day mortality. Its timing study [2] showed that the elevated risk persists for six weeks after diagnosis and is highest in patients with ongoing symptoms: 30-day mortality was 6.0% in patients still symptomatic at seven weeks or more, compared with 2.4% when symptoms had resolved and 1.3% in asymptomatic patients. A further analysis argued that prioritising preoperative vaccination of surgical patients would prevent more deaths than vaccinating the general population [11]. These findings established three levers for perioperative AI: who is at risk, when to operate and how closely to monitor afterwards.

2.2 Machine-learning risk prediction

General surgical risk. Before the pandemic, ML models had already matched or exceeded traditional calculators. MySurgeryRisk used electronic health record (EHR) data to predict eight postoperative complications with AUROC 0.82–0.94 [12]. The POTTER calculator, built with optimal classification trees, captured non-linear interactions in emergency surgery risk better than logistic regression [13].

COVID-19-specific models. The COVIDSurg mortality score [4] was developed on 8,492 patients. The authors compared logistic regression, decision trees and random forests and selected a four-variable logistic model (age, ASA grade, Revised Cardiac Risk Index and preoperative respiratory support), which achieved AUROC 0.73 in derivation and 0.80 in validation. More recently, McDermott et al. [5] trained nine algorithms on 15,882 COVID-19-positive patients from the ACS-NSQIP registry and reported AUROC of about 0.87–0.90 for a LightGBM model predicting a composite of 30-day morbidity and mortality, with older age, ASA class IV–V, low albumin and preoperative ventilatory support among the main SHAP-ranked predictors. For hospitalised (largely non-surgical) COVID-19 patients, the 4C Mortality Score [14] and multi-centre ML mortality models [15] showed that routinely collected physiology and laboratory values, such as C-reactive protein (CRP), lymphocyte count and oxygen saturation, add predictive value beyond demographics.

Quality of COVID-19 models. A living systematic review by Wynants et al. [9] appraised hundreds of COVID-19 prediction models and rated almost all at high risk of bias. Common problems were small

samples, poor handling of missing data, overfitting and absent calibration. Roberts et al. [10] reached a similar conclusion for imaging-based ML. Calibration, often called the Achilles heel of predictive analytics [16], is particularly important when predictions drive resource allocation, because a model that ranks patients well but over-estimates risk will over-allocate HDU beds.

2.3 Early warning of postoperative deterioration

Track-and-trigger systems such as NEWS2 [17] aggregate six vital signs and level of consciousness into a score with fixed thresholds. They are simple and widely adopted, but they ignore the patient's own baseline and the direction of change. This causes particular problems in COVID-19, where chronic hypoxaemia and tachypnoea can produce persistently high scores and alarm fatigue. Churpek et al. [6] showed that ML methods, particularly gradient-boosted trees and random forests, outperformed regression-based scores for ward deterioration. Escobar et al. [7] deployed an automated early-warning programme across 19 hospitals and found lower 30-day mortality in the intervention population. In COVID-19 specifically, Singh et al. [18] externally validated the proprietary Epic Deterioration Index in 392 inpatients and found moderate discrimination (AUROC 0.79). They cautioned that widely deployed proprietary models need independent local validation, a lesson reinforced by the poor external performance of a proprietary sepsis model [19].

Continuous wearable monitoring offers denser data for such models. A systematic review found that continuous ward monitoring may improve the detection of deterioration, but evidence on hard outcomes remains limited [20]. Michard and Kalkman [21] argued that intermittent nurse-measured vital signs leave long unobserved gaps on surgical wards and that wireless monitoring combined with smart alerting can close them. After discharge, virtual care with remote automated monitoring increased the detection and correction of drug errors and reduced pain, although it did not reduce the composite of hospital-based acute care [22].

2.4 Care-pathway and resource optimisation

Enhanced Recovery After Surgery (ERAS) pathways reduce complications and length of stay by standardising evidence-based perioperative care [23]. Postoperative pulmonary complications, the dominant complication in COVID-19 surgical patients, independently increase early mortality and length of stay in non-cardiothoracic surgery [24]. Targeted bundles such as I COUGH (incentive spirometry, coughing and deep breathing, oral care, education, head-of-bed elevation and ambulation) reduced postoperative pneumonia and unplanned intubation [25]. Variation in surgical mortality between hospitals is driven more by failure to rescue patients after complications than by complication rates themselves [26], which links early detection directly to survival. Discrete-event simulation is an established method for testing pathway changes and bed policies before implementation [8].

2.5 Explainability, fairness and governance

Clinicians need to know why a model flags a patient. Shapley additive explanations (SHAP) provide consistent local and global feature attributions for tree ensembles [27]. Algorithmic bias is a documented risk: Obermeyer et al. [28] showed that a commercial algorithm using healthcare cost as a proxy for need systematically under-served Black patients. Reporting guidance (TRIPOD+AI) [29] and WHO guidance on the ethics of AI for health [30] call for subgroup performance reporting, transparency and human oversight. Synthetic data can support method development and sharing while protecting privacy, but results must be validated on real patients [31].

2.6 Summary of gaps

Table 1 positions the present study against representative prior work.

Table 1: Representative AI studies relevant to postoperative care in COVID-19.

Study	Population	Task	Method	Reported performance	Operational impact assessed
COVIDSurg mortality score [4]	8,492 surgical patients, SARS-CoV-2 positive	30-day mortality	Logistic regression (vs RF, trees)	AUROC 0.73 (derivation), 0.80 (validation)	No
McDermott et al. [5]	15,882 NSQIP surgical patients, COVID-19	30-day morbidity or mortality	LightGBM (vs 8 others) + SHAP	AUROC about 0.87–0.90	No
Singh et al. [18]	392 COVID-19 ward inpatients	ICU, ventilation or death	Epic Deterioration Index	AUROC 0.79	No
Knight et al. [14]	Hospitalised COVID-19 patients (ISARIC)	In-hospital mortality	Point score from regression	Validated in large UK cohort	No
Escobar et al. [7]	19 hospitals, general ward	Deterioration	Automated early-warning programme	Lower 30-day mortality	Yes
This study	**Synthetic COVID-19 surgical cohort, 10,000**	**Complications, deterioration, pathway**	**LR, RF, XGBoost + SHAP; ML early warning; pathway simulation**	**Section 4**	**Yes (simulated)**

The gaps addressed here are the lack of (i) an integrated evaluation from preoperative triage to postoperative surveillance, (ii) operational outcome measures alongside discrimination and (iii) routine calibration, decision-curve and subgroup reporting in COVID-19 surgical AI.

3. MATERIALS AND METHODS

The study uses a simulation-based design with three linked components, shown in Fig. 1: (A) a preoperative risk-stratification model, (B) a postoperative ML early-warning model and (C) an AI-guided care pathway evaluated against standard care by Monte Carlo simulation. All analyses were scripted in Python 3 (scikit-learn 1.9, XGBoost 3.4, SHAP 0.54) with fixed random seeds, and reporting follows TRIPOD+AI [29] where applicable.

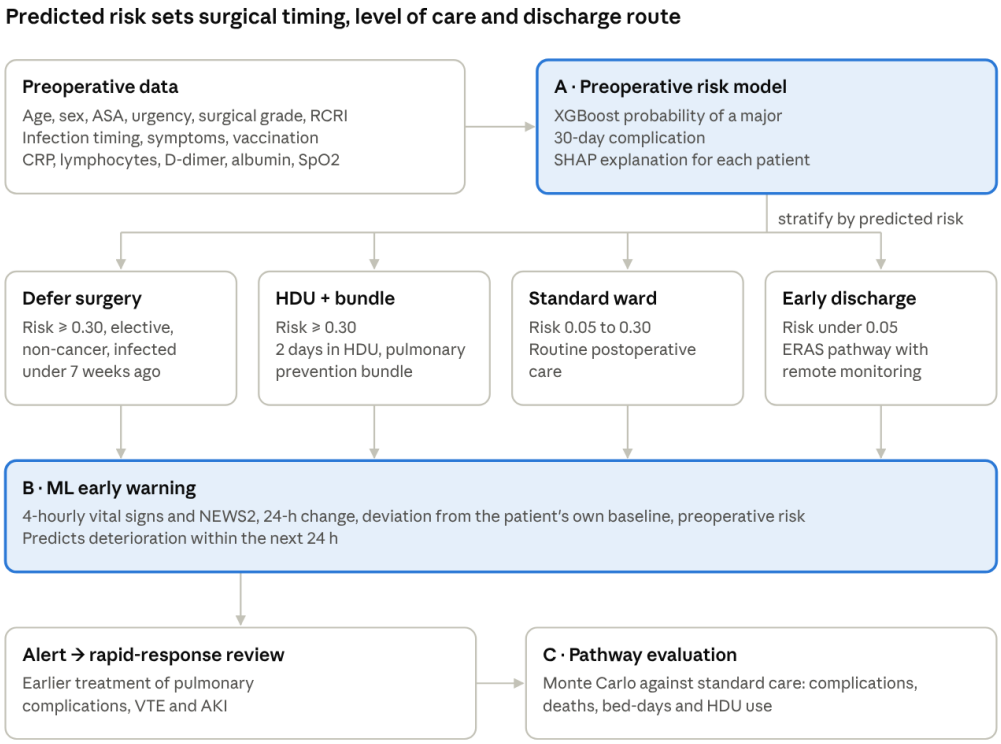


Figure 1: The proposed three-component framework. Preoperative risk (A) routes each patient to one of four pathways; ML early warning (B) monitors all patients after surgery; and pathway simulation (C) measures the combined effect against standard care.

3.1 Synthetic cohort

No identifiable patient data were used. A cohort of 10,000 adult surgical patients with perioperative SARS-CoV-2 infection was generated so that its risk structure follows published international data. Ethical approval was therefore not required, and the design allows the full pipeline to be shared and reproduced [31].

Variables. Each patient has 19 preoperative variables: age, sex, ASA grade, urgency (elective or emergency), surgical grade (minor or major), malignancy, Revised Cardiac Risk Index (RCRI), diabetes, chronic kidney disease, body mass index, vaccination status (≥ 2 doses), weeks from SARS-CoV-2 diagnosis to surgery (0–2, 3–4, 5–6, ≥ 7), symptoms at surgery, preoperative respiratory support, and five laboratory and physiological markers (CRP, lymphocyte count, D-dimer, serum albumin and room-air SpO_2). Marginal distributions and correlations were set to be clinically plausible; for example, symptoms, raised CRP, lymphopenia and lower SpO_2 are concentrated in patients infected within the previous four weeks.

Outcome. The primary outcome is a major 30-day postoperative complication: postoperative pulmonary complication, venous thromboembolism, acute kidney injury, Clavien–Dindo grade ≥ III complication or death. Outcomes were drawn from a logistic data-generating model. Its main-effect coefficients are the natural logarithms of published adjusted odds ratios: male sex 1.75, age ≥ 70 years 2.30, ASA 3–5 2.35, malignancy 1.55, emergency surgery 1.67 and major surgery 1.52 [1]; timing ORs of 4.1, 3.9, 3.6 and 1.5 for 0–2, 3–4, 5–6 and ≥ 7 weeks [2]. Vaccination (OR 0.6) and respiratory support (OR 2.5) were assumed. To reflect real clinical complexity, the model also contains non-linear and interaction terms: a sigmoid CRP effect above 60 mg/L, lymphopenia interacting with recent infection, hypoalbuminaemia (< 30 g/L), an SpO_2 deficit below 95% and an emergency × major × ASA ≥ 3 interaction. The intercept was calibrated to an overall event rate of 18%, between the 11.8–17.2% composite morbidity reported for COVID-19-positive NSQIP patients [5] and the much higher early-pandemic rates [1]. The realised rate was 17.97%.

3.2 Component A: preoperative risk stratification

The cohort was split 70:30 into training (n = 7,000) and held-out test (n = 3,000) sets, stratified by outcome. Four models were compared:

1. **Clinical score**: logistic regression on the four COVIDSurg mortality-score variables (age, ASA, RCRI, respiratory support) [4], as a benchmark.

2. **Logistic regression (LR)** on all 19 variables (standardised; CRP and D-dimer log-transformed).
3. **Random forest (RF)** [32], 400 trees; maximum depth and minimum leaf size tuned by 5-fold cross-validation.
4. **Extreme gradient boosting (XGBoost)** [33], 400 trees, learning rate 0.03, subsampling 0.8; maximum depth and minimum child weight tuned by 5-fold cross-validation.

Cross-validation selected an unrestricted-depth RF with minimum leaf size 5 (CV AUROC 0.821) and an XGBoost model of depth 2 (CV AUROC 0.831). Performance on the test set was assessed by:

- **Discrimination**: AUROC with 1,000-sample bootstrap 95% CI, and area under the precision–recall curve (AUPRC); paired bootstrap (2,000 samples) for differences in AUROC.
- **Calibration**: Brier score, calibration slope and calibration-in-the-large (observed minus expected event rate) [16, 34].
- **Clinical utility**: specificity, PPV and NPV at the threshold giving 80% sensitivity; decision-curve net benefit [35]:

$$NB(p_t) = \frac{TP}{N} - \frac{FP}{N} \cdot \frac{p_t}{1 - p_t} \quad (1)$$

- **Explainability**: mean absolute SHAP values from TreeSHAP [27].
- **Fairness**: AUROC, sensitivity and observed-to-expected (O/E) ratio by sex, age group, urgency and vaccination status.

3.3 Component B: postoperative early warning

For every test-set patient, vital signs (respiratory rate, SpO_2, supplemental oxygen, systolic blood pressure, heart rate, temperature and consciousness) were simulated every 4 h for 7 days, giving 19,160 patient-days and 114,961 observations. Each patient has an individual baseline, measurement noise, a transient early postoperative tachycardia and fever, and random benign spikes (4% of observations). Recently symptomatic COVID-19 patients start with higher respiratory and heart rates and lower SpO_2. In patients who develop a complication, onset time is drawn uniformly between 24 h and 150 h after surgery, and physiological deterioration begins 12–36 h before onset with patient-specific severity.

Two alerting approaches were compared:

- **NEWS2** [17], computed with Scale 1 thresholds; alerts at total scores ≥ 3 to ≥ 9, and the standard ≥ 5 or single red-parameter trigger.
- **ML early-warning model**: XGBoost (depth 4, 400 trees) trained on training-set trajectories to predict complication onset within the next 24 h. Features are the current vital signs and NEWS2, change from 24 h earlier, a 12-h rolling mean, deviation from the patient's own postoperative baseline, and the preoperative risk from Component A.

Patient-level sensitivity was defined as an alert in the 48 h before onset; lead time as the interval from the first such alert to onset; and false alerts as alerts in patients without complication, or more than 48 h before onset, per 100 patient-days.

3.4 Component C: AI-guided care pathway simulation

Two strategies were simulated for the 3,000 test patients over 500 Monte Carlo replications, and results were scaled to 1,000 patients.

Standard care. HDU admission for patients with ASA ≥ 4, those on preoperative respiratory support, or emergency major surgery in patients aged ≥ 70 years (26.5% of patients). Ward monitoring uses NEWS2.

AI-guided pathway.

1. **Timing**: elective, non-cancer patients with predicted risk ≥ 0.30 and infection < 7 weeks before surgery have surgery deferred to ≥ 7 weeks, by when their acute inflammatory markers are assumed to have resolved [2].
2. **Level of care**: patients with predicted risk ≥ 0.30 (18.4%) are admitted to HDU for 2 days and receive a pulmonary-complication prevention bundle (base-case relative risk reduction 15%) [25].
3. **Early discharge**: patients with predicted risk < 0.05 (25.0%) follow an ERAS pathway with remote monitoring, which shortens uncomplicated stay by 1 day [23].
4. **Surveillance**: the ML early-warning model replaces NEWS2, reducing failure to rescue (death after complication) by 15% from a baseline of 22% [26].

Length of stay (LOS) was modelled as log-normal (median 5 days for major and 2.2 days for minor surgery), plus gamma-distributed extra days after a complication (mean 8 days). 40% of complications require ICU (mean 5 days). Because intervention effect sizes are assumptions, a one-way sensitivity analysis varied the bundle effect (0–30%), the rescue gain (0–30%), deferral (on or off) and the HDU threshold (0.20–0.40).

4. RESULTS

The gradient-boosting model achieved the best discrimination (AUROC 0.840) and was well calibrated. The ML early-warning model detected 97% of deteriorating patients at a false-alert rate similar to NEWS2 ≥ 7, which detected 79%. In simulation, the AI-guided pathway reduced major complications by 18% and HDU bed-days by 30% relative to standard care.

4.1 Cohort characteristics

The 10,000 synthetic patients had a mean age of 57.8 years; 48.4% were male, 39.7% underwent emergency surgery and 55.3% major surgery, and 24.4% had a malignant diagnosis. Surgery took place within 2 weeks of SARS-CoV-2 diagnosis in 45% of patients, and 59.9% were vaccinated. The major-complication rate was 17.97%, falling from 24.8% in patients operated on within 2 weeks of diagnosis to 22.3% at 3–4 weeks, 13.2% at 5–6 weeks and 7.1% at ≥ 7 weeks, consistent with the timing gradient reported internationally [2].

4.2 Preoperative risk prediction

Table 2: Test-set performance of the preoperative models (n = 3,000; 539 events).

Model	AUROC (95% CI)	AUPRC	Brier score	Calibration slope	Specificity at 80% sensitivity	PPV	NPV
Clinical score (4-variable LR)	0.739 (0.713–0.762)	0.438	0.127	0.98	0.507	0.263	0.921
Logistic regression (19 variables)	0.830 (0.811–0.849)	0.568	0.110	0.99	0.691	0.362	0.941
Random forest	0.828 (0.808–0.847)	0.575	0.110	1.28	0.709	0.376	0.942
XGBoost	**0.840 (0.821–0.858)**	**0.592**	**0.107**	**1.14**	**0.735**	**0.399**	**0.945**

AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision–recall curve; PPV, positive predictive value; NPV, negative predictive value; LR, logistic regression. The best-performing model is shown in bold.

The upper bound of achievable discrimination, given by the true data-generating probabilities, was AUROC 0.850; XGBoost came within 0.010 of it. All three multivariable models improved substantially on the four-variable clinical score: XGBoost gained 0.101 in AUROC (paired bootstrap 95% CI 0.083–0.121). This gain comes from the laboratory, physiological and timing variables that the clinical score omits. XGBoost outperformed full logistic regression by a small but statistically consistent margin (ΔAUROC 0.010, 95% CI 0.002–0.017) and random forest by 0.012 (0.006–0.018). At 80% sensitivity, XGBoost flagged 36% of patients against 55% for the clinical score, so the same number of complications is captured with about a third fewer patients flagged.

Calibration-in-the-large was close to zero for all models ($|O - E| < 0.002$). The calibration slope of the random forest (1.28) shows that its predictions were too compressed towards the mean, a known property of bagged trees; XGBoost (1.14) was close to ideal, and Fig. 2 shows close agreement between predicted and observed risk across deciles.

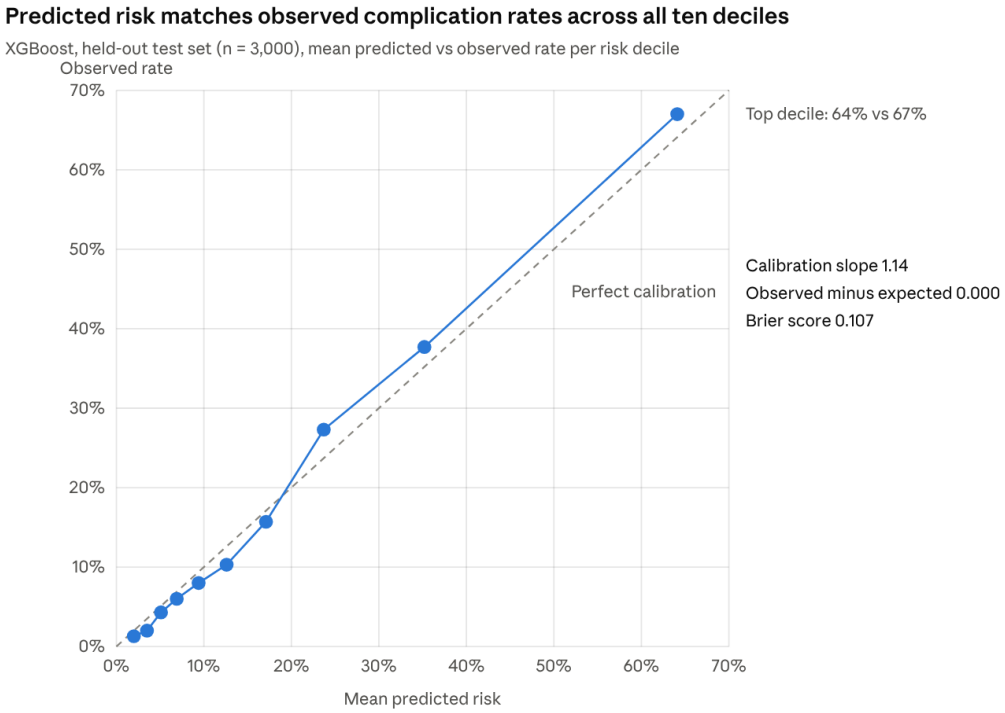


Figure 2: Calibration of the XGBoost model by decile of predicted risk; the dashed line marks perfect calibration.

In decision-curve analysis, XGBoost had the highest net benefit at every threshold probability from 0.10 to 0.40; for example, 0.104 at a 15% threshold, against 0.072 for the clinical score and 0.035 for treating all patients. This is equivalent to identifying 32 additional patients who will develop complications per 1,000 assessed, with no increase in unnecessary interventions, compared with the clinical score.

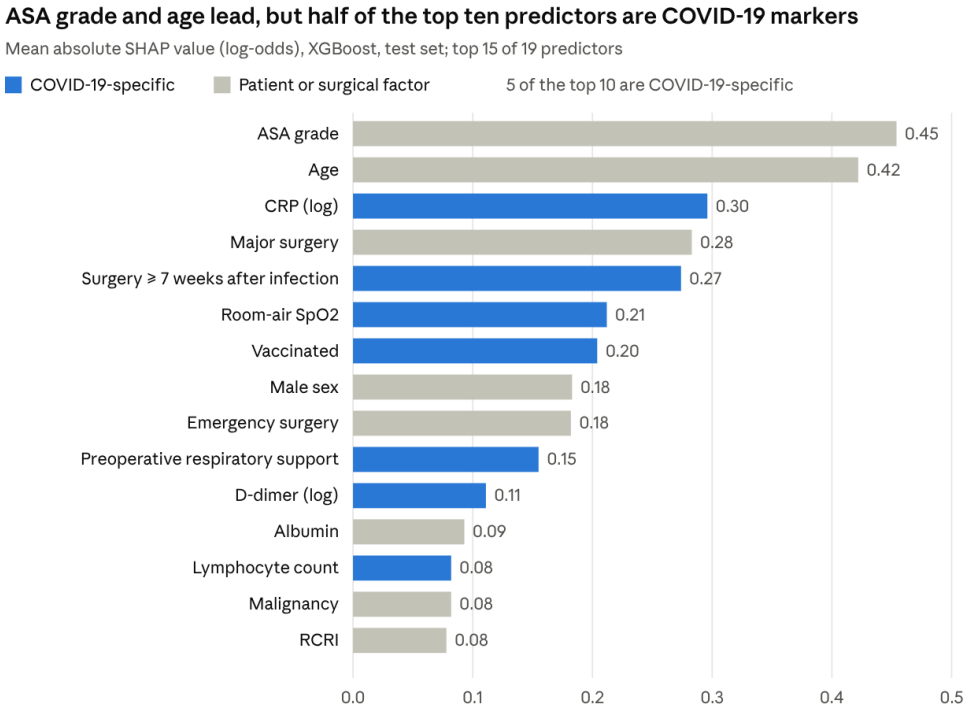


Figure 3: Global feature importance of the XGBoost model (mean absolute SHAP value). COVID-19-specific predictors are highlighted.

SHAP analysis (Fig. 3) ranked ASA grade and age as the most influential predictors, followed by CRP, surgical grade and surgery ≥ 7 weeks after infection (protective). Room-air SpO_2, vaccination and sex followed. These rankings agree with the published risk factors [1, 2, 5] and confirm that the model's reasoning can be inspected and challenged by clinicians.

Table 3: XGBoost performance by subgroup (test set).

Subgroup	n	Event rate	AUROC	Sensitivity at global threshold	O/E ratio
Female	1,538	15.0%	0.855	0.783	0.94
Male	1,462	21.1%	0.820	0.819	1.05
Age < 70 years	2,226	13.5%	0.834	0.713	0.98
Age ≥ 70 years	774	30.9%	0.796	0.916	1.03
Elective	1,819	14.7%	0.855	0.787	0.93
Emergency	1,181	22.9%	0.813	0.819	1.09
Vaccinated	1,789	15.4%	0.856	0.782	1.01
Unvaccinated	1,211	21.8%	0.814	0.826	0.99

O/E, observed-to-expected ratio.

Discrimination was lower in higher-risk groups (age ≥ 70, emergency, unvaccinated), as expected when the outcome is more frequent and risk is more homogeneous. The more important finding is the variation in sensitivity at a single global threshold: 71% in patients under 70 compared with 92% in those aged 70 or over. A threshold that is fair on average can therefore under-detect complications in younger patients. O/E ratios between 0.93 and 1.09 indicate acceptable calibration in all subgroups, with slight under-prediction for emergency patients.

4.3 Postoperative early warning

At the observation level, the ML early-warning model predicted complication onset within 24 h with AUROC 0.921, compared with 0.829 for NEWS2. The most important ML features were deviation of NEWS2 from the patient's own baseline, 24-h change in NEWS2, current NEWS2 and preoperative risk. This shows that the model's gain comes mainly from personalising and trending existing observations.

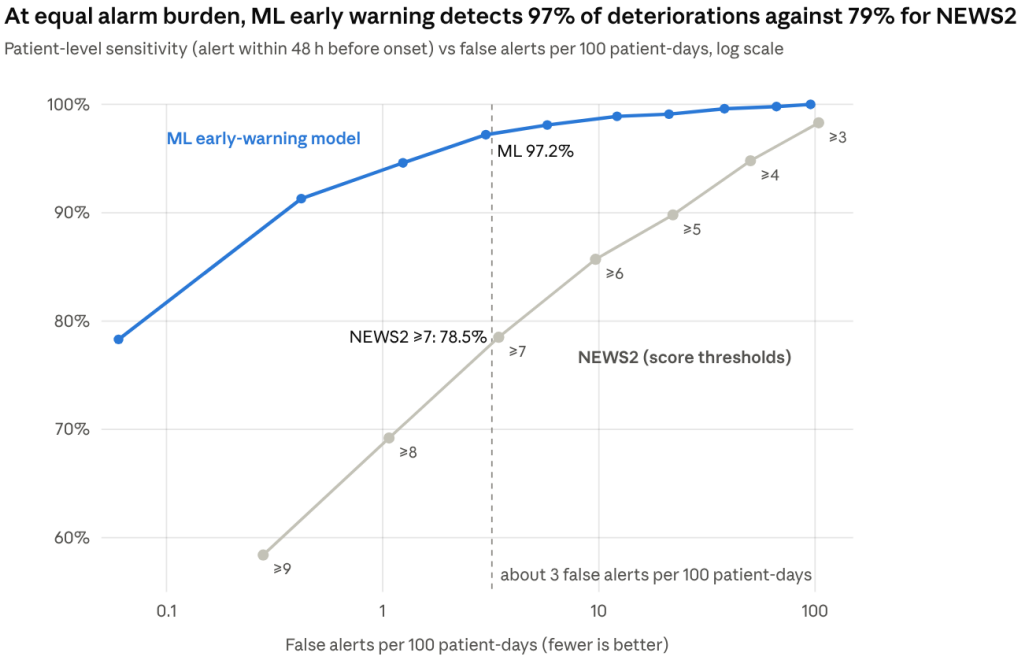


Figure 4: Trade-off between detection and alarm burden for NEWS2 score thresholds (grey) and the ML early-warning model across probability thresholds (blue).

Table 4: Patient-level performance at selected operating points.

Alerting rule	Sensitivity	False alerts per 100 patient-days	Median lead time (h)
NEWS2 ≥ 5	0.898	22.0	12.7
NEWS2 ≥ 7	0.785	3.4	7.7
ML, threshold 0.08	0.991	21.1	21.4
ML, threshold 0.30	0.972	3.0	10.6
ML, threshold 0.60	0.913	0.4	7.7

Fig. 4 shows the trade-off between detection and alarm burden. At a false-alert rate similar to NEWS2 ≥ 5 (about 21 per 100 patient-days), the ML model detected 99.1% of deteriorating patients with a median lead time of 21.4 h, against 89.8% and 12.7 h for NEWS2. At a false-alert rate similar to NEWS2 ≥ 7 (about 3 per 100 patient-days), the ML model detected 97.2% of patients against 78.5%, with 2.9 h more warning. Alternatively, the ML model matched the sensitivity of NEWS2 ≥ 5 (91%) with about 50 times fewer false alerts (0.4 against 22.0 per 100 patient-days).

4.4 AI-guided care pathway

Table 5: Simulated outcomes per 1,000 patients (mean of 500 replications; 95% simulation interval).

Outcome per 1,000 patients	Standard care	AI-guided pathway	Change
Major complications	180 (169–192)	148 (137–158)	−18.0%
Deaths	39.6 (33.3–47.0)	27.6 (21.8–33.2)	−30.3%
Total hospital bed-days	5,530 (5,393–5,667)	5,033 (4,902–5,159)	−9.0%
Mean length of stay (days)	5.53	5.03	−0.50 days
ICU bed-days	358 (308–410)	293 (252–340)	−18.0%
HDU bed-days	512	357	−30.3%
Patients admitted to HDU	265	184	−30.8%
Low-risk patients (< 10%) in HDU	36	0	−100%
Elective surgeries deferred ≥ 7 weeks	0	52	–
Patients on early-discharge pathway	0	250	–

Values are means with 95% simulation intervals in parentheses. HDU, high-dependency unit; ICU, intensive care unit.

The AI pathway used fewer HDU beds while targeting them better. Given the same HDU capacity as standard care (26.5% of patients), ranking by predicted risk would capture 69.0% of all complications in HDU compared with 53.1% for the rule-based criteria, and raise the event rate among HDU patients from 35.9% to 46.7%. The 25% of patients routed to early discharge had an observed complication rate of only 2.4%, supporting the safety of a reduced-intensity pathway for this group.

Table 6: One-way sensitivity analysis of the AI-guided pathway (change relative to standard care).

Scenario	Complications	Deaths	Bed-days	HDU bed-days
Base case	−18.1%	−30.5%	−8.9%	−30.3%
Prevention bundle, RRR 0%	−11.4%	−24.9%	−7.2%	−30.3%
Prevention bundle, RRR 30%	−24.7%	−36.1%	−10.6%	−30.3%
Rescue gain 0%	−18.1%	−18.3%	−8.9%	−30.3%
Rescue gain 30%	−18.1%	−42.6%	−8.9%	−30.3%
No surgical deferral	−8.4%	−22.4%	−6.4%	−30.3%
HDU threshold 0.20	−23.2%	−35.1%	−10.3%	+15.0%
HDU threshold 0.40	−13.9%	−26.9%	−7.8%	−53.9%

RRR, relative risk reduction.

The direction of benefit was robust: complications, deaths and bed-days fell in every scenario. Surgical deferral was the largest single contributor to complication reduction (removing it halved the effect, from −18.1% to −8.4%), even though it affected only 5.2% of patients. The HDU threshold is a capacity lever. Lowering it to 0.20 prevented more complications (−23.2%) but used 15% more HDU bed-days than standard care. Raising it to 0.40 halved HDU use while still reducing complications by 13.9%.

5. DISCUSSION

5.1 Principal findings

This study suggests that AI can improve postoperative care for COVID-19 surgical patients at three points, and that the operational value comes from linking them. Preoperative ML risk models identified patients at risk far better than a four-variable clinical score. ML early warning caught deterioration earlier with fewer false alarms than NEWS2. Using both to set surgical timing, level of care and discharge pathway reduced simulated complications, deaths and bed use together. Efficiency and safety moved in the same direction: the AI pathway used 30% fewer HDU bed-days and still prevented more complications, because it moved monitoring capacity away from low-risk patients selected by blunt rules towards high-risk patients selected by individual risk.

5.2 Model choice: data matter more than algorithm

The largest gain came from richer inputs, not from a more complex algorithm. Adding laboratory, physiological and timing variables raised AUROC by about 0.09 (0.739 to 0.830 with the same logistic method), while switching from logistic regression to XGBoost added only 0.010. This mirrors the COVIDSurg experience, where tree-based methods did not outperform logistic regression on a small set of predictors [4], and the finding of McDermott et al. [5] that gradient boosting performs best on rich registry data. For hospitals, the practical implication is that the reliable capture of CRP, lymphocyte count, SpO_2, vaccination status and the date of SARS-CoV-2 diagnosis matters more than the choice of model. Well-specified logistic regression remains a credible, more transparent alternative where data or governance capacity is limited.

5.3 Early warning: personalise, trend and set thresholds deliberately

NEWS2 uses fixed population thresholds, which penalises COVID-19 patients whose baseline respiratory rate and oxygen saturation are already abnormal. The ML model's most important features were deviations from the patient's own baseline and 24-h changes, in line with prior ward-deterioration work [6, 7]. In practice, the alert threshold should be chosen by the clinical team as a trade-off between missed deterioration, lead time and nursing workload, using a curve such as Fig. 4, rather than inherited from a vendor default. The early-warning results are, however, the most optimistic in this study, because the simulated deterioration is a smooth, monotonic drift. Real deterioration is noisier, observations are missed, and vital signs are affected by interventions such as oxygen titration. External validation on real trajectories is therefore essential before any claim of superiority over NEWS2, as the experience with proprietary deterioration and sepsis models shows [18, 19].

5.4 Where the efficiency gains come from

The sensitivity analysis isolates three mechanisms.

1. **Timing.** Deferring high-risk elective, non-cancer surgery beyond seven weeks contributed most to complication reduction, although it affected only about 5% of patients. This is consistent with COVIDSurg guidance [2], and AI makes it selective, so that only patients whose individual risk justifies the delay are deferred and backlogs do not grow unnecessarily.
2. **Targeted escalation.** Ranking by risk captured 69% of complications in HDU compared with 53% for rule-based criteria at the same capacity. HDU capacity can then be tuned as an explicit lever, from −54% to +15% HDU use across the thresholds tested.
3. **De-escalation.** One in four patients had a predicted risk below 5% and an observed complication rate of 2.4%. These patients are candidates for ERAS early discharge with remote monitoring [22, 23], which frees beds for the backlog.

The mortality reduction depends most on the assumed rescue gain (−18% to −43% across 0–30% gain). This reflects the evidence that failure to rescue drives variation in surgical mortality [26], but it is also the least certain parameter and the one most in need of prospective measurement.

5.5 Implementation, fairness and governance

Several conditions need to be met before such a framework can be used clinically.

- **Validation and recalibration.** Models must be validated locally and recalibrated as variants, vaccination coverage and treatments change the baseline risk. Calibration should be monitored continuously, because a model that drifts towards over-prediction will silently fill HDU beds.
- **Subgroup thresholds.** The lower sensitivity in patients under 70 (71% against 92%) shows that one global threshold can distribute missed cases unevenly. Subgroup reporting [29] and, where justified, group-specific thresholds or recalibration should be routine. Proxy variables that encode access to care rather than need must be avoided [28].

- **Explainability and human oversight.** SHAP explanations allow surgeons and anaesthetists to see why a patient was flagged and to override the recommendation. The model should support decisions, not make them [30, 36].
- **Regulation.** Software that informs triage and level-of-care decisions is likely to be regulated as a medical device and, in the European Union, as a high-risk AI system under the AI Act [37], with requirements for risk management, data governance, transparency and post-market monitoring.
- **Workflow.** Alerts must reach a clinician who can act on them; an early warning without a rapid-response pathway adds workload without reducing failure to rescue.

5.6 Limitations

The main limitation is that the cohort, vital-sign trajectories and intervention effects are simulated. The results therefore show the framework's potential performance under stated assumptions, not observed clinical effects. Specifically: (i) discrimination is bounded by the data-generating model (oracle AUROC 0.850) and real-world performance will probably be lower; (ii) published odds ratios were used as generative coefficients, which simplifies their confounding structure; (iii) the rescue gain and bundle effect are assumptions informed by the literature [25, 26], tested only by one-way sensitivity analysis; (iv) missing data, EHR measurement error, temporal drift and alert fatigue were not modelled; and (v) the cost of HDU, remote monitoring and the effect of deferral on cancer progression or patient preference were not modelled. Each limitation defines a step in the validation path set out in Section 6.

6. CONCLUSION AND FUTURE WORK

This paper proposed and evaluated an integrated AI framework for postoperative care of surgical patients with perioperative SARS-CoV-2 infection. It covers preoperative risk stratification, ML-based postoperative early warning and an AI-guided care pathway. On a 10,000-patient synthetic cohort whose risk structure follows published international data, the main findings were:

1. **Risk prediction.** Gradient boosting (XGBoost) achieved AUROC 0.840 with good calibration, compared with 0.739 for a four-variable clinical score. Most of the gain came from routinely available laboratory, physiological and infection-timing variables, and the model's explanations agreed with established risk factors.
2. **Early warning.** A personalised, trend-based ML model detected 97% of deteriorating patients at about 3 false alerts per 100 patient-days, compared with 79% for NEWS2 ≥ 7 at a similar alert burden, and gave longer warning.
3. **Care pathway.** Acting on these predictions through selective deferral, targeted HDU admission with prevention bundles, early discharge for low-risk patients and ML surveillance reduced simulated major complications by 18%, deaths by 30%, total bed-days by 9% and HDU bed-days by 30%. The direction of benefit was robust across all sensitivity scenarios.
4. **Fairness.** Sensitivity at a single global threshold varied between subgroups (71% in patients under 70 against 92% in those 70 or older), so subgroup monitoring must be part of any deployment.

These results come from a simulation and show what the framework could achieve. Their value lies in the design they establish (what to measure, which thresholds matter and which assumptions drive the benefit) rather than in the effect sizes themselves.

Future work. The next steps are:

1. retrospective external validation of the preoperative and early-warning models on multi-centre EHR data from COVID-19 surgical patients, with temporal validation across pandemic waves and variants;
2. a silent prospective deployment to measure real-world calibration, alert burden and lead time without influencing care;
3. a stepped-wedge cluster trial of the AI-guided pathway, with failure to rescue, length of stay and HDU utilisation as outcomes;
4. extension to continuous wearable monitoring and to post-discharge remote monitoring;
5. health-economic evaluation, including the cost of HDU days, monitoring technology and deferral; and
6. transfer of the framework to other respiratory infections, such as influenza and RSV, and to future pandemic preparedness.

REFERENCES

1. COVIDSurg Collaborative. Mortality and pulmonary complications in patients undergoing surgery with perioperative SARS-CoV-2 infection: an international cohort study. Lancet. 2020;396(10243):27–38.
2. COVIDSurg Collaborative, GlobalSurg Collaborative. Timing of surgery following SARS-CoV-2 infection: an international prospective cohort study. Anaesthesia. 2021;76(6):748–758. doi:10.1111/anae.15458
3. COVIDSurg Collaborative. Elective surgery cancellations due to the COVID-19 pandemic: global predictive modelling to inform surgical recovery plans. Br J Surg. 2020;107(11):1440–1449.
4. COVIDSurg Collaborative. Machine learning risk prediction of mortality for patients undergoing surgery with perioperative SARS-CoV-2: the COVIDSurg mortality score. Br J Surg. 2021;108(11):1274–1292. PMCID: PMC8344569

5. McDermott SP, Zhou L, Kaynar AM. Development and validation of a machine learning model to predict postoperative morbidity and mortality in patients with COVID-19 who underwent major surgery. Cureus. 2025;17(8):e91247. PMCID: PMC12477459
6. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Crit Care Med. 2016;44(2):368–374.
7. Escobar GJ, Liu VX, Schuler A, Lawson B, Greene JD, Kipnis P. Automated identification of adults at risk for in-hospital clinical deterioration. N Engl J Med. 2020;383(20):1951–1960.
8. Vázquez-Serrano JI, Peimbert-García RE, Cárdenas-Barrón LE. Discrete-event simulation modeling in healthcare: a comprehensive review. Int J Environ Res Public Health. 2021;18(22):12262.
9. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020;369:m1328.
10. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021;3:199–217.
11. COVIDSurg Collaborative, GlobalSurg Collaborative. SARS-CoV-2 vaccination modelling for safe surgery to save lives: data from an international prospective cohort study. Br J Surg. 2021;108(9):1056–1063.
12. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, Motaei A, Madkour M, Pardalos PM, et al. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. Ann Surg. 2019;269(4):652–662.
13. Bertsimas D, Dunn J, Velmahos GC, Kaafarani HMA. Surgical risk is not linear: derivation and validation of a novel, user-friendly, and machine-learning-based Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) calculator. Ann Surg. 2018;268(4):574–583.
14. Knight SR, Ho A, Pius R, Buchan I, Carson G, Drake TM, et al. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. BMJ. 2020;370:m3339.
15. Bertsimas D, Lukin G, Mingardi L, Nohadani O, Orfanoudaki A, Stellato B, et al. COVID-19 mortality risk assessment: an international multi-center study. PLoS One. 2020;15(12):e0243262.
16. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230.
17. Royal College of Physicians. National Early Warning Score (NEWS) 2: standardising the assessment of acute-illness severity in the NHS. Report of a working party. London: RCP; 2017.
18. Singh K, Valley TS, Tang S, Li BY, Kamran F, Sjoding MW, et al. Evaluating a widely implemented proprietary deterioration index model among hospitalized patients with COVID-19. Ann Am Thorac Soc. 2021;18(7):1129–1137.
19. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med. 2021;181(8):1065–1070.
20. Downey CL, Chapman S, Randell R, Brown JM, Jayne DG. The impact of continuous versus intermittent vital signs monitoring in hospitals: a systematic review and narrative synthesis. Int J Nurs Stud. 2018;84:19–27.
21. Michard F, Kalkman CJ. Rethinking patient surveillance on hospital wards. Anesthesiology. 2021;135(3):531–540.
22. McGillion MH, Parlow J, Borges FK, Marcucci M, Jacka M, Adili A, et al. Post-discharge after surgery Virtual Care with Remote Automated Monitoring-1 (PVC-RAM-1) technology versus standard care: randomised controlled trial. BMJ. 2021;374:n2209.
23. Ljungqvist O, Scott M, Fearon KC. Enhanced recovery after surgery: a review. JAMA Surg. 2017;152(3):292–298.
24. Fernandez-Bustamante A, Frendl G, Sprung J, Kor DJ, Subramaniam B, Martinez Ruiz R, et al. Postoperative pulmonary complications, early mortality, and hospital stay following noncardiothoracic surgery: a multicenter study by the Perioperative Research Network investigators. JAMA Surg. 2017;152(2):157–166.
25. Cassidy MR, Rosenkranz P, McCabe K, Rosen JE, McAneny D. I COUGH: reducing postoperative pulmonary complications with a multidisciplinary patient care program. JAMA Surg. 2013;148(8):740–745.
26. Ghaferi AA, Birkmeyer JD, Dimick JB. Variation in hospital mortality associated with inpatient surgery. N Engl J Med. 2009;361(14):1368–1375.
27. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NeurIPS 2017); 2017. p. 4765–4774.
28. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–453.
29. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
30. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: WHO; 2021.
31. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng. 2021;5:493–497.
32. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
33. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 785–794.

34. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. 2nd ed. Cham: Springer; 2019.
35. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565–574.
36. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44–56.
37. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Off J Eur Union. 2024 Jul 12;L series.