Cureus
Part of SPRINGER NATURE

Open Access Review Article

Artificial Intelligence Versus Logistic Regression for Postoperative Mortality Prediction: A Systematic Review and Meta-Analysis

Review began 09/03/2026
Review ended 10/14/2026
Published 10/20/2026

© **Copyright** 2026
Escandón-Villalobos et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.118237

Edson A. Escandón-Villalobos [1], Rebeca G. Elguezabal-Rodelo [2], Miguel A. Pardiño Vega [1], Lia Medina-Montalvo [1], Gyselle A. Godinez-Castillo [1], Eduardo E. Montalvo-Jave [1], Regina Faes Petersen [1], Eduardo Villegas-Tovar [1]

1. General Surgery, Hospital Médica Sur, Mexico City, MEX 2. Anesthesiology and Perioperative Medicine, Hospital Médica Sur, Mexico City, MEX

Corresponding author: Edson A. Escandón-Villalobos, edson.escandon@gmail.com

Abstract

Postoperative mortality remains a major global health burden, and accurate risk prediction is essential for perioperative decision-making and quality improvement. We conducted a systematic review and meta-analysis comparing artificial intelligence (AI)-based models with logistic regression (LR) for postoperative mortality prediction in adult surgical populations. PubMed, Excerpta Medica database (Embase), and the Cochrane Library were searched from inception to February 2, 2026. The review was registered in the International Prospective Register of Systematic Reviews (PROSPERO) on May 5, 2026 (registration number: CRD420261365093). Eligible studies directly compared AI-based and LR models within the same adult surgical population and reported the area under the receiver operating characteristic curve (AUC) for 30-day and/or 90-day postoperative mortality. Risk of bias was assessed using the Prediction model Risk Of Bias Assessment Tool (PROBAST). A random-effects meta-analysis with restricted maximum likelihood and Hartung-Knapp adjustment was performed to pool the between-model difference in AUC (ΔAUC). The search identified 977 records; after removal of 70 duplicates, 907 records were screened, 68 reports were assessed for eligibility, and 11 retrospective cohort studies including 372,343 participants were included. All studies showed high or unclear risk of bias. Seven studies contributed to the 30-day mortality synthesis, in which AI-based models showed better discrimination than LR (ΔAUC: 0.0575, 95% confidence interval (CI): 0.0128-0.1022; I^2 = 89.0%). Six studies contributed to the 90-day mortality synthesis, also favoring AI-based models (ΔAUC: 0.0509, 95% CI: 0.0078-0.0940; I^2 = 39.9%). AI-based models showed modest pooled improvements in discrimination compared with LR for postoperative mortality prediction. However, confidence in these findings is limited by high or unclear risk of bias, the small number of eligible studies, and heterogeneity, particularly in the 30-day analysis.

Categories: General Surgery, Healthcare Technology
Keywords: ai and machine learning, artificial intelligence, artificial intelligence in surgery, perioperative mortality, surgery outcomes, systematic review and meta-analysis

Introduction And Background

Postoperative mortality represents a major global health burden, although its precise magnitude is difficult to quantify across healthcare systems. One modeling study estimated that approximately 4.2 million deaths occur within 30 days of surgery each year, which, if accurate, would place postoperative mortality among the leading causes of death worldwide [1]. Reliable preoperative risk stratification is therefore essential to guide perioperative decision-making, optimize patient outcomes, support shared decision-making, and improve health system efficiency.

Traditional surgical risk prediction models have relied predominantly on logistic regression (LR), a conventional statistical modeling approach that estimates the probability of an outcome from a specified relationship between predictors and the log-odds of that outcome. LR remains widely used because of its interpretability, simplicity, and clinical familiarity. However, its predictive performance depends on appropriate specification of the relationships between predictors and the log-odds of the outcome, including relevant non-linear effects and interactions. Inadequate handling of these relationships may limit model performance, particularly in heterogeneous patient populations [2].

In contrast, artificial intelligence (AI)-based models, including machine learning (ML) approaches, offer flexible methods for handling high-dimensional data and modeling complex relationships among predictors [3,4]. In this context, ML refers to data-driven modeling approaches that learn predictive relationships from observed data. AI/ML approaches are increasingly being investigated for perioperative and surgical risk prediction, but comparative studies have reported variable incremental performance over conventional methods. Chiew et al. evaluated several ML algorithms for predicting 30-day postoperative mortality and prolonged intensive care unit admission in more than 90,000 surgical patients and reported improved predictive performance compared with established surgical risk models [5]. In a direct comparison using the

same preoperative predictors, Michelsen et al. found only a modest improvement in discrimination with boosted decision trees compared with logistic regression, illustrating that the advantage of ML may be context-dependent [6].

This uncertainty is also reflected in the broader literature on perioperative prediction models. A systematic review of preoperative models for 30-day perioperative mortality found substantial variability in discrimination, with c-statistics ranging from 0.82 to 0.96. Calibration was reported in only a minority of studies, and all model-development studies were judged to have unclear or high risk of bias [7]. These findings suggest that ML may improve discrimination in selected surgical populations, but the magnitude of its incremental performance over LR remains uncertain and appears to depend on the population, predictors, modeling strategy, and validation approach. The previous review focused on 30-day postoperative mortality across heterogeneous non-cardiac surgical populations [7], whereas evidence from direct within-cohort comparisons of AI/ML and LR models, particularly for 90-day mortality, has been less extensively synthesized.

To address this gap, we conducted a systematic review and meta-analysis of studies directly comparing AI/ML-based models with LR within the same surgical population, with separate analyses for 30-day and 90-day postoperative mortality. The review question was structured using a PICO-style framework adapted to prediction-model research: the population (P) comprised adults undergoing abdominal or visceral surgery; the index prediction approach (I) consisted of AI/ML-based models; the comparator (C) was LR evaluated within the same study population; and the outcome (O) was discriminative performance for 30-day or 90-day postoperative mortality. Discrimination was measured using the area under the receiver operating characteristic curve (AUC), which reflects a model's ability to distinguish between patients who do and do not experience the outcome.

Review

Materials and methods

This meta-analysis was performed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [8], and all methodological procedures followed the recommendations of the Cochrane Handbook for Systematic Reviews of Interventions, version 6.3 [9]. The protocol was registered in the International Prospective Register of Systematic Reviews (PROSPERO) database on May 5, 2026 (registration number: CRD420261365093). Because this review analyzed previously published aggregate data and did not involve individual patient participation, ethics committee approval and informed consent were not required. The study was conducted in accordance with the ethical standards of the Declaration of Helsinki as revised in Brazil in 2013 [10].

Search Strategy

A systematic literature search was conducted in PubMed, Excerpta Medica database (EMBASE), and the Cochrane Library from database inception through February 2, 2026. The search strategy combined database-specific controlled vocabulary terms and free-text keywords related to AI, ML, LR, surgery, postoperative mortality, and risk prediction. Boolean operators ("AND" and "OR") were used to combine the search concepts as appropriate for each database. The complete search syntax for each database, including the specific search terms, field tags, and Boolean combinations used, is provided in the Appendices to allow reproducibility. In addition, the reference lists of all included studies and relevant reviews were manually screened to identify potentially eligible studies not retrieved through the electronic database searches.

Study Selection

All retrieved records were imported into Zotero reference management software (Corporation for Digital Scholarship, Vienna, VA, USA) [11], and duplicate records were removed. Study selection was conducted independently by two reviewers (L.M.M. and G.G.C.) in two stages: first, titles and abstracts were screened against the predefined eligibility criteria described in the following section; second, the full texts of potentially eligible articles were assessed for inclusion. Disagreements were first resolved through discussion between the two reviewers. If consensus could not be reached, a third reviewer (E.A.E.V.) was consulted; unresolved disagreements were subsequently discussed with the remaining two authors (M.A.P.V. and R.G.E.R.) to reach a final decision. Reasons for exclusion at the full-text stage were documented and are summarized in the PRISMA flow diagram.

Eligibility Criteria

Studies were eligible for inclusion if they met all of the following criteria: (i) original research studies, including randomized controlled trials or observational cohort studies, published in English; (ii) inclusion of adult patients undergoing abdominal or visceral surgical procedures, including general, gastrointestinal, colorectal, hepatobiliary, bariatric, and abdominal oncological surgery; (iii) direct comparison of an AI/ML-based prediction model with an LR model evaluated within the same study population; and (iv) reporting of discriminative performance for 30-day and/or 90-day postoperative mortality using the AUC, with sufficient

information to extract or calculate the between-model difference in AUC and its corresponding measure of variance.

Studies were excluded if they were systematic reviews, meta-analyses, editorials, letters, conference abstracts, case reports, case series, or animal studies. Studies evaluating only one modeling approach without a direct within-cohort AI/ML-versus-LR comparison, or those lacking sufficient data to derive the required effect estimate or measure of uncertainty, were also excluded. When multiple publications were derived from the same cohort or database, the publication providing the most comprehensive data and methodological detail was retained to avoid duplicate inclusion.

Data Extraction and Model Selection

Data extraction was performed independently by two reviewers (L.M.M. and G.G.C.) using a standardized form based on the Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies (CHARMS) framework [12]. Extracted information included study and population characteristics, including age and sex; type of surgery; sample size; number of outcome events; mortality time frame; modeling approach; predictor variables; validation strategy; and model-performance measures. For the quantitative synthesis, the AUC and its corresponding 95% confidence interval (CI), standard error (SE), or other reported measure of uncertainty were extracted when available. Calibration metrics were also recorded when reported. Missing or unclear information was recorded as such in the original studies.

When multiple AI/ML algorithms were reported within the same study, the model designated by the original authors as the primary or final model was selected. If no model was explicitly designated, the model evaluated using the most rigorous validation strategy and, where possible, in the same validation cohort as the LR comparator was selected. Only one AI/ML-versus-LR comparison per study and mortality time point was included in each meta-analysis.

Risk of Bias and Applicability Assessment

Risk of bias and applicability were independently assessed by two reviewers (L.M.M. and G.G.C.) using the Prediction model Risk Of Bias Assessment Tool (PROBAST) [13]. PROBAST evaluates four domains: participants, predictors, outcome, and analysis. Each domain was rated as having low, high, or unclear risk of bias according to the corresponding signaling questions and guidance. Particular attention was given to methodological aspects relevant to prediction-model studies, including the handling of missing data, model overfitting, and validation strategy.

Because the included studies employed AI/ML methods, additional AI-specific considerations consistent with PROBAST-AI guidance [14] were considered, including model-tuning procedures, feature selection, and potential data leakage. Overall study-level risk of bias was determined according to PROBAST recommendations; a study was classified as being at high overall risk of bias if at least one domain was rated as high risk. Disagreements between reviewers were resolved through discussion and consensus. Individual study-level risk-of-bias and applicability assessments, as well as overall domain-level summaries, were graphically presented using the Risk-Of-Bias VISualization (robvis) web application [15].

Statistical Analysis

A pairwise meta-analysis was performed to compare the discriminative performance of AI-based models with that of conventional LR models for predicting 30-day and 90-day mortality. The primary outcome was the difference in AUC (ΔAUC), calculated for each study as the AUC of the AI-based model minus the AUC of the LR model. Studies were allocated to each synthesis according to the mortality time frame reported in the original article, with separate meta-analyses for 30-day and 90-day outcomes.

Effect sizes were derived from reported AUC values. When 95% CIs were available, SEs were approximated as (upper confidence limit - lower confidence limit)/(2 × 1.96), assuming approximate normality. When neither an SE nor a CI was reported, the SE of the AUC was estimated from the reported AUC and the numbers of participants with and without the outcome using the Hanley-McNeil approximation for the sampling variance of the area under the receiver operating characteristic curve [16]. When studies reported variability in AUC estimates across repeated resampling or validation procedures, the reported standard deviation (SD) of those repeated AUC estimates was used as an empirical approximation of uncertainty. Patient-level SDs were not converted into SEs.

Because AI and LR models were generally evaluated within the same patient cohorts, their AUC estimates were considered correlated, consistent with methods for comparing receiver operating characteristic areas derived from the same cases [17]. The variance of the difference in AUC was calculated as $SE^2(AI) + SE^2(LR) - 2\rho \times SE(AI) \times SE(LR)$, where ρ represents the within-study correlation between the paired AUC estimates. Because the covariance between the paired AUC estimates was not reported in the primary studies and could not be recovered from the available aggregate data, $\rho = 0.6$ was assumed as a pragmatic representation of moderate-to-high positive dependence. This value was treated as an analytical assumption rather than an

empirically validated universal correlation coefficient.

Meta-analyses were conducted using the inverse-variance method within a random-effects framework. Between-study variance (τ^2) was estimated with the restricted maximum likelihood estimator. Given the limited number of studies and the potential for heterogeneity, pooled Cls were calculated using the Hartung-Knapp adjustment to provide more conservative and robust inference. Publication bias and small-study effects were not formally assessed because fewer than 10 studies contributed to each meta-analysis, limiting the statistical power and interpretability of funnel-plot asymmetry tests. Statistical significance was defined as a two-sided p-value below 0.05. All analyses were conducted in R (R Foundation for Statistical Computing, Vienna, Austria) using the meta package [18].

Assessment of Heterogeneity

Heterogeneity was assessed using Cochran's Q test, with a p-value below 0.05 considered indicative of statistically significant heterogeneity. The I^2 statistic was used to quantify the percentage of variability in effect estimates attributable to between-study heterogeneity rather than sampling error and was interpreted as low (≤25%), moderate (26%-50%), substantial (51%-75%), or considerable (>75%).

Results

The database search identified 977 records. After removal of 70 duplicate records, 907 records were screened by title and abstract. Of these, 68 reports underwent full-text review against the predefined criteria. Eleven studies were ultimately included in the quantitative synthesis [19-29]. The PRISMA flow diagram (Figure *1*) summarizes the selection process, including the number of studies included and the reasons for exclusion at full-text review.

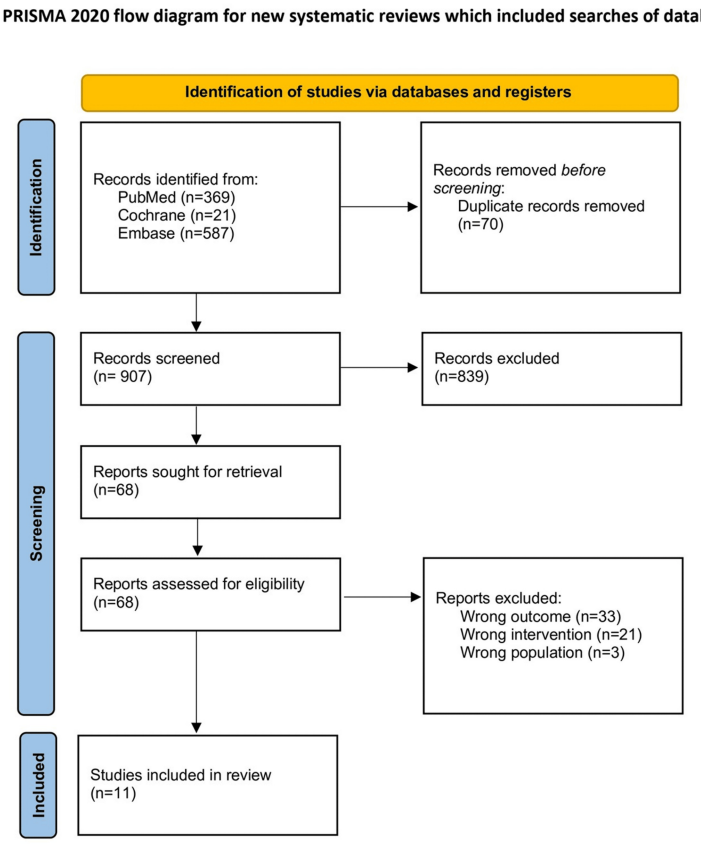


FIGURE 1: Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 flow diagram.

The diagram shows the identification, screening, eligibility assessment and inclusion of studies, together with the reasons for exclusion at full-text review.

General Study Characteristics

All included studies were observational cohort studies with retrospective data collection. More than half used multicentre datasets (seven of 11, 63.6%), and the remainder were single-centre studies (four of 11, 36.4%). The studies encompassed a range of populations undergoing abdominal procedures, including general, colorectal, bariatric, oncological, and upper gastrointestinal surgery, reflecting variability in clinical settings and patient populations (Table *1*).

First author	Year	Journal	Type of surgery	Study design	Sample size	Outcome time frame (days)
Abdollahzade et al. [19]	2025	Journal of Liver Transplantation	Liver transplantation	Single-centre retrospective cohort	454	30 and 90
van den Bosch et al. [20]	2021	JAMA Network Open	Oncological surgery	Multicentre retrospective cohort	62,501	30
Bunn et al. [21]	2020	Surgery	Appendicectomy	Multicentre retrospective cohort	223,214	30
Riepe et al. [22]	2025	International Journal of Surgery	General surgery	Multicentre retrospective cohort	7,711	30 and 90
Dal Cero et al. [23]	2024	Cancers	Oncological surgery	Multicentre retrospective cohort	2,546	90
Kong et al. [24]	2020	International Journal of Surgery	General surgery	Single-centre retrospective cohort	1,495	90
Pera et al. [25]	2022	Annals of Surgery	Gastric surgery	Multicentre retrospective cohort	3,182	90
Sahara et al. [26]	2021	Hepatobiliary Surgery and Nutrition	Oncological surgery	Multicentre retrospective cohort	63,507	30
Vasavada [27]	2023	medRxiv (preprint)	General surgery	Single-centre retrospective cohort	504	90
Wise et al. [28]	2023	Surgical Endoscopy	Bariatric surgery	Multicentre retrospective cohort	2,907	30
Yan et al. [29]	2022	Anesthesia & Analgesia	General surgery	Single-centre retrospective cohort	4,322	30

TABLE 1: Main characteristics of the included studies.

The primary outcome in all studies was postoperative mortality. Five studies (45.5%) assessed 30-day mortality, four (36.4%) evaluated 90-day mortality, and the remaining two (18.2%) reported both time frames. All studies developed predictive models from preoperative variables, and three studies (27.3%) also incorporated intraoperative data, demonstrating heterogeneity in model design and variable selection (Table *2*).

Study characteristic	n (%) or median (range)
Study design	
Cohort study	11 (100)
Outcome time frame	
30-day mortality	5 (45.5)
90-day mortality	4 (36.4)
Both	2 (18.2)
Type of data collection	
Retrospective	11 (100)
Prospective	0 (0)
Participant sampling	
Multicentre	7 (63.6)
Single centre	4 (36.4)
Sample size, median (range)	3,182 (454-223,214)
Predictors	
Preoperative only	8 (72.7)
Preoperative and intraoperative	3 (27.3)

TABLE 2: Summary of the study characteristics.

All studies evaluated at least one ML algorithm, most frequently in comparison with LR. Non-penalized multivariable LR was the most common regression approach (eight of 11, 72.7%), whereas penalized regression, specifically elastic net, was used in two studies (18.2%). In one study (9.1%), LR was incorporated into an ensemble framework rather than used as a standalone predictive model. The most frequently applied ML methods were random forest (seven of 11, 63.6%) and gradient boosting (six of 11, 54.5%), followed by support vector machines (four of 11, 36.4%) and neural network-based models (three of 11, 27.3%). Less frequently used approaches included decision trees (two of 11, 18.2%) and k-nearest neighbours (one of 11, 9.1%). Several studies also incorporated regularized regression and ensemble-based approaches within ML frameworks (Table *3*). Overall, the included studies showed substantial heterogeneity in population, sampling, sample size, and modeling strategy.

Model	n (%)
Logistic regression	
Non-penalized	8 (72.7)
Elastic net	2 (18.2)
Not used as a standalone model	1 (9.1)
Machine learning	
Random forest	7 (63.6)
Gradient boosting (XGBoost or GBM)	6 (54.5)
Support vector machine	4 (36.4)
Neural network (ANN or CNN)	3 (27.3)
Elastic net (within machine learning framework)	3 (27.3)
Decision tree	2 (18.2)
k-nearest neighbours	1 (9.1)
Other	3 (27.3)

TABLE 3: Overview of the modelling approaches used in the included studies.

Abbreviations: ANN, artificial neural network; CNN, convolutional neural network; GBM, gradient boosting machine; XGBoost, extreme gradient boosting. Categories are not mutually exclusive because studies could evaluate more than one model.

Risk of Bias Assessment

No study was classified as having a low overall risk of bias. Five studies (45.5%) were rated as high risk and six (54.5%) as unclear risk. The analysis domain was most frequently affected, primarily because of insufficient reporting of model-development procedures, absence of external validation, and potential overfitting. Several studies also showed unclear risk in the participants, predictors, and outcome domains, largely because of incomplete methodological reporting (Figures *2*, *3*).

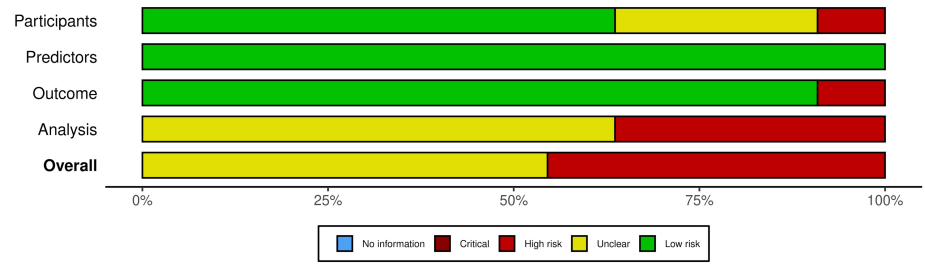


FIGURE 2: Summary of risk of bias across the 11 included studies.

Risk of bias was assessed with the Prediction model Risk Of Bias Assessment Tool (PROBAST). Bars show the proportion of studies rated as being at low, high or unclear risk of bias within each domain (participants, predictors, outcome and analysis) and overall.

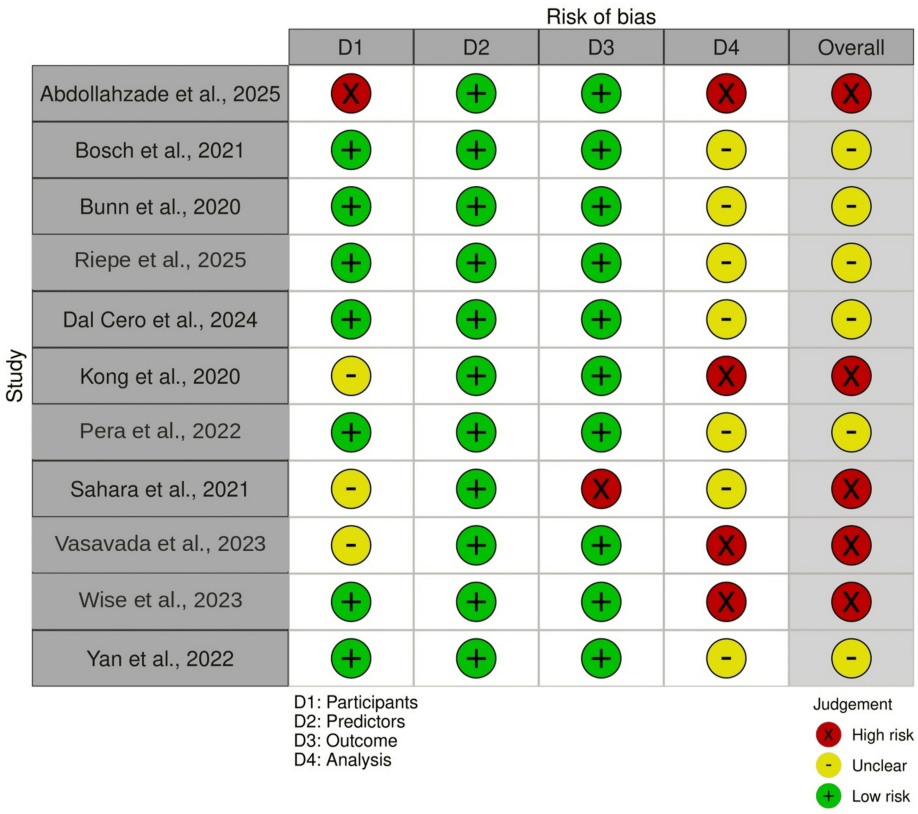


FIGURE 3: Study-level risk of bias assessment for each included study across the four Prediction model Risk Of Bias Assessment Tool (PROBAST) domains.

Studies included: [19-29]

In contrast, applicability concerns were generally low. Most studies showed adequate alignment between their populations, predictors, and outcomes and the objectives of this review. Minor concerns arose in a small number of studies, primarily because of highly selected populations or single-centre designs; these concerns were not considered sufficient to substantially limit overall applicability (Figures *4*, *5*).

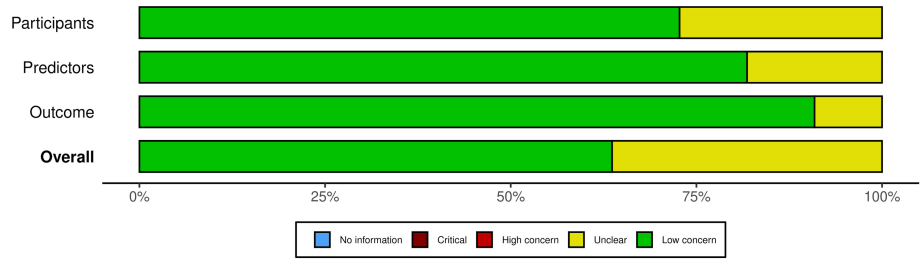


FIGURE 4: Summary of applicability concerns across the 11 included studies.

Applicability was assessed with the Prediction model Risk Of Bias Assessment Tool (PROBAST). Bars show the proportion of studies rated as being of low, high or unclear concern within each domain (participants, predictors and outcome) and overall.

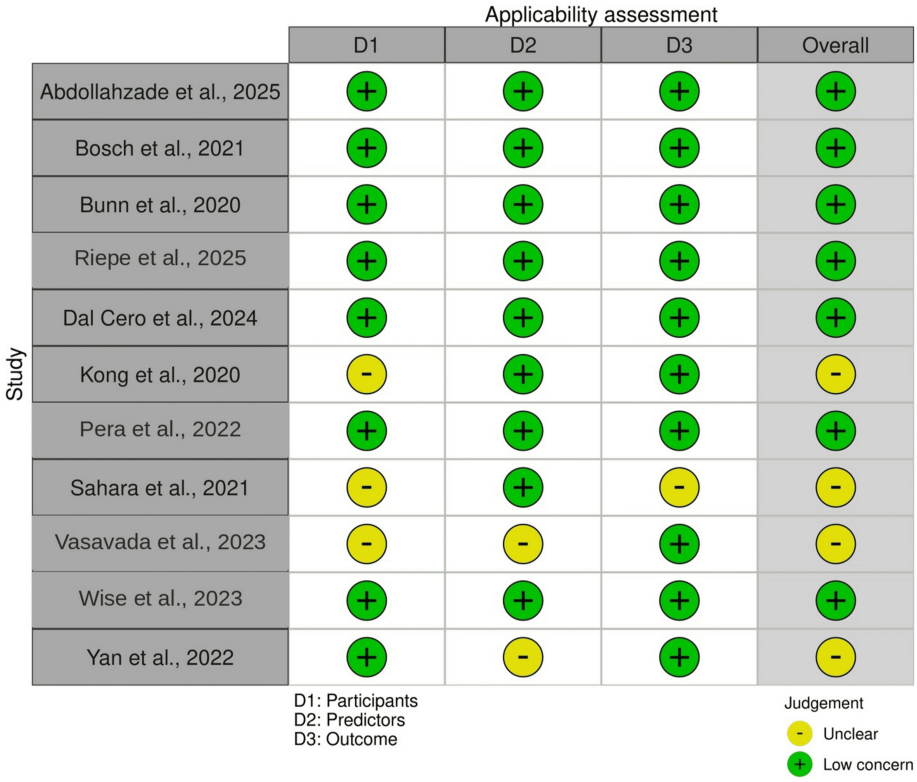


FIGURE 5: Study-level applicability assessment for each included study across the Prediction model Risk Of Bias Assessment Tool (PROBAST) domains.

Studies included: [19-29]

Thirty-Day Mortality Prediction Performance

Seven studies were included in the 30-day mortality analysis, comprising predominantly retrospective cohort studies with variable sample sizes and overall high or unclear risk of bias. The random-effects model with Hartung-Knapp adjustment showed a statistically significant improvement in discrimination favoring AI-based models over LR, with a pooled ΔAUC of 0.0575 (95% CI: 0.0128-0.1022; P = 0.020) (Figure *6*), corresponding to an average increase of approximately 0.06 in AUC. Considerable heterogeneity was observed (I^2 = 89.0%), indicating substantial variability in effect sizes across studies. Although the direction of effect consistently favored AI-based models, the magnitude of improvement differed across cohorts, suggesting contextual or methodological differences between studies.

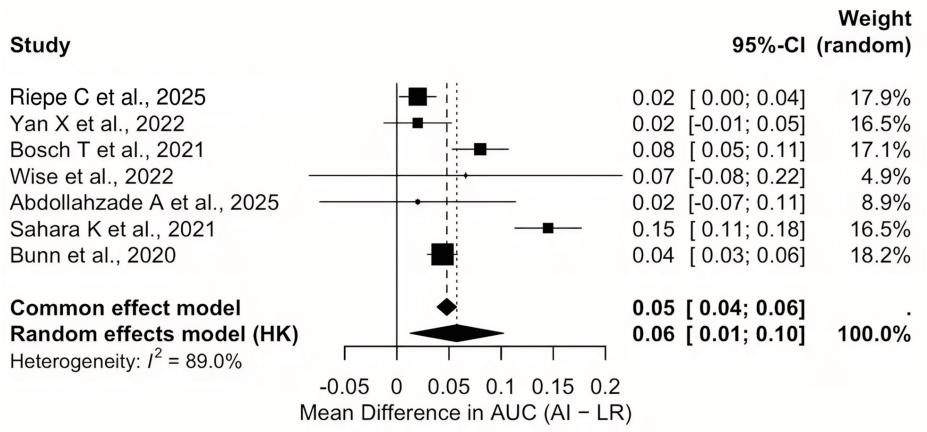


FIGURE 6: Forest plot of the difference in the area under the receiver operating characteristic curve (ΔAUC) between artificial intelligence-based models and logistic regression for 30-day postoperative mortality.

Squares represent study-level ΔAUC estimates and horizontal lines the corresponding 95% confidence intervals; the diamond represents the pooled random-effects estimate with the Hartung-Knapp (HK)-adjusted 95% confidence interval. Between-study variance (τ²) was estimated by restricted maximum likelihood. Values above zero favour artificial intelligence-based models.

Studies included: [19-22,26,28,29]

Ninety-Day Mortality Prediction Performance

Six retrospective cohort studies with overall high or unclear risk of bias were included in the 90-day mortality analysis. The random-effects model with Hartung-Knapp adjustment showed a pooled ΔAUC of 0.0509 (95% CI: 0.0078-0.0940; P = 0.029) (Figure 7), again demonstrating a statistically significant but modest improvement in discrimination with AI-based models. In contrast to the 30-day analysis, between-study heterogeneity was lower (I^2 = 39.9%), and Cochran's Q test was not statistically significant, indicating greater consistency of effect estimates across studies evaluating 90-day mortality.

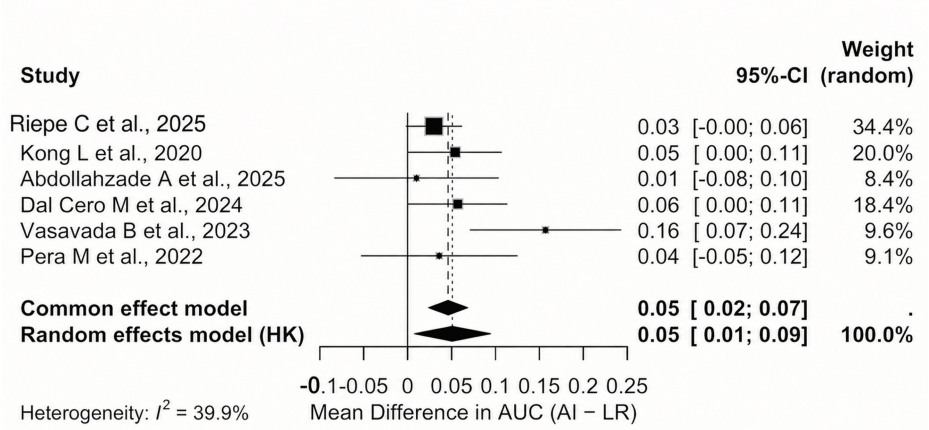


FIGURE 7: Forest plot of the difference in the area under the receiver operating characteristic curve (ΔAUC) between artificial intelligence-based models and logistic regression for 90-day postoperative mortality.

Squares represent study-level ΔAUC estimates and horizontal lines the corresponding 95% confidence intervals; the diamond represents the pooled random-effects estimate with the Hartung-Knapp (HK)-adjusted 95% confidence interval. Between-study variance (τ²) was estimated by restricted maximum likelihood. Values above zero favour artificial intelligence-based models.

Studies included: [19,22-25,27]

Overall Comparative Performance

Overall, AI-based models were associated with modest but statistically significant improvements in AUC for both short-term and longer-term mortality prediction. Although the magnitude of benefit was similar for 30-day and 90-day outcomes, heterogeneity was substantially higher in the 30-day analysis, suggesting that short-term mortality prediction may be more sensitive to differences in study design, population characteristics, or modeling strategy.

Discussion

In this systematic review and meta-analysis, ML models showed a statistically significant but modest improvement in discrimination compared with LR for postoperative mortality prediction, with pooled ΔAUC values of 0.0575 for 30-day mortality and 0.0509 for 90-day mortality. Previous work has emphasized that heterogeneity among surgical populations may influence the predictive performance of perioperative risk models [7]. Our findings extend this observation by showing a similarly modest incremental improvement with ML across two mortality horizons despite markedly different levels of between-study heterogeneity. The 30-day analysis showed considerable heterogeneity (I^2 = 89.0%), whereas heterogeneity was substantially lower for 90-day mortality (I^2 = 39.9%); nevertheless, the pooled ΔAUC was similar in magnitude. This pattern suggests that between-study heterogeneity alone is unlikely to explain the limited incremental discrimination observed with ML and highlights the importance of how prediction models are evaluated and reported.

Previous comparative studies and reviews have also shown that the relative performance of ML and conventional regression varies across populations and modeling settings [30-34]. ML-based surgical risk calculators have been developed for specific surgical settings, but performance in individual cohorts does not establish consistent superiority across settings [35,36].

The clinical significance of these improvements should be interpreted cautiously. Discrimination alone does not establish clinical usefulness, and an increase in AUC does not necessarily translate into improved decision-making or patient outcomes [32]. Calibration, which reflects agreement between predicted and observed risk, is also an important component of model evaluation [37]. Decision curve analysis assesses net clinical benefit across clinically relevant thresholds [38]. However, few included studies reported calibration metrics or decision curve analysis, precluding quantitative synthesis of these domains. Consequently, whether the observed improvements in discrimination translate into meaningful clinical benefit remains uncertain.

These findings suggest that the central limitation of the current literature may not be the modest difference

in discrimination between ML and LR itself, but the incomplete assessment of model performance beyond discrimination. Future studies should prioritize transparent reporting of calibration, decision curve analysis, and external validation to determine whether statistically improved discrimination translates into clinically meaningful benefit.

Limitations

This study has several limitations. First, the number of studies eligible for quantitative synthesis was limited, reflecting the scarcity of direct within-cohort comparisons between ML and LR for postoperative mortality prediction. This may have reduced the precision of the pooled estimates and limited the scope of sensitivity analyses. Despite restricting inclusion to direct within-cohort comparisons, substantial heterogeneity remained, particularly in the 30-day analysis. Differences in patient populations, surgical procedures, sample size, case mix, predictor selection, ML algorithms, feature engineering, and validation strategies are likely contributors to this variability. Accordingly, the elevated I^2 in the 30-day analysis should be interpreted as reflecting both clinical and methodological diversity across prediction-model studies. Because the covariance between paired AI and LR AUC estimates was unavailable, estimating the variance of ΔAUC required an assumed within-study correlation, which introduced additional uncertainty into study weights and CIs.

Second, most included studies primarily reported discrimination metrics, with limited and inconsistent reporting of calibration and decision curve analysis. Because AUC was the most consistently reported and comparable performance measure across studies, the meta-analysis necessarily focused on discrimination. The absence of sufficient calibration and clinical utility data limits conclusions about the real-world applicability of the observed performance differences.

Third, PROBAST assessment identified methodological concerns in several included studies, particularly in the analysis domain, including unclear reporting of model development, lack of external validation, and potential overfitting. These issues may have influenced individual model-performance estimates and contributed to between-study variability.

Additionally, the literature search was restricted to PubMed, Embase, and the Cochrane Library and did not include dedicated searches of grey-literature sources or trial registries. Although the reference lists of included studies and relevant reviews were manually screened and preprints were eligible for inclusion, relevant unpublished or non-indexed studies may have been missed, potentially introducing retrieval or publication bias. Studies published or indexed after the search date of February 2, 2026, are not represented. Publication bias and small-study effects could not be formally evaluated because few studies contributed to each meta-analysis; selective publication of studies reporting more favorable model performance therefore cannot be excluded. Finally, although risk of bias was formally assessed using PROBAST, the overall certainty of evidence was not evaluated using a structured framework such as Grading of Recommendations Assessment, Development, and Evaluation (GRADE) [39]. Confidence in the pooled estimates is therefore limited by methodological concerns in the included studies, the small number of studies available for each synthesis, and the observed heterogeneity.

Conclusions

In this systematic review and meta-analysis, ML models demonstrated a statistically significant but modest improvement in discriminative performance compared with LR for postoperative mortality prediction. Although these findings suggest a potential advantage of ML approaches, the magnitude of improvement is limited and may not translate into meaningful clinical benefit without adequate calibration and demonstrated clinical utility. The current evidence base is characterized by substantial methodological heterogeneity and incomplete reporting of key performance domains, limiting full assessment of the clinical impact and generalizability of these models. Future research should prioritize rigorous model development and validation, including routine assessment of calibration, decision curve analysis, and external validation across diverse patient populations. Adoption of ML in perioperative risk prediction should therefore be guided by improvements in discrimination together with evidence of clinical utility, calibration, and reproducibility across settings.

Appendices

Database	Search string
	(("artificial intelligence"[Title/Abstract] OR AI[Title/Abstract] OR "machine learning"[Title/Abstract] OR "deep learning"[Title/Abstract] OR "neural network*"[Title/Abstract] OR "artificial neural network*"[Title/Abstract] OR "convolutional neural network*"[Title/Abstract] OR CNN[Title/Abstract] OR RNN[Title/Abstract] OR LSTM[Title/Abstract] OR transformer*[Title/Abstract] OR "large language model*"[Title/Abstract] OR LLM[Title/Abstract] OR "foundation model*"[Title/Abstract] OR random forest*[Title/Abstract] OR "support vector machine*"[Title/Abstract] OR SVM[Title/Abstract] OR XGBoost[Title/Abstract] OR LightGBM[Title/Abstract] OR CatBoost[Title/Abstract] OR gradient boosting[Title/Abstract] OR AdaBoost[Title/Abstract] OR "decision tree*"[Title/Abstract] OR ensemble*[Title/Abstract] OR "Bayesian network*"

PubMed	[Title/Abstract] OR "fuzzy logic"[Title/Abstract] OR "expert system*"[Title/Abstract] OR "generative pre-trained transformer" [Title/Abstract] OR "logistic regression"[Title/Abstract]) AND ("mortality"[Title/Abstract] OR "death"[Title/Abstract] OR "postoperative mortality"[Title/Abstract] OR "risk prediction"[Title/Abstract] OR "predictive model*"[Title/Abstract] OR "prognostic model*"[Title/Abstract]) AND (("elective"[Title/Abstract] OR "planned"[Title/Abstract] OR "non-emergent" [Title/Abstract]) AND ("abdominal surgery"[Title/Abstract] OR "general surgery"[Title/Abstract] OR "visceral surgery" [Title/Abstract] OR "gastrointestinal surgery"[Title/Abstract] OR "colorectal surgery"[Title/Abstract] OR "coloproctology" [Title/Abstract] OR "bowel surgery"[Title/Abstract] OR "intestinal surgery"[Title/Abstract] OR anastomosis[Title/Abstract] OR "hepatic surgery"[Title/Abstract] OR "hepatobiliary surgery"[Title/Abstract] OR cholecystectomy[Title/Abstract] OR appendectomy[Title/Abstract] OR appendicectomy[Title/Abstract] OR gastrectomy[Title/Abstract] OR pancreatectomy[Title/Abstract] OR splenectomy[Title/Abstract] OR laparotomy[Title/Abstract] OR laparoscopy[Title/Abstract])))
Embase	('gastrointestinal surgery'/exp OR 'esophagus surgery'/exp OR 'liver surgery'/exp OR 'abdominal surgery'/de OR (((gastrointestin* OR 'digestive tract*' OR 'digestive system*' OR anus OR bariatric* OR intestin* OR pancrea* OR stomach* OR esophag* OR oesophag* OR gastric OR liver* OR hepatic OR colon OR rectal OR rectum OR colorect* OR abdominal OR abdomen) NEAR/3 (surger* OR resection* OR operation* OR 'surgical proced*' OR reconstruction*)):ab,ti,kw) OR esophagectom*:ab,ti,kw OR oesophagectom*:ab,ti,kw OR gastrectom*:ab,ti,kw OR gastroresection*:ab,ti,kw OR gastroduodenostom*:ab,ti,kw OR gastroenterostom*:ab,ti,kw OR gastrojejunostom*:ab,ti,kw OR pancreatectom*:ab,ti,kw OR colectom*:ab,ti,kw OR appendectom*:ab,ti,kw OR appendicectom*:ab,ti,kw OR hepatectom*:ab,ti,kw OR pancreaticoduodenectom*:ab,ti,kw) AND ('machine learning'/de OR 'supervised machine learning'/de OR 'support vector machine'/exp OR 'random forest'/de OR 'artificial intelligence'/exp OR 'natural language processing'/de OR 'artificial neural network'/exp OR 'statistical model'/de OR 'logistic regression analysis'/exp OR 'learning algorithm'/de OR 'bayesian learning'/de OR 'deep learning'/de OR 'computational intelligence'/de OR (((machin* OR bayesian OR deep OR hierarchical) NEAR/3 learning*):ab,ti,kw) OR (((artificial* OR machin* OR computational) NEAR/3 intelligen*):ab,ti,kw) OR ((support* NEAR/3 vector*):ab,ti,kw) OR ((natural NEAR/3 languag* NEAR/3 processing*):ab,ti,kw) OR ((neural NEAR/3 network*):ab,ti,kw)) AND ((((logistic OR 'proportional odds') NEAR/3 (model* OR regression*)):ab,ti,kw) OR ((random NEAR/3 forest*):ab,ti,kw) OR ((learning NEAR/3 algorithm*):ab,ti,kw) OR ((naive NEAR/3 bayes):ab,ti,kw) OR (((computer OR automated OR machine) NEAR/3 reasoning):ab,ti,kw) OR nlp:ab,ti,kw) AND ('controlled clinical trial'/exp OR random*:ti,ab OR placebo*:ti,ab OR trial:ti OR (('controlled clinical trial'/exp OR random*:ti,ab OR placebo*:ti,ab OR trial:ti) AND [embase]/lim)) AND ('morbidity'/de OR 'mortality'/de OR 'surgical mortality'/de OR 'perioperative mortality'/de OR 'clavien dindo classification'/de OR 'postoperative complication'/de OR 'survival'/de OR 'surgical outcome'/de OR 'mortality rate'/exp OR 'hospital mortality'/de OR morbidit*:ab,ti,kw OR mortalit*:ab,ti,kw OR 'clavien dindo*':ab,ti,kw OR survival:ab,ti,kw OR (((postoperativ* OR perioperativ* OR surgical OR 'post operat*' OR 'peri operativ*' OR postoperat* OR 'post surg*' OR postsurg*) NEAR/3 (complicat* OR outcome*)):ab,ti,kw) OR (((predict* OR model*) NEAR/3 outcome*):ab,ti,kw)) NOT (('animal'/exp OR animal*:de OR 'nonhuman'/de) NOT 'human'/exp) NOT (('juvenile'/exp OR juvenil*:ab,ti,kw OR adolescen*:ab,ti,kw OR preadolescen*:ab,ti,kw OR youth*:ab,ti,kw OR child*:ab,ti,kw OR schoolchild*:ab,ti,kw OR minors:ab,ti,kw OR teen:ab,ti,kw OR teens:ab,ti,kw OR teenager*:ab,ti,kw OR infan*:ab,ti,kw OR toddler*:ab,ti,kw OR pediatr*:ab,ti,kw OR paediatr*:ab,ti,kw OR puber*:ab,ti,kw OR baby:ab,ti,kw OR babies:ab,ti,kw OR girl*:ab,ti,kw OR boy*:ab,ti,kw OR newborn*:ab,ti,kw OR neonate*:ab,ti,kw OR premature*:ab,ti,kw OR 'pre matur*':ab,ti,kw OR kid:ab,ti,kw OR kids:ab,ti,kw OR underag*:ab,ti,kw OR kindergar*:ab,ti,kw OR pubescen*:ab,ti,kw OR prepubesc*:ab,ti,kw OR school*:ab,ti,kw OR preschool*:ab,ti,kw OR highschool*:ab,ti,kw OR suckling:ab,ti,kw OR picu:ab,ti,kw OR nicu:ab,ti,kw OR picus:ab,ti,kw OR nicus:ab,ti,kw) NOT ('adult'/exp OR adult*:ab,ti,kw OR elderl*:ab,ti,kw OR man:ab,ti,kw OR men:ab,ti,kw OR woman:ab,ti,kw OR women:ab,ti,kw)) AND [english]/lim NOT (([conference abstract]/lim OR [conference review]/lim) AND 1800-2019/py) NOT 'editorial'/de NOT ('case report'/de OR 'case report':ti)
Cochrane Library	(("artificial intelligence":ti,ab,kw OR "machine learning":ti,ab,kw OR "deep learning":ti,ab,kw OR "neural network*":ti,ab,kw OR "artificial neural network*":ti,ab,kw OR "convolutional neural network*":ti,ab,kw OR CNN:ti,ab,kw OR RNN:ti,ab,kw OR LSTM:ti,ab,kw OR transformer*:ti,ab,kw OR "large language model*":ti,ab,kw OR "foundation model*":ti,ab,kw OR random forest*:ti,ab,kw OR "support vector machine*":ti,ab,kw OR SVM:ti,ab,kw OR XGBoost:ti,ab,kw OR gradient boosting:ti,ab,kw OR AdaBoost:ti,ab,kw OR "decision tree*":ti,ab,kw OR ensemble*:ti,ab,kw OR "Bayesian network*":ti,ab,kw OR "fuzzy logic":ti,ab,kw OR "expert system*":ti,ab,kw OR "logistic regression":ti,ab,kw) AND (Mortality:ti,ab,kw OR death:ti,ab,kw OR "postoperative mortality":ti,ab,kw OR "risk prediction":ti,ab,kw OR "predictive model*":ti,ab,kw OR "prognostic model*":ti,ab,kw) AND ((elective:ti,ab,kw OR planned:ti,ab,kw OR nonemergent:ti,ab,kw) AND ("abdominal surgery":ti,ab,kw OR "general surgery":ti,ab,kw OR "visceral surgery":ti,ab,kw OR "gastrointestinal surgery":ti,ab,kw OR "colorectal surgery":ti,ab,kw OR "bowel surgery":ti,ab,kw OR "intestinal surgery":ti,ab,kw OR anastomosis:ti,ab,kw OR "hepatic surgery":ti,ab,kw OR "hepatobiliary surgery":ti,ab,kw OR cholecystectomy:ti,ab,kw OR appendectomy:ti,ab,kw OR gastrectomy:ti,ab,kw OR pancreatectomy:ti,ab,kw OR splenectomy:ti,ab,kw OR laparotomy:ti,ab,kw OR laparoscopy:ti,ab,kw)))

TABLE 4: Database search strategies

Full electronic search strategy for PubMed, Embase and the Cochrane Library.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Edson A. Escandón-Villalobos, Rebeca G. Elguezabal-Rodelo, Miguel A. Pardiño Vega, Eduardo E. Montalvo-Jave, Eduardo Villegas-Tovar , Regina Faes Petersen

Acquisition, analysis, or interpretation of data: Edson A. Escandón-Villalobos, Rebeca G. Elguezabal-Rodelo, Miguel A. Pardiño Vega, Lia Medina-Montalvo, Gyselle A. Godinez-Castillo

Drafting of the manuscript: Edson A. Escandón-Villalobos, Rebeca G. Elguezabal-Rodelo, Miguel A. Pardiño Vega, Lia Medina-Montalvo, Gyselle A. Godinez-Castillo

Critical review of the manuscript for important intellectual content: Edson A. Escandón-Villalobos, Rebeca G. Elguezabal-Rodelo, Miguel A. Pardiño Vega, Lia Medina-Montalvo, Eduardo E. Montalvo-Jave, Eduardo Villegas-Tovar , Regina Faes Petersen

Supervision: Edson A. Escandón-Villalobos, Rebeca G. Elguezabal-Rodelo, Eduardo E. Montalvo-Jave, Eduardo Villegas-Tovar , Regina Faes Petersen

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

ChatGPT (GPT-5; OpenAI, San Francisco, California, USA) was used to assist with journal-specific formatting and language editing. It was not used to generate or analyze scientific data, conduct the statistical analyses, assess risk of bias, or determine or modify the scientific interpretation. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content of the manuscript.

References

1. Ng-Kamstra JS, Nepogodiev D, Lawani I, Bhangu A, Workneh RS: Perioperative mortality as a meaningful indicator: challenges and solutions for measurement, interpretation, and health system improvement. Anaesth Crit Care Pain Med. 2020, 39:673-81. 10.1016/j.accpm.2019.11.005
2. Ma J, Dhiman P, Qi C, Bullock G, van Smeden M, Riley RD, Collins GS: Poor handling of continuous predictors in clinical prediction models using logistic regression: a systematic review. J Clin Epidemiol. 2023, 161:140-51. 10.1016/j.jclinepi.2023.07.017
3. Fahim YA, Hasani IW, Kabba S, Ragab WM: Artificial intelligence in healthcare and medicine: clinical applications, therapeutic advances, and future perspectives. Eur J Med Res. 2025, 30:848. 10.1186/s40001-025-03196-w
4. Haug CJ, Drazen JM: Artificial intelligence and machine learning in clinical, 2023 . N Engl J Med. 2023, 388:1201-8. 10.1056/NEJMra2302038
5. Chiew CJ, Liu N, Wong TH, Sim YE, Abdullah HR: Utilizing machine learning methods for preoperative prediction of postsurgical mortality and intensive care unit admission. Ann Surg. 2020, 272:1133-9. 10.1097/SLA.0000000000003297
6. Michelsen C, Jørgensen CC, Heltberg M, et al.: Machine-learning vs. logistic regression for preoperative prediction of medical morbidity after fast-track hip and knee arthroplasty-a comparative study. BMC Anesthesiol. 2023, 23:391. 10.1186/s12871-023-02354-z
7. Vernooij JE, Koning NJ, Geurts JW, Holewijn S, Preckel B, Kalkman CJ, Vernooij LM: Performance and usability of pre-operative prediction models for 30-day peri-operative mortality risk: a systematic review. Anaesthesia. 2023, 78:607-19. 10.1111/anae.15988
8. Page MJ, McKenzie JE, Bossuyt PM, et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Rev Esp Cardiol (Engl Ed). 2021, 74:790-9. 10.1016/j.rec.2021.07.010
9. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, Thomas J: Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. Cochrane Database Syst Rev. 2019, 10:ED000142. 10.1002/14651858.ED000142
10. World Medical Association: World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA. 2013, 310:2191-4. 10.1001/jama.2013.281053
11. Zotero: your personal research assistant. (2026). Accessed: August 27, 2026: https://www.zotero.org/.
12. Moons KG, de Groot JA, Bouwmeester W, et al.: Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014, 11:e1001744. 10.1371/journal.pmed.1001744
13. Wolff RF, Moons KG, Riley RD, et al.: PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019, 170:51-8. 10.7326/M18-1376
14. Collins GS, Dhiman P, Andaur Navarro CL, et al.: Protocol for development of a reporting guideline

(TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021, 11:e048008. 10.1136/bmjopen-2020-048008
15. McGuinness LA, Higgins JP: Risk-of-bias VISualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments. Res Synth Methods. 2021, 12:55-61. 10.1002/jrsm.1411
16. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982, 143:29-36. 10.1148/radiology.143.1.7063747
17. Hanley JA, McNeil BJ: A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983, 148:839-43. 10.1148/radiology.148.3.6878708
18. Balduzzi S, Rücker G, Schwarzer G: How to perform a meta-analysis with R: a practical tutorial . Evid Based Ment Health. 2019, 22:153-60. 10.1136/ebmental-2019-300117
19. Abdollahzade A, Rahimi H, Ahmadzade AM, et al.: Predictive models for post-liver transplant survival using machine learning techniques in three critical time intervals. J Liver Transpl. 2025, 17:100253. 10.1016/j.liver.2024.100253
20. van den Bosch T, Warps AK, de Nerée Tot Babberich MP, et al.: Predictors of 30-day mortality among Dutch patients undergoing colorectal cancer surgery, 2011-2016. JAMA Netw Open. 2021, 4:e217737. 10.1001/jamanetworkopen.2021.7737
21. Bunn C, Kulshrestha S, Boyda J, et al.: Application of machine learning to the prediction of postoperative sepsis after appendectomy. Surgery. 2021, 169:671-7. 10.1016/j.surg.2020.07.045
22. Riepe C, van de Water R, Winter A, et al.: 90-day mortality prediction in elective visceral surgery using machine learning: a retrospective multicenter development, validation, and comparison study. Int J Surg. 2025, 111:3742-51. 10.1097/JS9.0000000000002372
23. Dal Cero M, Gibert J, Grande L, et al.: International external validation of risk prediction model of 90-day mortality after gastrectomy for cancer using machine learning. Cancers (Basel). 2024, 16:2463. 10.3390/cancers16132463
24. Kong L, Lv T, Jiang L, Yang J, Yang J: A simple four-factor preoperative recipient scoring model for prediction of 90-day mortality after adult liver transplantation: a retrospective cohort study. Int J Surg. 2020, 81:26-31. 10.1016/j.ijsu.2020.07.021
25. Pera M, Gibert J, Gimeno M, et al.: Machine learning risk prediction model of 90-day mortality after gastrectomy for cancer. Ann Surg. 2022, 276:776-83. 10.1097/SLA.0000000000005616
26. Sahara K, Paredes AZ, Tsilimigras DI, et al.: Machine learning predicts unpredicted deaths with high accuracy following hepatopancreatic surgery. Hepatobiliary Surg Nutr. 2021, 10:20-30. 10.21037/hbsn.2019.11.30
27. Vasavada B: Use of supervised machine learning algorithms in predicting postoperative mortality in gastrointestinal and hepatopancreatobiliary surgeries [PREPRINT]. medRxiv. 2023, 10.1101/2023.07.05.23292033
28. Wise E, Leslie D, Amateau S, Hocking K, Scott A, Dutta N, Ikramuddin S: Prediction of thirty-day morbidity and mortality after duodenal switch using an artificial neural network. Surg Endosc. 2023, 37:1440-8. 10.1007/s00464-022-09378-5
29. Yan X, Goldsmith J, Mohan S, et al.: Impact of intraoperative data on risk prediction for mortality after intra-abdominal surgery. Anesth Analg. 2022, 134:102-13. 10.1213/ANE.0000000000005694
30. Weng SF, Vaz L, Qureshi N, Kai J: Prediction of premature all-cause mortality: a prospective general population cohort study comparing machine-learning and standard epidemiological approaches. PLoS One. 2019, 14:e0214365. 10.1371/journal.pone.0214365
31. Benedetto U, Dimagli A, Sinha S, et al.: Machine learning improves mortality risk prediction after cardiac surgery: systematic review and meta-analysis. J Thorac Cardiovasc Surg. 2022, 163:2075-2087.e9. 10.1016/j.jtcvs.2020.07.105
32. Mann J, Lyons M, O'Rourke J, Davies S: Machine learning or traditional statistical methods for predictive modelling in perioperative medicine: a narrative review. J Clin Anesth. 2025, 102:111782. 10.1016/j.jclinane.2025.111782
33. Jing B, Boscardin WJ, Deardorff WJ, Jeon SY, Lee AK, Donovan AL, Lee SJ: Comparing machine learning to regression methods for mortality prediction using Veterans Affairs electronic health record clinical data. Med Care. 2022, 60:470-9. 10.1097/MLR.0000000000001720
34. Hu Y, Zhang X, Slavin V, Belsti Y, Tiruneh SA, Callander E, Enticott J: Beyond comparing machine learning and logistic regression in clinical prediction modelling: shifting from model debate to data quality. J Med Internet Res. 2025, 27:e77721. 10.2196/77721
35. Bertsimas D, Dunn J, Velmahos GC, Kaafarani HM: 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, 268:574-83. 10.1097/SLA.0000000000002956
36. Liu R, Lai X, Wang J, Zhang X, Zhu X, Lai PB, Guo CR: A non-linear ensemble model-based surgical risk calculator for mixed data from multiple surgical fields. BMC Med Inform Decis Mak. 2021, 21:88. 10.1186/s12911-021-01450-9
37. Kantidakis G, Hazewinkel AD, Fiocco M: Neural networks for survival prediction in medicine using prognostic factors: a review and critical appraisal. Comput Math Methods Med. 2022, 2022:1176060. 10.1155/2022/1176060
38. Vickers AJ, Holland F: Decision curve analysis to evaluate the clinical benefit of prediction models . Spine J. 2021, 21:1643-8. 10.1016/j.spinee.2021.02.024
39. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ: GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008, 336:924-6. 10.1136/bmj.39489.470347.AD