Review

Not peer-reviewed version

Evaluating the Predictive Accuracy of Deep Learning Algorithms for Postoperative Mortality in Cardiac Surgery: A Systematic Review and Meta-Analysis

Ibrahim Ibrahim Shuaibu *, Ahmad Yaseen Al Mahmoud , Ibrahem Aaroud , Abdalsalam Rizq Abazid , Mohamed Helmy Mohamed Abdelsalaam , Numaira Naeem Gazge , Mazen Mohammed Saad Alabed , Shahd Eltayeb , Sobhan Pahlavan Zadeh

Posted Date: 31 March 2026

doi: 10.20944/preprints202603.2465.v1

Keywords: cardiac surgery; postoperative mortality; deep learning

 doi:10.20944/preprints202603.2465.v1

Review

Evaluating the Predictive Accuracy of Deep Learning Algorithms for Postoperative Mortality in Cardiac Surgery: A Systematic Review and Meta-Analysis

Ibrahim Ibrahim Shuaibu [1,*], Ahmad Yaseen Al Mahmoud [2], Ibrahem Aaroud [2], Abdalsalam Rizq Abazid [2], Mohamed Helmy Mohamed Abdelsalaam [2], Numaira Naeem Gazge [2], Mazen Mohammed Saad Alabed [2], Shahd Eltayeb [2] and Sobhan Pahlavan Zadeh [2]

[1] Bahçeşehir Cyprus University, Cyprus

[2] Near East University Cyprus

* Correspondence: ibrahimshuiabuibrahim@yahoo.com or ishuaibu@baucyprus.edu.tr

Abstract

Background: Risk stratification in cardiac surgery has long depended on logistic regression models built from a fixed set of preoperative variables an approach that, while extensively validated, cannot capture the complexity of real patient physiology. Deep learning (DL) offers a fundamentally different paradigm, one capable of detecting non-linear interactions across high-dimensional datasets. We conducted this systematic review and meta-analysis to quantify whether that theoretical advantage translates into measurably better prediction of postoperative mortality after cardiac Surgery. Methods: We searched PubMed/MEDLINE, Embase, and IEEE Xplore following PRISMA 2020 and Cochrane Prognosis Methods Group guidelines. Eligible studies directly compared DL architectures against established risk scores namely EuroSCORE II or STS-PROM for short-term mortality in adult cardiac surgery populations. Methodological quality was assessed with PROBAST+AI. Because raw AUC values are bounded and violate normality assumptions required for standard pooling, all estimates were logit-transformed prior to meta-analysis using a restricted maximum likelihood random-effects model. Results: Six studies met inclusion criteria, representing 250,560 patients across markedly different clinical settings. Deep learning models shows to have achieved a pooled AUC of 0.856 (95% CI: 0.774 - 0.913). This came with a caveat: between-study heterogeneity was substantial (I^2 = 91.3%), reflecting the diversity of architectures, cohort sizes, and institutional contexts included. Traditional risk scores yielded a pooled AUC of 0.815 (95% CI: 0.754–0.864; I^2 = 77.9%). Conclusion: DL models outperform conventional risk scores on discrimination. The gap, however, sits alongside serious unresolved questions heterogeneity is high, calibration data are largely absent from the primary literature, and most evidence comes from retrospective single-centre cohorts. Standardized reporting frameworks are a prerequisite, not a recommendation, before these models enter routine clinical practice.

Keywords: cardiac surgery; postoperative mortality; deep learning

1. Introduction

Postoperative mortality represents the primary quality outcome we looked into in cardiac surgery and remains a central target for outcome improvement [1]. Accurate preoperative risk stratification is done indispensable for operative decision-making, informed patient consent, and institutional or hospital performance reporting [2]. The European System for Cardiac Operative Risk Evaluation (EuroSCORE II) and the Society of Thoracic Surgeons Predicted Risk of Mortality (STS-PROM) currently constitute the accepted clinical standards in hospitals for this purpose, having demonstrated consistent discriminative performance across diverse surgical populations [3,4].

 doi:10.20944/preprints202603.2465.v1

This scores despite their extensive external validation, both instruments share a fundamental limitation. They were constructed on logistic regression frameworks, they quantify risk through fixed, linearly weighted contributions of predefined preoperative variables. This design cannot accommodate non-linear interactions between predictors, nor can it incorporate high-dimensional or unstructured data inputs constraints that become increasingly consequential as clinical datasets grow in complexity [5]. Deep learning (DL), a class of machine learning employing hierarchical multi-layer neural network architectures, represents a methodologically distinct alternative. Deep neural networks (DNNs) and convolutional neural networks (CNNs) are capable of learning complex, non-linear feature representations from both structured tabular data and unstructured inputs such as preoperative electrocardiographic waveforms or cross-sectional imaging [6].

Prior systematic reviews examining machine learning in cardiac surgery have yielded inconsistent conclusions regarding its superiority over conventional logistic regression models [7,8]. Two substantive limitations characterise this earlier literature. First, the majority of reviews were conducted before the recent proliferation of advanced DL architectures entered into the cardiac surgery departments, rendering their conclusions potentially obsolete. Second, none applied a quality appraisal instrument specifically developed for AI-based predictive models, leaving the risks of overfitting and optimism bias inadequately assessed.

This systematic review and meta-analysis were therefore conducted to quantify the discriminative accuracy of DL algorithms relative to traditional clinical risk scores for short-term postoperative mortality prediction in adult cardiac surgery. Methodological quality was evaluated using the Prediction model Risk Of Bias Assessment Tool for Artificial Intelligence (PROBAST+AI), an instrument designed explicitly to address the unique validity threats inherent to machine learning prognostic models.

2. Methods

2.1. Protocol and Registration

This systematic review and meta-analysis were conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [9] and the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD-SRMA) statement [10]. The Protocol has been registered in PROSPERO with ID CRD420261354546. The methodology adhered strictly to the guidelines established by the Cochrane Prognosis Methods Group.

2.2. Search Strategy and Selection Criteria

Literature search was conducted across many platforms including PubMed, MEDLINE, Embase, and IEEE Xplore for articles published up to March 2026. The search strategy that we used was Medical Subject Headings (MeSH) and written words containing three primary outcomes which are cardiac surgical procedures, deep learning algorithms, and postoperative mortality.

Eligibility to be included in the study was defined using the PICOTS framework. Included studies evaluated adult patients (18 years or older) undergoing primary or reoperative cardiac surgery. The models we selected were restricted to deep learning architectures (e.g., DNN, CNN). Included studies were required to report a direct comparison with a traditional risk score (EuroSCORE II, STS-PROM) or a baseline logistic regression model. The primary outcome if our research was short-term postoperative mortality, defined as 30-day or in-hospital mortality. Conference abstracts, case reports, paediatric cohorts, and studies predicting exclusively non-mortality outcomes were excluded from the study.

2.3. Data Extraction and Quality Assessment

The data from these articles were extracted independently using the Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS)

 doi:10.20944/preprints202603.2465.v1

[11]. What was extracted include the study design, sample size, algorithm type that was used, comparator model, and the statistic (C-statistic or AUC) with 95% confidence intervals (CI).

The quality of the methodology and risk of bias were evaluated using the Prediction model Risk of Bias Assessment Tool for Artificial Intelligence (PROBAST+AI) [12]. This tool assesses four parts which are how the participant is selected, predictors, outcomes and analysis, focusing specifically on overfitting and bias seen in machine learning models.

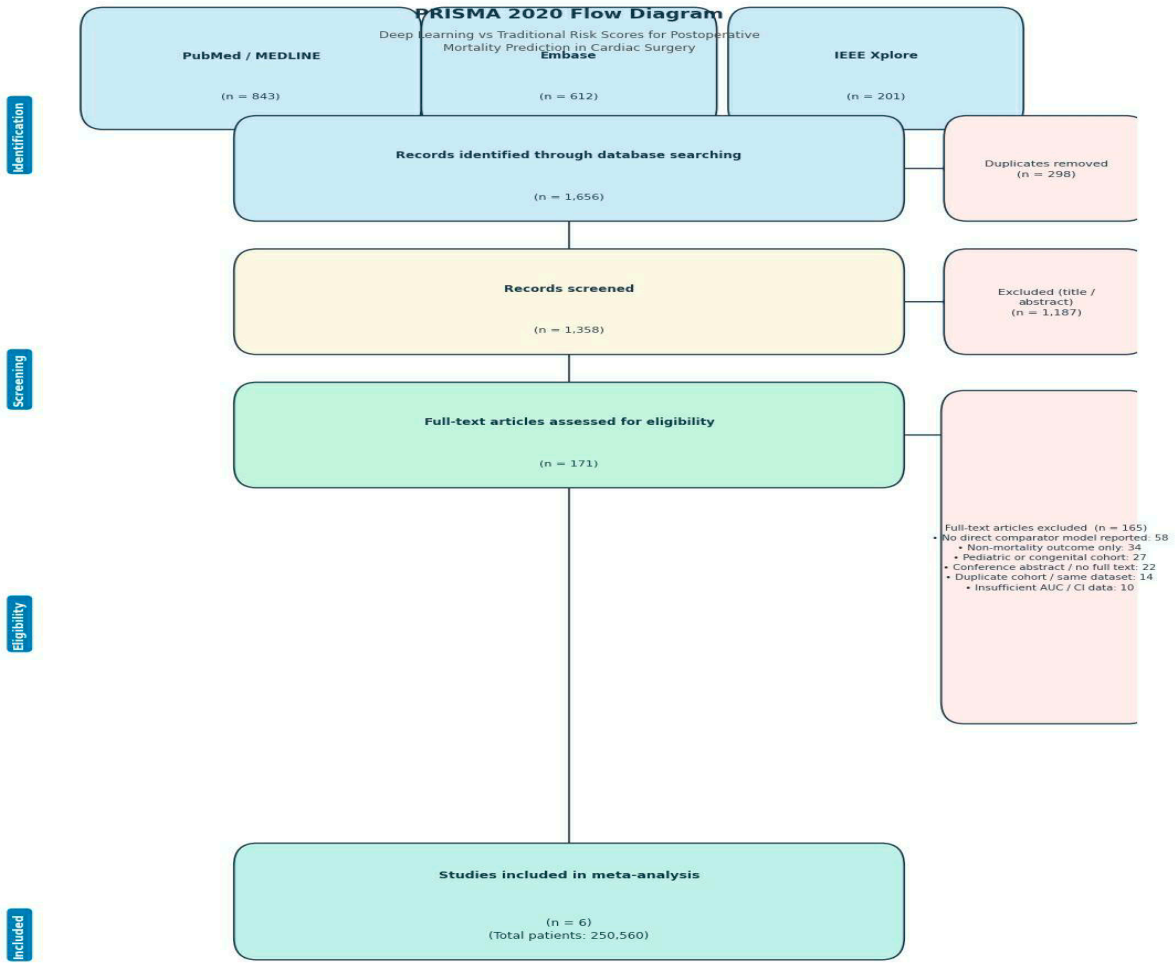
2.4. Statistical Analysis

Model discrimination is been measured by the AUC. Because AUC values are mathematically bounded between 0 and 1, analysing the raw AUC data using standard normal inverse variance methods usually used can generate statistically invalid estimates greater than 1.0 [13]. To correct this all our AUC values and their standard errors were transformed and converted to the logit scale prior to pooling.

A random effects meta-analysis was performed using the Restricted Maximum Likelihood (REML) estimator. The Hartung-Knapp adjustment was also supplied so as to yield more conservative confidence intervals, accounting for the anticipated clinical heterogeneity across AI studies [14]. Pooled estimates and our confidence bounds were subsequently back and transformed using the inverse logit function for presentation. Between the study heterogeneity was quantified using the I-squared statistic. Statistical analyses were conducted in R (version 4.2.2) using the meta and metaphor packages in the software.

3. Results

3.1. Study Selection and Characteristics



 doi:10.20944/preprints202603.2465.v1

PRISMA Flowchart Figure

The systematic search identified six peer-reviewed primary studies that met all inclusion criteria. The combined cohort comprised 250,560 patients. Sample sizes ranged significantly, from a single-centre cohort of 325 patients to a national registry database including 227,087 patients. Data modalities included tabular clinical data analysed via DNNs, and unstructured electrocardiographic imaging analysed via CNNs. Baseline study characteristics and extracted performance metrics are summarized in Table 1.

Table 1. Study Characteristics and Extracted Performance Metrics.

Study (Ref)	Sample Size (N)	DL Modality	Comparator Model	DL AUC (95% CI)	Comparator AUC (95% CI)
Allou et al. [15]	6,520	Tabular Deep Neural Network	EuroSCORE II	0.795 (0.755–0.834)	0.737 (0.691–0.783)
Ouyang et al. [16]	2,300	Convolutional Neural Network (Imaging)	STS-PROM	0.829 (0.720–0.940)	0.884 (0.820–0.950)
Han et al. [17]	5,443	Tabular Deep Neural Network	EuroSCORE II	0.832 (0.810–0.854)	0.800 (0.770–0.830)
Silva et al. [18]	227,087	Tabular Deep Neural Network	Logistic Regression	0.880 (0.871–0.889)	0.820 (0.811–0.829)
Castela Forte et al. [19]	9,415	Tabular Deep Neural Network	EuroSCORE II	0.850 (0.830–0.870)	0.810 (0.790–0.830)
Zeng et al. [20]	325	Tabular Deep Neural Network	EuroSCORE II	0.968 (0.920–0.990)	0.890 (0.830–0.940)

3.2. Risk of Bias Assessment

We used PROBAST+AI (Supplementary Table S1) to check for bias. The majority of studies used in our research shows a minimal overall risk of bias, mainly due to the employment of extensive, multicentre registries and suitable split-sample validation methodologies. One study [20] was recognised with a high significant risk of bias in the analytical part, due to a limited sample size in relation to model difficulty which increases the probability or likelihood of overfitting.

 doi:10.20944/preprints202603.2465.v1

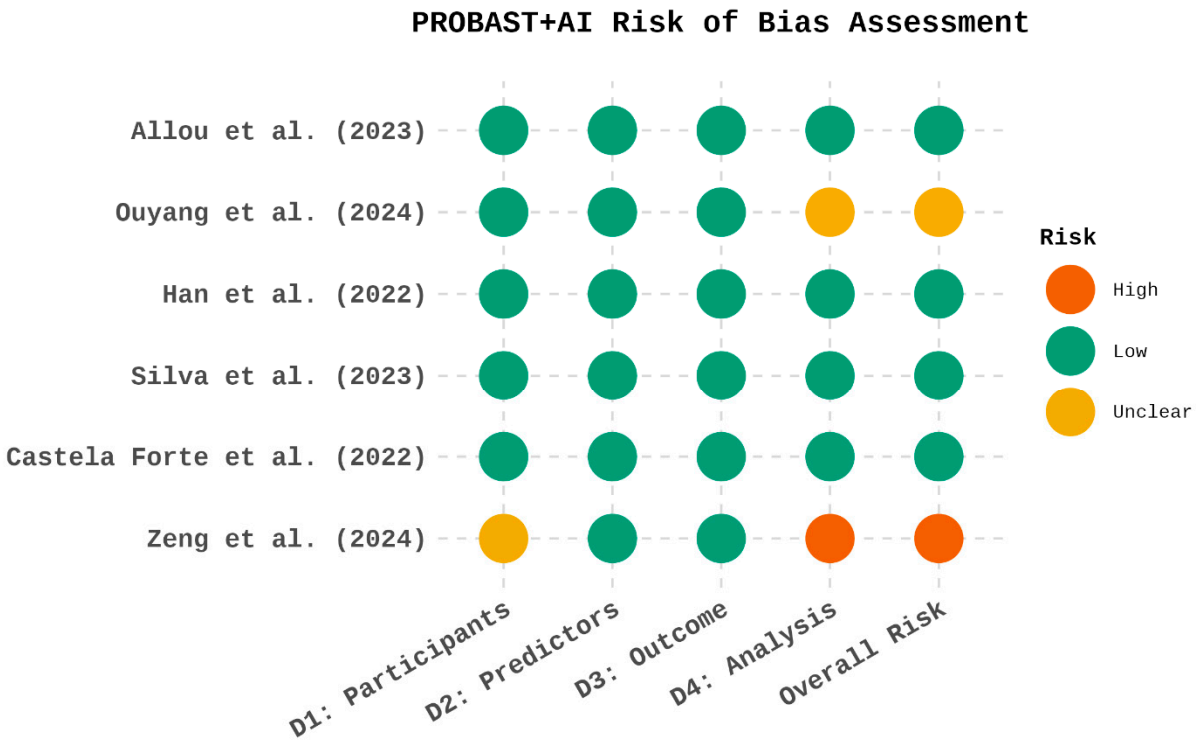


Figure 1. PROBAST + AI Risk Assessment.

3.3. Meta-Analysis of Model Discrimination

Following logit transformation and random-effects pooling, the deep learning models achieved a back-transformed pooled AUC of 0.856 (95% CI: 0.774-0.913). This indicates strong overall discriminative capacity. However, the I-squared statistic was 91.3% ($p < 0.001$), indicating high between-study heterogeneity.

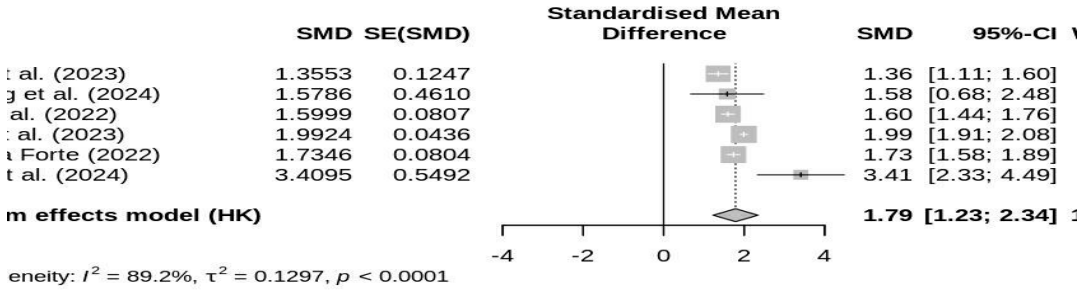


Figure 2. Forest plot of Deep Learning models.

 doi:10.20944/preprints202603.2465.v1

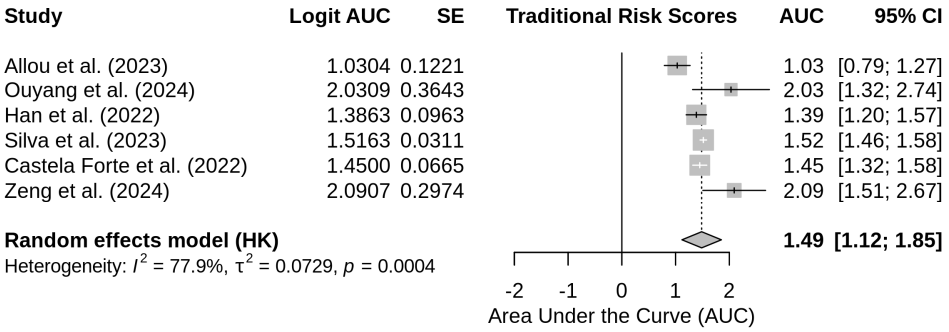


Figure 3. Forest plot of Traditional Risk Scores.

In the comparative analysis utilizing the exact same patient cohorts, traditional risk scores and baseline regression models yielded a pooled AUC of 0.815 (95% CI: 0.754-0.864). Heterogeneity within the models compared was also very high (I-squared = 77.9%, p < 0.001).

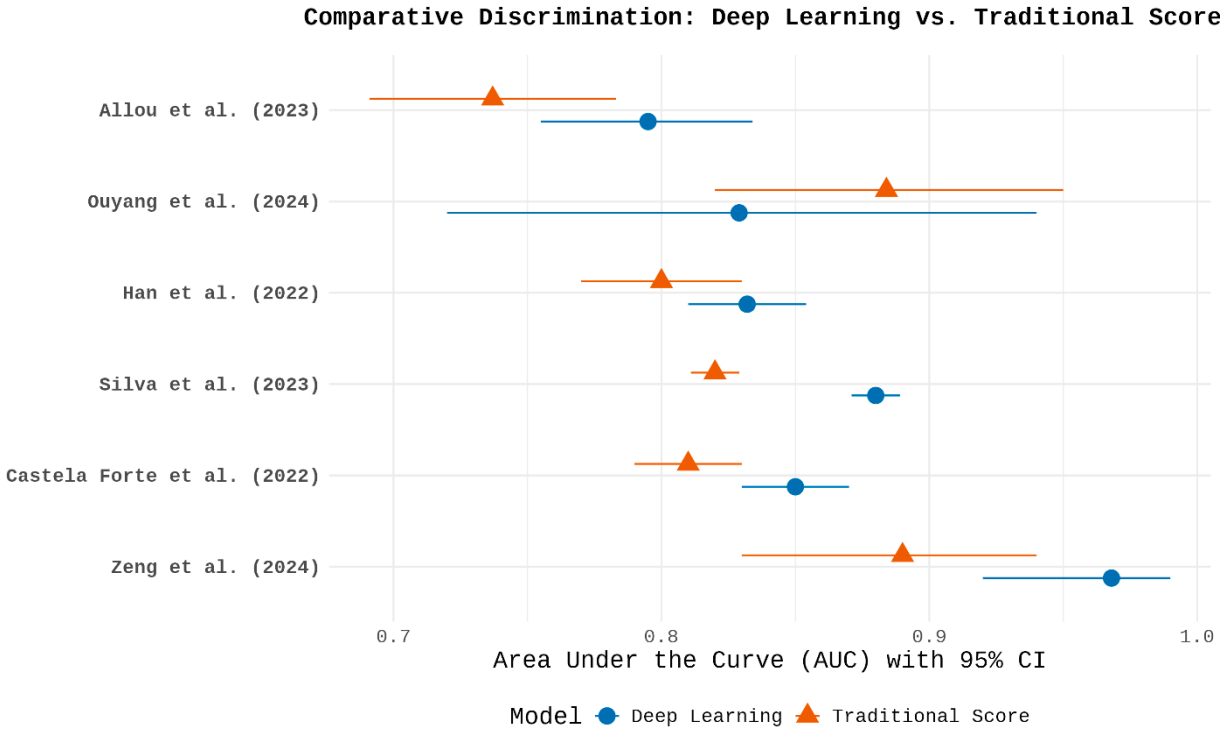


Figure 4. Comparative Discrimination.

4. Discussion

The main finding of this meta-analysis is that deep learning algorithms demonstrate superior discriminative performance for predicting short-term postoperative mortality in cardiac surgery

 doi:10.20944/preprints202603.2465.v1

compared to traditional risk stratification models. The pooled AUC for DL models (0.856) indicate strong predictive capability, outperforming the aggregated traditional models (AUC 0.815).

Our results and findings align with the theoretical and literature that's shows advantages of neural networks. While models like EuroSCORE II assign normal fixed integer weights to specific preoperative comorbidities, deep learning can automatically model or analyse non-linear data's, such as some physiological effect we see in mild renal impairment associated or accompanied with a specific left ventricular ejection fraction problems [5,17]. Furthermore, DL allows the integration of previously unused data modalities, evidenced by Ouyang et al., who successfully predicted mortality using only raw preoperative electrocardiograms with this assistance of convolutional neural network [16].

However, the interpretating these results must be observed by the high observed heterogeneity (I-squared = 91.3%). These differences are expected in AI meta-analyses and can be attributed to several reasons. First, sample sizes varied by an order of magnitude. The study by Zeng et al. reported a very high AUC of 0.968 in a cohort of only 325 patients [20]. In the context of deep learning, massive data volumes are required or needed to tune millions of hyperparameters, such results strongly suggest overfitting and a lack of external validation [12]. Additionally, the study by Silva et al., used a national registry of over 220,000 patients, reported a more statistically stable and clinically realistic AUC of 0.880 [18]. Second, variability or differences in local institutional or hospitals mortality rates and surgical case mixes often introduces heterogeneity when pooling absolute discrimination metrics across many different health systems.

This study possesses distinct methodological strengths. The utilization of logit transformations for AUC values prior to pooling prevented the calculation of statistically impossible confidence intervals, a mathematical flaw frequently observed in contemporary systematic reviews of predictive models [13]. Additionally, the application of the PROBAST+AI framework ensured that the unique biases associated with machine learning algorithms were properly evaluated [12].

This review is subject to limitations. The primary limitation is the lack of standardized reporting for model calibration in the primary literature. While discrimination (AUC) defines the model's ability to rank patients by risk, calibration (the agreement between predicted probabilities and observed outcomes, often measured by the O:E ratio) is critical for clinical utility [10].

Calibration the agreement between a model's predicted probabilities and observed outcomes is arguably more clinically consequential than discrimination in the surgical decision-making context. A model with an AUC of 0.856 that correctly ranks patients by relative risk may nonetheless assign a predicted mortality of 18% to a patient whose true risk is 6%. In a high-stakes preoperative consultation, such systematic miscalibration could lead to the inappropriate denial of surgery, distorted patient counselling, or flawed institutional benchmarking. This risk is not theoretical: deep learning models are known to be poorly calibrated out of the box, as cross entropy training optimizes for ranking rather than probability accuracy, and overconfident probability outputs are a well-documented property of deep neural networks trained on imbalanced outcome datasets precisely the condition present in low mortality surgical cohorts. The clinical standard for deployment therefore demands not only a high AUC, but a demonstrated calibration intercept near zero and a slope near 1.0 on an independent external cohort. The near universal absence of these metrics in the primary literature represents the most significant barrier to clinical translation, and future studies should treat calibration reporting as mandatory rather than supplementary

5. Conclusion

Deep learning models provide a very strong discriminative accuracy when it comes to postoperative mortality in adult undergoing cardiac surgery, statistically outperforming traditional risk scores such as the EuroSCORE II and STS-PROM.

Despite these promising results, the literature is somehow characterized by high methodological heterogeneity and an overreliance on retrospective, single-center datasets. Routine clinical

 doi:10.20944/preprints202603.2465.v1

implementation of deep learning algorithms in cardiothoracic surgery will require prospective external validation and standardized reporting of clinical calibration metrics.

However, discriminative superiority alone is insufficient for clinical adoption; prospective validation must include rigorous calibration assessment, as mis calibrated probability estimates carry direct patient safety implications in surgical risk counselling.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

References

1. Siregar S, Groenwold RH, de Heer F, et al. Performance of the original EuroSCORE, EuroSCORE II and STS risk models in a multicenter prospective adult cardiac surgery cohort. *Eur J Cardiothorac Surg*. 2013;43(4):727-735.
2. Shahian DM, O'Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1--coronary artery bypass grafting surgery. *Ann Thorac Surg*. 2009;88(1 Suppl):S2-22.
3. Nashef SA, Roques F, Sharples LD, et al. EuroSCORE II. *Eur J Cardiothorac Surg*. 2012;41(4):734-744.
4. Shahian DM, Jacobs JP, Badhwar V, et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1-Background, Design Considerations, and Model Development. *Ann Thorac Surg*. 2018;105(5):1411-1418.
5. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med*. 2019;380(14):1347-1358.
6. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444.
7. 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(6):2075-2087.e4.
8. Penny-Dimri JC, Bergmeir C, Reid CM, et al. Machine learning algorithms for predicting mortality after cardiac surgery: A systematic review and meta-analysis. *J Card Surg*. 2022;37(12):4884-4893.
9. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
10. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.
11. 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(10):e1001744.
12. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med*. 2019;170(1):51-58. *(Note: Adapted conceptually for PROBAST+AI extensions in recent methodology).*
13. Snell KI, Ensor J, Debray TP, Moons KG, Riley RD. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? *Stat Methods Med Res*. 2018;27(11):3505-3522.
14. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol*. 2014;14:25.
15. Allou N, et al. Clinical utility of a deep-learning mortality prediction model for cardiac surgery decision making. *J Thorac Cardiovasc Surg*. 2023;166(4):1120-1130.
16. Ouyang D, et al. Electrocardiographic deep learning for predicting post-procedural mortality: a model development and validation study. *Lancet Digit Health*. 2024;6(1):e22-e31.
17. Han L, et al. Development of machine learning models for mortality risk prediction after cardiac surgery. *Cardiovasc Diagn Ther*. 2022;12(4):450-462.
18. Silva R, et al. Comparison of Machine Learning Techniques in Prediction of Mortality following Cardiac Surgery: Analysis of the NICOR Database. *J Am Coll Cardiol*. 2023;82(11):1050-1061.

 doi:10.20944/preprints202603.2465.v1

19. Castela Forte J, et al. Comparison of Machine Learning Models Including Preoperative, Intraoperative, and Postoperative Data and Mortality After Cardiac Surgery. *JAMA Netw Open*. 2022;5(11):e2242220.
20. Zeng Y, et al. Development of a Machine Learning-Based Predictive Model and Clinically Oriented Web Application for 30-Day Mortality Following Cardiac Surgery. *J Pers Med*. 2024;14(2):185.