

Machine learning improves mortality risk prediction after cardiac surgery: Systematic review and meta-analysis

Umberto Benedetto, MD, PhD,^a Arnaldo Dimagli, MD,^a Shubhra Sinha, MD,^a Lucia Cocomello, MD,^a Ben Gibbison, MD,^a Massimo Caputo, MD,^a Tom Gaunt, PhD,^b Matt Lyon, MSc,^b Chris Holmes, PhD,^c and Gianni D. Angelini, MD^a

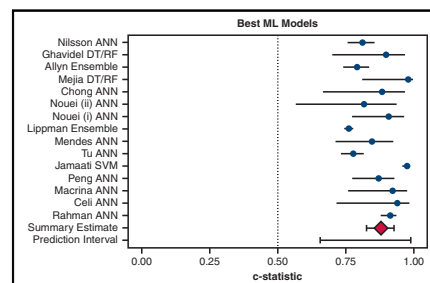
ABSTRACT

Background: Interest in the usefulness of machine learning (ML) methods for outcomes prediction has continued to increase in recent years. However, the advantage of advanced ML model over traditional logistic regression (LR) remains controversial. We performed a systematic review and meta-analysis of studies comparing the discrimination accuracy between ML models versus LR in predicting operative mortality following cardiac surgery.

Methods: The present systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis statement. Discrimination ability was assessed using the C-statistic. Pooled C-statistics and its 95% credibility interval for ML models and LR were obtained using a Bayesian framework. Pooled estimates for ML models and LR were compared to inform on difference between the 2 approaches.

Results: We identified 459 published citations of which 15 studies met inclusion criteria and were used for the quantitative and qualitative analysis. When the best ML model from individual study was used, meta-analytic estimates showed that ML were associated with a significantly higher C-statistic (ML, 0.88; 95% credibility interval, 0.83-0.93 vs LR, 0.81; 95% credibility interval, 0.77-0.85; $P = .03$). When individual ML algorithms were instead selected, we found a nonsignificant trend toward better prediction with each of ML algorithms. We found no evidence of publication bias ($P = .70$).

Conclusions: The present findings suggest that when compared with LR, ML models provide better discrimination in mortality prediction after cardiac surgery. However, the magnitude and clinical influence of such an improvement remains uncertain. (J Thorac Cardiovasc Surg 2020; ■:1-13)



Pooled C-statistic for mortality prediction by the best machine learning models.

CENTRAL MESSAGE

When compared to logistic regression models, machine learning appears able to provide better discrimination power in mortality prediction after cardiac surgery.

PERSPECTIVE

Mortality risk prediction is of crucial importance, especially when the benefit of surgery is difficult to assess and when individualized decision making is complex. Interest in the usefulness of new approaches based on machine learning has bloomed in recent years. We found that prediction models based on machine learning were associated with significantly better prediction accuracy.

See Commentary on page XXX.

From the Department of ^aTranslational Health Sciences, Bristol Heart Institute, and ^bPopulation Health Sciences, University of Bristol, London, United Kingdom; and ^cDepartment of Statistics, University of Oxford, Oxford, United Kingdom. Supported by the UK National Institute for Health Research Bristol Biomedical Research Centre and the British Heart Foundation.

Drs Benedetto and Dimagli contributed equally to this article.

Received for publication Jan 23, 2020; revisions received July 16, 2020; accepted for publication July 30, 2020.

Address for reprints: Umberto Benedetto, MD, PhD, Translational Health Sciences, Bristol Heart Institute, University of Bristol, Office Room 84, Level 7, Bristol Royal Infirmary, Upper Maudlin St, London, BS2 8HW United Kingdom (E-mail: umberto.benedetto@bristol.ac.uk).

0022-5223/\$36.00

Copyright © 2020 by The American Association for Thoracic Surgery

<https://doi.org/10.1016/j.jtcvs.2020.07.105>

Cardiac surgery is at high risk of intraoperative and postoperative complications. The benefit of surgery is sometimes difficult to predict and the decision to proceed on an individual basis is complex and therefore mortality risk evaluation has been increasingly emphasized in cardiac surgery. The aims of developing risk models include quality monitoring



Scanning this QR code will take you to the article title page to access supplementary information.

Abbreviations and Acronyms

AUC	= area under the receiver operating characteristic curve
CABG	= coronary artery bypass graft
LR	= logistic regression
MHR	= medical health record
ML	= machine learning

of surgical performance, counseling patients to aid with decision making and cost-benefit analysis. Several risk stratifications models have been developed to support clinical decision making such as the European System for Cardiac Operative Risk Evaluation (EuroSCORE)^{1,2} and the North American Society of Thoracic Surgeons.³ However, some of these scores, such as the EuroSCORE have shown major limitations as they tend to overestimate the actual risk.^{4,5} This can potentially translate into inappropriate risk adverse practice that denies surgery to patients who would benefit from surgery, falsely reassuring conclusions about surgeon and center performance, patients and their doctors not being fully informed during the process of shared decision making.

All models in current use are based on logistic regression (LR), which relies on the modeler input to manually specify interactions, such as complex interactions. Missing those relationships during the development of the scores may result in model misspecification. In this context, machine learning (ML) approaches automatically learn the relationships from the data and do not require input from the modeler to specify interactions.⁶ Interest on the usefulness of these methods has continued to increase in the recent years although ML has not been widely adopted in clinical practice yet. Moreover, recent reports, including a variety of clinical conditions have challenged the additional value of ML in the development of clinical prediction models.⁶ The objective of this systematic review and meta-analysis was to compare the accuracy of prediction methods using ML with conventional models based on LR in predicting operative mortality after cardiac surgery.

METHODS

The study was registered with PROSPERO (CRD42019155549). We followed the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) statement.

Search Strategy

We screened citations obtained from MEDLINE (1966 to October 2019), OVID Healthstar (1975 to October 2019), EMBASE (1980 to October 2019), The Cochrane Library (all databases) (October 2019) and SciSearch (1980 to October 2019). The search strategy is presented in [Appendix E1](#).

The reviewers screened reference lists of included studies. The search is updated to October 17, 2019.

Selection of Studies

All abstracts were independently screened by 2 reviewers (AD and SS); conflicts were resolved by a third reviewer (UB). The full text of selected abstracts was independently assessed for eligibility by 3 reviewers (AD, LC, UB), and conflicts were resolved by consensus.

Inclusion and Exclusion Criteria

Studies were eligible in the case that the article described the development of a prognostic prediction model for individualized prediction of operative mortality (in-hospital or within 30 days from surgery) in patients undergoing cardiac surgery, the article compared prediction models based on ML versus LR model. Studies were excluded if a new modelling approach was introduced (ie, dynamic modeling), no validation was carried out, the models made predictions for individual images or signals rather than participants, models were developed based on high-dimensional data modalities, the primary interest was assessing risk factors rather than prediction modeling, they were reviews of the literature, and full text was not available. In the case of studies with overlapping population, we predetermined that the study with the largest sample was to be included.

Data Extraction and Risk of Bias

Two reviewers (AD and UB) independently abstracted qualitative and quantitative data from selected studies. The list of extraction items was based on the CHARMS checklist⁷ and the QUADAS⁸ risk of bias tool. The extracted items included general study characteristics, applied algorithms and their characteristics, data-driven variable selection, and model performance.

Model performance was primarily assessed in terms of discrimination ability for operative mortality. Discrimination refers to a prediction model's ability to distinguish between subjects developing and not developing the outcome and is quantified by the concordance (C)-statistic, which corresponds to the area under the receiver operating characteristics curve (AUC).⁹ The C-statistic is an estimated conditional probability that for any pair of a subject who experienced and a subject who did not experience the outcome, the predicted risk of an event is higher for the former. C-statistics were from external validation (ie, validation sample was not used for model training) or from internal validation analysis (ie, k-fold cross-validation or bootstrapping). The standard error (SE) of C-statistic was calculated using the following formula:¹⁰

$$SE = \sqrt{\frac{c(1-c) + (N_1 - 1)(Q_1 - c^2) + (N_2 - 1)(Q_2 - c^2)}{N_1 N_2}}$$

where:

$$Q1 = \frac{c}{2-c} \quad Q2 = \frac{2c^2}{1+c}$$

Based on the extracted data, we classified ML algorithms into 5 broad groups¹¹: classification trees/random forests, artificial neural networks, support vector machines, Naïve Bayes, and other algorithms. We also collected the c-statistic for LR models and traditional risk scores (ie, EuroSCORE).

As proposed by Christodoulou and colleagues,⁶ from each article, we defined 5 signalling items to indicate potential bias ([Table E1](#)): unclear or biased validation of model performance, difference in whether data-driven variable selection was performed (yes/no) before applying LR and ML algorithms, difference in handling of continuous variables before applying LR and ML algorithms, different predictors considered for LR and ML algorithms, and whether corrections for imbalanced outcomes were used only for LR or only for ML algorithms. Each bias item was scored as no (not present), unclear, or yes (present). We considered a comparison at low risk of bias if the answer was no for all 5 signalling items. If the answer was unclear or yes for at least 1 item, we assumed high risk of bias.

TABLE 1. Study characteristics

Author, y	Geographic area	Population	Age (y)	Male sex (%)	Source of data	Sample size	Operative mortality (%)
Nilsson, 2006 ¹⁴	Europe	Unselected	62.6 ± 10.7	72	Retrospective (EuroSCORE database and MHR)	18,362	4.9
Ghavidel, 2014 ¹⁵	Asia	CABG and valve surgery	45-60 (58.1%) 61-75 (39.5%) >76 (2.4%)	86	Retrospective (MHR)	948	3.8
Allyn, 2017 ¹⁶	Europe	Unselected	63.4 ± 14.4	68	Retrospective (MHR)	6520	6.3
Mejia, 2018 ²³	South America	Rheumatic valve disease	51.2 ± 14.9 for survivors 54.4 ± 17.4 for nonsurvivors	NR	Prospective	2919	3.5
Chong, 2003 ²⁴	Asia	CABG	64.0 ± 10.4 for training set 63.3 ± 9.7 for testing set	70	Retrospective (MHR)	563	7.5
Nouei, 2016 ²⁵	Asia	CABG	58.24 ± 9.74 for survivors 62.07 ± 9.47 for nonsurvivors	70	Retrospective (MHR)	824	3.5
Nouei, 2014 ²⁶	Asia	CABG	58.62 ± 10.18 for survivors 61.82 ± 10.72 for nonsurvivors	NR	Retrospective (MHR)	1811	3.3
Lippman, 1997 ²⁷	North America	CABG	NR	73.4% survivors 62.3% nonsurvivors	Retrospective (STS database)	80,606	3.4
Mendes, 2015 ²⁸	South America	CABG	60.4 ± 9.6 in the training set 61.1 ± 9.8 in the testing set	68	Prospective	1315	8.6
Jamaati, 2015 ¹⁷	Asia	CABG	57	51	Prospective	2220	12.2
Tu, 1998 ¹⁸	North America	CABG	NR	NR	Retrospective (Cardiac Care Network of Ontario)	15,608	3.0
Rahman, 2012 ¹⁹	Asia	Unselected	18-40 (9.4%) 40-60 (53.2%) >60 (37.3%)	77	Retrospective (MHR)	1209	17.3
Celi, 2012 ²⁰	Oceania	Unselected	>80 (100%)	NR	Retrospective (Registry of Cardiac Surgery Patients in Dunedin Hospital)	165	7.4
Macrina, 2009 ²¹	Europe	Acute aortic dissection	61 ± 12 for survivors 66 ± 10 for nonsurvivors	63% for survivors 66% for nonsurvivors	Retrospective (MHR)	208	25.5
Peng, 2008 ²²	Asia	Unselected	63.2 ± 13.6 in the training set 64.8 ± 13.8 in the testing set	76% in the training set 70% in the testing set	Retrospective (MHR)	952	10.7

EuroSCORE, European System for Cardiac Operative Risk Evaluation; MHR, medical health record; CABG, coronary artery bypass graft; NR, not reported; STS, Society of Thoracic Surgeons.

TABLE 2. Study methodological characteristics

Author, y (reference no.)	Model tested	No. Predictors	Handling of missing data	Type of validation	Split ratio of training to testing sample	Calibration	Statistical software for ML
Nilsson, 2006 ¹⁴	ANN LR EuroSCORE	34	Missing excluded for mandatory variable Imputation for other variables (statistical mode or mean substitution)	Sample splitting and k-fold cross-validation plus external validation	75:25	Unclear	MatLab 7, Neural Network Toolbox, Stata
Ghavidel, 2014 ¹⁵	EEF-DT EEC-DT LR EuroSCORE	19	Missing excluded for the analysis	Sample splitting and k-fold cross-validation	70:30	NR	MATLAB and SPSS
Allyn, 2017 ¹⁶	GBM RF NB SVM Ensemble LR EuroSCORE EuroSCORE II	17	NR	Sample splitting and k-fold cross-validation	70:30	NR	SAS macro and R packages XGBoost, ExtraTrees, and e1071
Mejia, 2018 ²³	RF ANN SVM NB LR EuroSCORE II	10	Missing negligible, imputation not performed	K-fold cross-validation	–	NR	R package caret
Chong, 2003 ²⁴	ANN LR	18	Coded as missing for categorical variables and mean substitution for continuous variables	Sample splitting and k-fold cross-validation	75:25	NR	STATISTICA Neural Networks from StatSoft Inc
Nouei, 2016 ²⁵	ANN LR	40	Missing excluded for the analysis	Sample splitting	70:30	NR	NR
Nouei, 2014 ²⁶	ANN LR	40	Missing excluded for the analysis	Sample splitting	70:30	NR	MATLAB
Lippman, 1997 ²⁷	ANN Ensemble LR	36	Imputation for variables (statistical mode or mean substitution)	Sample splitting and k-fold cross-validation	50:50	Performed using χ^2 for comparison	LNKnet software
Mendes, 2015 ²⁸	ANN LR	12	NR	Sample splitting	80:20	NR	Accord NET Framework

(Continued)

TABLE 2. Continued

Author, y (reference no.)	Model tested	No. Predictors	Handling of missing data	Type of validation	Split ratio of training to testing sample	Calibration	Statistical software for ML
Jamaati, 2015 ¹⁷	SVM LR	17	NR	NR	-	Hosmer-Lemeshow goodness-of-fit statistic	SPSS
Tu, 1998 ¹⁸	ANN LR	17	NR	Sample splitting and k-fold cross-validation	65:35	NR	Stata
Rahman, 2012 ¹⁹	ANN DT LR	12	NR	Sample splitting	NR	NR	SPSS PASW Modeler 13
Celi, 2012 ²⁰	ANN BN LR	6	NR	Sample splitting and k-fold cross-validation	70:30	Hosmer-Lemeshow goodness-of-fit statistic	Weka and R
Macrina, 2009 ²¹	ANN LR	22	NR	External validation	-	NR	NCSS and MedCalc
Peng, 2008 ²²	ANN LR	16	NR	Sample splitting	70:30	Hosmer-Lemeshow goodness-of-fit statistic	STATISTICA from StatSoft Inc

ML, Machine learning; ANN, artificial neural networks; LR, logistic regression; EuroSCORE, European System for Cardiac Operative Risk Evaluation; EEFDT, entropy error fuzzy decision tree; EECDT, entropy error crisp decision tree; NR, not reported; GBM, gradient boosting machine; RF, random forest; NB, naive Bayesian SVM, support vector machine.

Data Analysis

Once all relevant studies were identified and corresponding results were extracted, the retrieved estimates of *C*-statistic for ML and LR models were summarized into a weighted average to provide an overall summary of their performance. A Bayesian estimation framework was used to calculate meta-analytic estimates (Appendix E1).¹²

For the main analysis, ML and LR models were extracted and pooled from each study. For studies reporting on multiple ML models, the ML model with best discrimination ability was selected. Pooled *C*-statistics for ML models and LR were then compared using the method described by Hanley and colleagues.¹³ As secondary analysis, we pooled *C*-statistics from models based on same ML algorithm and these were compared with pooled estimate from relative LR models. As a sensitivity analysis, we repeated the main comparison including studies at low and high risk of bias separately. We also stratified the analysis based on year of publication (before 2010 vs 2010 and after), validation method (external vs internal), and total sample size (≥ 1000 vs < 1000 patients). Conventional risk scoring systems (eg, EuroSCORE) pooled *C*-statistics was also reported. The presence of small-study effects was verified by visual inspection of the funnel plot and tested by fitting a regression directly to the data using the treatment effect as the dependent variable, and standard error as the independent variable for ML models performance. All analyses were performed using R version 3.5.1 and *metamisc* and *rjags* packages. All statistical tests were 2-sided.

RESULTS

Our search identified 458 citations published between June 1997 and July 2018, of which 295 studies were excluded based on title or abstract (Figure E1). Thirteen

studies were excluded during full-text screening, and 15 studies¹⁴⁻²² met inclusion criteria and were used for the quantitative and qualitative analysis. No study was found to have overlapping population with another study.

General Study Characteristics

Study characteristics are reported in Table 1 and Table 2. Notably, the first article comparing ML methods versus LR in a cardiac surgery setting was published in 1997²⁷ and it used the Society of Thoracic Surgeons (STS) database for training and testing. However, most studies were published from 2014 to 2018. Study geographic areas were Europe ($n = 3$),^{14,16,21} Asia ($n = 7$),^{15,17,19,22,24-26} North America ($n = 2$),^{18,27} South America ($n = 2$),^{23,28} and New Zealand ($n = 1$).²⁰ A total of 5 studies included unselected cardiac procedures,^{14,16,19,20,22} 7 studies focused on patients undergoing coronary artery bypass (CABG) only,^{17,18,24-28} and the remaining 3 articles included only patients undergoing heart valve surgery for rheumatic heart valve disease,²³ a combination of CABG and valve surgery,¹⁵ and type A ascending aorta dissection surgery,²¹ respectively.

In 10 studies, data were retrospectively obtained from medical health records^{14-16,19,21,22,24-26} or international surgical databases (EuroSCORE or STS).^{14,27} One study used data from the Cardiac Care Network of Ontario¹⁸ and another the Registry of Cardiac Surgery Patients in

TABLE 3. Model performance characteristics

Study, y	ML model	Testing all	Testing deaths	C-statistic	SE C-statistic
Nilsson, 2006 ¹⁴	ANN	1246	112	0.81	0.03
	LR	1246	112	0.80	0.03
	EuroSCORE	1246	112	0.79	0.03
Ghavidel, 2014 ¹⁵	DT/RF	298	12	0.90	0.06
	DT/RF (2)	298	12	0.86	0.07
	LR	298	12	0.78	0.08
	EuroSCORE	298	12	0.77	0.08
Allyn, 2017 ¹⁶	GBM	1956	123	0.78	0.02
	DT/RF	1956	123	0.79	0.02
	Naïve Bayes	1956	123	0.75	0.03
	SVM	1956	123	0.74	0.03
	Ensemble	1956	123	0.80	0.02
	LR	1956	123	0.74	0.03
	EuroSCORE	1956	123	0.72	0.03
	EuroSCORE II	1956	123	0.74	0.03
Mejia, 2018 ²³	DT/RF	584	20	0.98	0.02
	ANN	584	20	0.95	0.03
	SVM	584	20	0.95	0.04
	Naïve Bayes	584	20	0.93	0.04
	LR	584	20	0.89	0.05
	EuroSCORE II	584	20	0.86	0.05
Chong, 2003 ²⁴	ANN	140	11	0.89	0.07
	LR	140	11	0.81	0.08
Nouei (ii), 2016 ²⁵	ANN	247	8	0.82	0.09
	LR	247	8	0.62	0.11
Nouei (i), 2014 ²⁶	ANN	543	20	0.91	0.05
	LR	543	20	0.72	0.07
Lippman, 1997 ²⁷	ANN	40,126	1374	0.76	0.01
	Naïve Bayes	40,126	1374	0.75	0.01
	Ensemble	40,126	1374	0.76	0.01
	LR	40,126	1374	0.76	0.01
Mendes, 2015 ²⁸	ANN	262	22	0.85	0.05
	LR	262	22	0.86	0.05
Tu, 1998 ¹⁸	ANN	5517	173	0.78	0.02
	LR	5517	173	0.77	0.02
Jamaati, 2015 ¹⁷	SVM	2220	270	0.98	0.01
	LR	2220	270	0.84	0.02
Peng, 2008 ²²	ANN	315	37	0.87	0.04
	LR	315	37	0.85	0.04
Macrina, 2009 ^{21*}	ANN	87	20	0.93	0.05
	LR	87	20	0.88	0.06
Celi, 2012 ²⁰	ANN	165	12	0.94	0.05
	Naïve Bayes	165	12	0.93	0.06
	LR	165	12	0.85	0.07
	EuroSCORE	165	12	0.65	0.09
Rahman, 2012 ^{19†}	ANN	1209	209	0.91	0.01
	DT/RF	1209	209	0.91	0.01
	LR	1209	209	0.89	0.02

ML, Machine learning; SE, standard error; ANN, artificial neural networks; LR, logistic regression; EuroSCORE, European System for Cardiac Operative Risk Evaluation; DT/RF, decision tree/random forests; GBM, gradient boosting machine; SVM, support vector machine. *Derived from Gini coefficient as reported in original article. †Derived from sensitivity and specificity as reported in original article.

TABLE 4. Meta-analytic estimates

Variable	No. studies	ML pooled C-statistic (95% credible interval)	LR pooled C-statistic (95% credible interval)	Net benefit (%)	P value
Best ML model overall	15	0.88 (0.83-0.93)	0.81 (0.77-0.85)	+7	.03
Artificial neural network	12	0.86 (0.81-0.91)	0.81 (0.76-0.86)	+5	.15
Decision trees/random forest	4	0.89 (0.76-0.98)	0.80 (0.63-0.90)	+9	.30
Support vector machine	3	0.92 (0.75-1.00)	0.82 (0.65-0.96)	+10	.27
Naïve Bayes	4	0.81 (0.69-0.96)	0.78 (0.68-0.91)	+3	.8
Other	2	0.77 (0.70-0.87)	0.76 (0.54-0.96)	+1	.92
Best ML model-low risk of bias	10	0.85 (0.79-0.91)	0.79 (0.73-0.85)	+6	.15
Best ML model-high risk of bias	5	0.92 (0.82-0.98)	0.84 (0.79-0.90)	+8	.10
Best ML model \geq 2010	9	0.91 (0.84-0.97)	0.81 (0.74-0.88)	+10	.02
Best ML model <2010	6	0.81 (0.75-0.89)	0.78 (0.74-0.85)	+3	.5
Best ML model EV	2	0.85 (0.64-0.99)	0.82 (0.61-0.99)	+3	.8
Best ML model IV	13	0.88 (0.82-0.94)	0.80 (0.76-0.85)	+8	.04
Best ML model \geq 1000 patients	9	0.89 (0.79-0.95)	0.80 (0.75-0.86)	+9	.07
Best ML model < 1000 patients	6	0.88 (0.82-0.94)	0.82 (0.70-0.90)	+6	.13

ML, Machine learning; LR, logistic regression; EV, external validation; IV, internal validation.

Dunedin Hospital.²⁰ Data were prospectively collected only in 3 studies.^{17,23,28} Sample size ranged from 165 to 80,606 patients and operative mortality from 3.0% to 25.5%. ML models developed were artificial neural network (n = 12),^{14,18-28} decision tree analysis (n = 2),^{15,19} random forest (n = 2),^{16,23} support vector machine (n = 3),^{16,17,23} naïve Bayes (n = 3),^{16,20,23} gradient boost machine (n = 1),¹⁶ and ensemble of models (n = 2).^{16,27} ML models are described in Table E2. With the exception of 1 study,²⁴ all studies performed LR model using the same set of variables to compare its performance with ML models. The only traditional scoring systems for cardiac surgery evaluated was EuroSCORE either the original (n = 3)¹⁴⁻¹⁶ or the updated version (EuroSCORE II) (n = 2).^{16,23} C-statistic was the performance measure used in 13 studies,^{14-18,20,22-28} sensitivity and specificity,¹⁹ and Gini coefficient²¹ were used in the remaining 2 studies, and C-statistic was derived using conversion equations (details are in Appendix E1).

For ML models, the C-statistic ranged from 0.736 to 0.982 (Figure E2) and for LR from 0.620 to 0.890. The number of variables included in the models ranged from 6 to 40. Validation was performed using both sample splitting and k-fold cross-validation in 6 studies^{15,16,18,20,24,27} and sample splitting only in 5^{19,22,25,26,28} studies. Other validation methods adopted were k-fold cross-validation only (n = 1),²³ combination of sample splitting and k-fold cross-validation and external validation (n = 1),¹⁴ and external validation only (n = 1).²¹ In 1 study, the validation method was not reported.¹⁷ Calibration was reported only in 4 studies^{17,20,22,27} (details regarding calibration assessment are presented in Table E3).

Information on handling of missing data was lacking or unclear in 8 studies.^{16-22,28} In the remaining studies, missing

data were handled using complete case analysis (n = 4),^{15,23,25,26} single imputation (n = 2),^{24,27} and a combination of complete case for mandatory variables and single imputation for other variables (n = 1).¹⁴ Statistical software used for ML modeling was reported in all but 1 study.

Methodological Quality

Ten (67%) studies were at low risk of bias,^{14-16,18-20,23,25-27} whereas the remaining 5 (33%) were classified as at high risk of bias (Table E1). In the study by Chong and colleagues,²⁴ although the original number of input variables included in ML and LR models was 21, it was unclear why the final number of input variable predictors in the ML model was 18. In the study by Mendes and colleagues,²⁸ it was unclear whether input variables scaling into centered unit interval and correction for imbalanced outcomes was used to develop ML methods but no LR. In the study by Jamaati and colleagues,¹⁷ it was unclear to assess any validation methodology used. Peng and colleagues²² ran a data-driven variable selection for LR model but not for ML and similarly, the study by Macrina and colleagues.²¹

Comparison Between Performance of ML and LR Models

Individual study reported or derived C-statistics with relative standard error are presented in Table 3. Meta-analytic estimates with relative 95% credibility interval (CrI) for ML and LR models across different analyses are reported in Table 4. The main analysis based on best-performing ML models from each study, showed that when compared to LR, ML models were associated with a statistically significant improvement in C-statistic (ML, 0.88; 95% CrI, 0.83-0.93 vs LR, 0.81; 95% CrI,

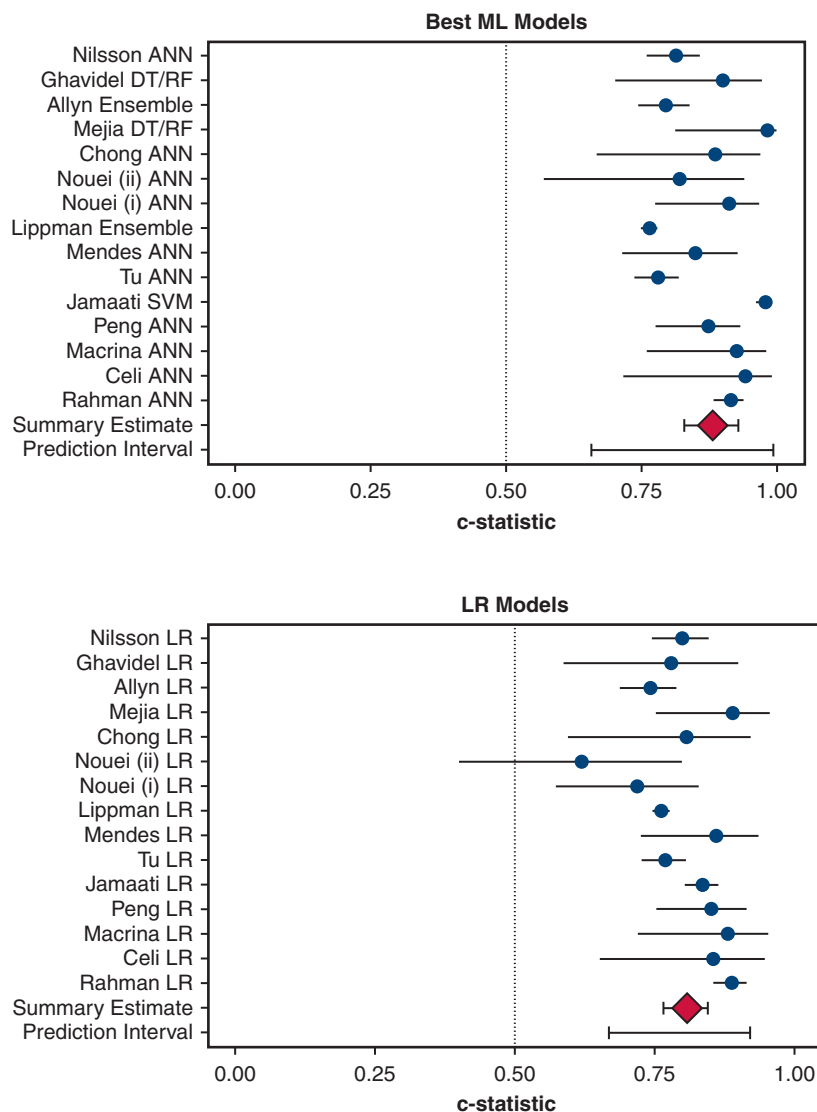
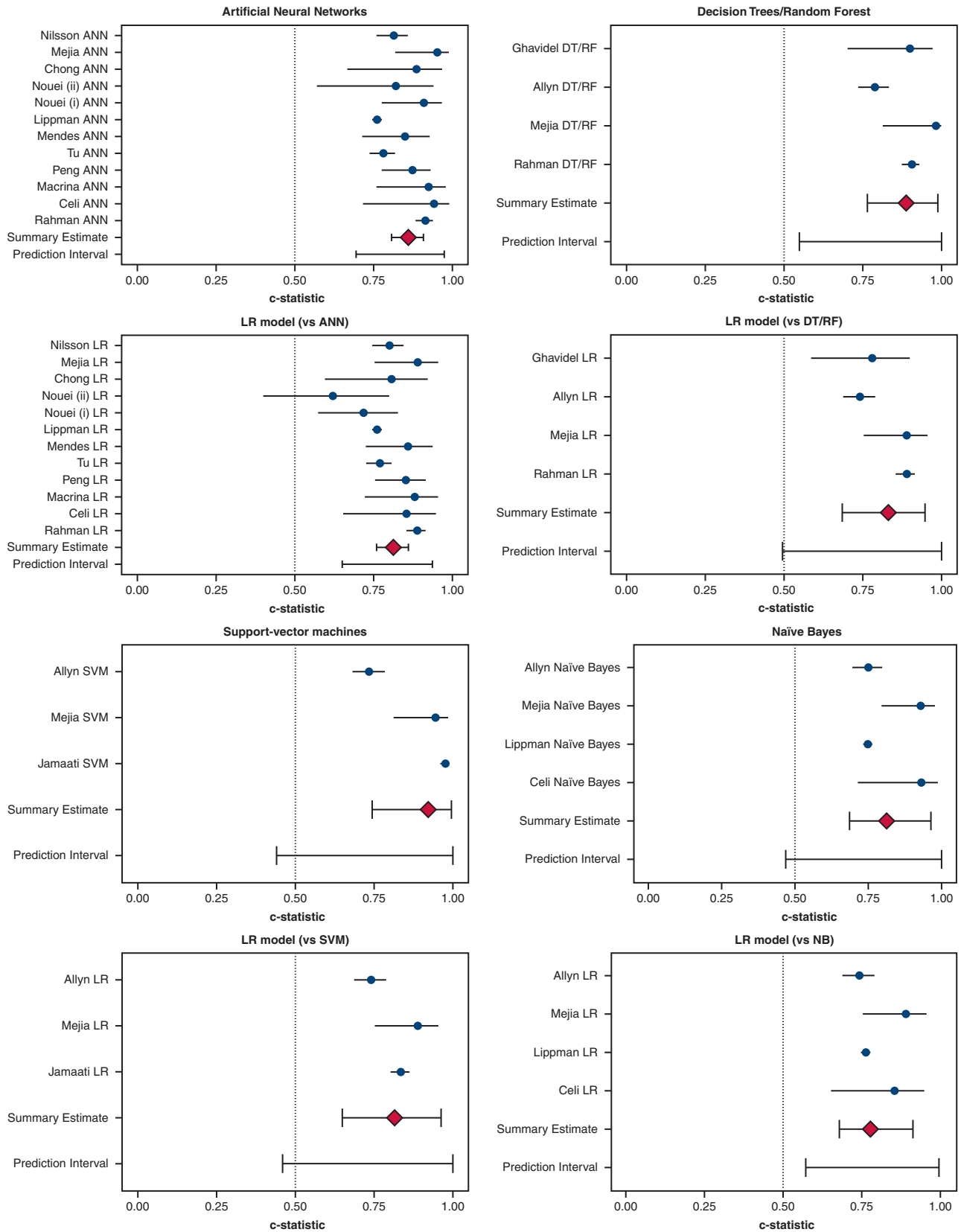


FIGURE 1. Forest plot comparing discrimination accuracy (ie, *C*-statistic) in mortality prediction by selecting machine learning (*ML*) models (*top*) with best performance versus logistic regression (*LR*) (*bottom*). *ANN*, Artificial neural networks; *DT*, decision tree; *RF*, random forests; *SVM*, support vector machine.

0.77-0.85; $P = .03$) (Figure 1). When the analysis was stratified by individual ML categories, artificial neural networks (0.86; 95% CrI, 0.81-0.91 vs LR, 0.81; 95% CrI, 0.76-0.86; $P = .15$), decision trees/random forest (0.89; 95% CrI, 0.76-0.98 vs LR, 0.80; 95% CrI, 0.63-0.90; $P = .30$), support vector machine (0.92; 95% CrI, 0.75-1.00 vs LR, 0.82; 95% CrI, 0.65-0.96; $P = .27$), and naïve Bayes (0.81; 95% CrI, 0.69-0.96 vs LR, 0.78; 95% CrI, 0.68-0.91; $P = .8$) achieved higher *C*-statistics but improvement was non statistically significant (Figure 2).

Sensitivity analysis showed that in studies at high risk of bias, both ML model and LR showed a higher *c*-statistic (ML, 0.92; 95% CrI, 0.82-0.98 vs LR 0.84; 95% CrI, 0.79-0.90; $P = .15$) than studies low risk of bias (ML,

0.85; 95% CrI, 0.79-0.91 vs LR, 0.79; 95% CrI, 0.73-0.85; $P = .10$) (Figure E1). Furthermore, when compared with LR, ML models achieved a better discrimination accuracy in studies published from 2010 onward (ML, 0.9; 95% CrI, 0.84-0.97 vs LR, 0.81; 95% CrI, 0.74-0.88; $P = .02$) than in studies published before 2010 (ML, 0.81; 95% CrI, 0.75-0.89 vs LR, 0.78; 95% CrI, 0.74-0.85; $P = .5$). We found a trend toward better ML model performance when the models were developed using internal validation and larger samples. Funnel plot and regression test showed no evidence of small study effect ($P = .70$) (Figure E3). Information on original EuroSCORE and EuroSCORE II performance was available in 4 and 2 studies, respectively, and pooled *C*-statistics was 0.74 (95% CrI, 0.61-0.86) and (0.78; 95% CrI, 0.53-0.99), respectively. Assessment of



MIS

FIGURE 2. Forest plot comparing discrimination accuracy (ie, C-statistic) in mortality prediction by selecting machine learning (ML) models based on the same algorithm versus logistic regression (LR). ANN, Artificial neural networks; DT, decision tree; RF, random forests; SVM, support vector machine; NB, naive Bayesian.

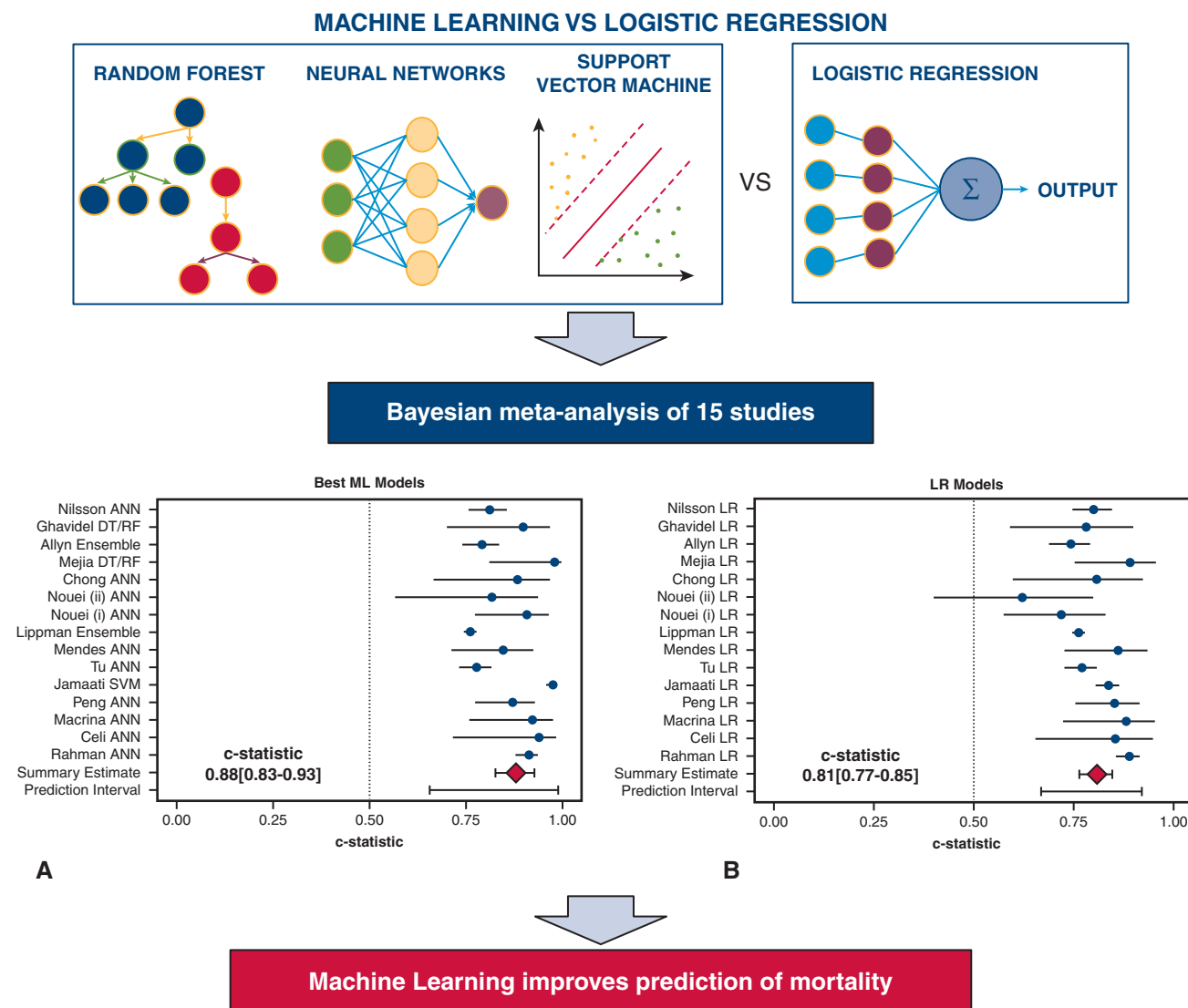


FIGURE 3. Machine learning (ML) algorithms (ie, random forest, neural networks, and support vector machine) were compared with traditional logistic regression in the prediction of mortality after cardiac surgery using a Bayesian meta-analysis of 15 studies. Model performance was estimated using C-statistics. ML models achieved a better prediction than logistic regression. ANN, Artificial neural networks; DT, decision tree; RF, random forests; SVM, support vector machine.

model calibration was reported only by a limited number of studies and different methodologies were used preventing any meta-analytic estimation. A descriptive summary of assessment of model calibration for studies reporting on this information is presented in [Table E3](#).

DISCUSSION

The present meta-analysis showed that ML models can achieve significantly better discrimination ability than LR when both models are on the same features ([Figure 3](#)). A significant improvement could be demonstrated only when the best performing ML model among all ML models investigated was selected from individual studies; however, we could not demonstrate a superiority from a specific ML

model. We also found a trend toward improved performance with ML models over LR in more recently published studies. This may be related to recent improvement in ML algorithms and increased popularity of dedicated statistical software. There has been a growing interest in risk-prediction models for clinical use to aid in multidisciplinary shared decision making. They are also used for both benchmarking outcomes and monitoring innovations. The clinical use is gaining increasing importance, especially in an era of expanding multimodal therapy for coronary artery and aortic valve disease; risk prediction plays an important role in determining which patients would benefit most from surgery or percutaneous therapy. National cardiac surgical registries have been established in many countries and

have developed risk prediction models suitable for local populations.

Risk stratification in cardiac surgery patients is usually performed using EuroSCORE II² and the STS-Predicted Risk of Mortality Score,³ which were developed based on LR. However, EuroSCORE II, as well as the logistic EuroSCORE have been shown to overestimate the actual risk especially in high-risk but also in low-risk subgroups and therefore they offer little information and guidance to the clinicians' judgment.^{4,5,29} Poor performance of current models can be partially attributed to the fact that these models require modeler input to specify complex interactions among variables. For instance, the contribution of a feature; for example, age, to the risk of mortality may not be equal and constant across the spectrum of coexisting comorbidities and surgical procedures. Although simplified models are associated with lower variance, they may also result in miscalibrated estimates. Due to the need for more precise and accurate risk predictions, the application of ML approaches for the development of clinical prediction rules has been increasingly investigated. Risk models based on ML have mainly focused on mortality prediction after cardiac surgery, but also on the development of other adverse events such as acute kidney injury,³⁰ major bleeding,³¹ and prolonged ventilation.³²

The potential advantage from ML models over traditional LR is their ability to capture nonlinearity and the interactions among features without the need for the modeler to manually specify all interactions, as needed with LR. Moreover, compared with traditional statistical methods, ML algorithms can handle missing data more efficiently because they do not rely on data distribution assumptions and are capable of more complex calculation.³³ The present findings support the hypothesis that ML models can achieve better discrimination in the prediction of mortality after cardiac surgery when compared with LR. However, a significant improvement with ML models was demonstrated only when the best ML model from each study was selected, thus pooling different type of ML algorithms. When the analysis focused on individual ML model categories, such an improvement was not significant. This can be partially related to lower power of subgroup analysis. However, this also support the so-called No Free-Lunch theorem in ML,³⁴ which states that there is no 1 model that works best for every problem or every dataset. The assumptions of a good model for 1 problem may not hold for another problem, so it is common in ML to try multiple models and find 1 that works best for a problem. This is because ML algorithms make some assumptions (known as learning bias) about the relationships between the predictor and target variables, introducing bias into the model. The assumptions made by ML algorithms mean that some algorithms will fit certain data sets better than others.

Therefore, the magnitude and clinical influence of improvement using ML remains uncertain. ML modeling needs far more events per variable to achieve a stable C-statistic than LR and should only be considered if very large data sets with many events are available.³⁵ Both ML and LR models can perform poorly when the prediction tool is developed using a data set that is small and/or has a low incidence of events. Substantial gain in prediction is unlikely to be determined by the application of ML algorithms alone, in particular when we can rely on a small subset of structured clinical data. Moreover, ML algorithms tend to produce unsatisfactory classifiers when faced with an imbalanced dataset, when the number of observations belonging to 1 class is significantly lower than those belonging to the other classes. This is because ML algorithms are designed to maximize accuracy (ie, proportion of correct predictions) and reduce error. However, in the presence of class imbalance, ML models can predict the value of the majority class for all predictions and achieve a high classification accuracy, but this model may present a high probability of misclassification of the minority class. This is called accuracy paradox.³⁶ In these cases, it may be desirable to select a model with a lower accuracy because it has a greater predictive power on the problem. Class imbalance can be tackled with different strategies such as over- and undersampling or algorithm-centered approaches that modify the algorithm to favor its prediction toward the less-represented class.³⁷ The problem of class imbalance may be particularly relevant when ML is applied for prediction in cardiac surgery because the incidence of adverse events is very low. For LR models, unbalanced training data affects only the estimate of the model intercept which can be corrected using a rare events correction to the intercept.³⁸ Moreover, traditional risk models are developed using structured dataset (ie, EuroSCORE or STS score).^{2,3} These databases contain only a restricted number of prespecified variables limiting the capability of ML that may perform best by exploiting high dimensional data from electronic medical records.³⁹

Better quantification of mortality risk is likely to be associated with the identification of other variables that explain more of the variance observed. Moreover, as a significant amount of patient data are available in unstructured formats like images and clinical notes, modeling approaches (such as deep learning) that can automatically extract novel features from these sources represent an emerging and attractive strategy to significantly improve risk prediction and provide reliable and objective tool in decision making.

Limitations

We focused on the performance of individual ML and LR algorithms based on the same set of variables, all predictive of the outcome of interest. Limiting the number of variables in the ML models may have reduced their discrimination

power. In fact, it is possible that some ML models can further improve prediction using many variables without incurring in overfitting, which is more frequent and detrimental for parametric models, such as LR.

Studies included a range of different cardiac surgical procedures and different populations from different continents and this can cause significant variation in model performance.

Five out of 15 studies had poor methodology and reporting.^{17,21,22,24,28} When studies at high risk of bias were removed from the analysis the advantage from ML models over LR was further reduced. Four studies out of 15 evaluated model performance in terms of calibration (whether risk estimates are accurate)^{17,20,22,27} and only 1 study assessed clinical utility for decision making by decision curve analysis,¹⁶ which is increasingly used in medical applications.⁴⁰

Also, all studies involving the use of ML to derive information from images/signals were excluded and this may have limited the benefit from applying ML approaches.

Moreover, reporting of articles that compare both types of algorithms needs to improve. Correct validation procedures are needed, with assessment of calibration and clinical utility in addition to discrimination, to define situations where modern methods have advantages over traditional approaches.

CONCLUSIONS

The present meta-analysis showed that when compared with LR, ML models achieved better discrimination ability in predicting operative mortality after cardiac surgery. However, the clinical implication of this finding remains unclear.

Conflict of Interest Statement

The authors reported no conflicts of interest.

The *Journal* policy requires editors and reviewers to disclose conflicts of interest and to decline handling or reviewing manuscripts for which they may have a conflict of interest. The editors and reviewers of this article have no conflicts of interest.

The authors thank Dr Giovanni Morlino for his support with statistical analysis.

References

- Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. Euro-pean system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardio-thorac Surg.* 1999;16:9-13.
- Nashef SAM, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. EuroSCORE II. *Eur J Cardiothorac Surg.* 2012;41:734-5.
- Shahian DM, O'Brien SM, Filardo G, Ferraris VA, Haan CK, Rich JB, 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.
- Gummert JF, Funkat A, Osswald B, Beckmann A, Schiller W, Krian A, et al. EuroSCORE overestimates the risk of cardiac surgery: results from the national registry of the German Society of Thoracic and Cardiovascular Surgery. *Clin Res Cardiol.* 2009;98:363-9.
- Ad N, Holmes SD, Patel J, Pritchard G, Shuman DJ, Halpin L. Comparison of EuroSCORE II, original EuroSCORE, and the Society of Thoracic Surgeons risk score in cardiac surgery patients. *Ann Thorac Surg.* 2016;102:573-9.
- Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol.* 2019;110:12-22.
- Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLOS Med.* 2014;11:e1001744.
- Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol.* 2003;3:25.
- Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010;21:128-38.
- NCSS Statistical Software. Confidence intervals for the area under an ROC curve. Chapter 26. Available at: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/PASS/Confidence_Intervals_for_the_Area_Under_an_ROC_Curve.pdf. Accessed August 14, 2020.
- Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol.* 2019;19:64.
- Debray TP, Damen JA, Riley RD, Snell K, Reitsma JB, Hooft L, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res.* 2019;28:2768-86.
- 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.
- Nilsson J, Ohlsson M, Thulin L, Hoglund P, Nashef SAM, Brandt J. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. *J Thorac Cardiovasc Surg.* 2006;132:12-9.
- Ghavidel AA, Javadikasgari H, Maleki M, Karbassi A, Omrani G, Noohi F. Two new mathematical models for prediction of early mortality risk in coronary artery bypass graft surgery. *J Thorac Cardiovasc Surg.* 2014;148:1291-8.e1.
- Allyn J, Allou N, Augustin P, Philip I, Martinet O, Belghiti M, et al. A comparison of a machine learning model with EuroSCORE II in predicting mortality after elective cardiac surgery: a decision curve analysis. *PLoS One.* 2017;12:e0169772.
- Jamaati H, Najafi A, Kahe F, Karimi Z, Ahmadi Z, Bolursaz M, et al. Assessment of the EuroSCORE risk scoring system for patients undergoing coronary artery bypass graft surgery in a group of Iranian patients. *Indian J Crit Care Med.* 2015;19:576-9.
- Tu JV, Weinstein MC, McNeil BJ, Naylor CD. Predicting mortality after coronary artery bypass surgery: what do artificial neural networks learn? The steering committee of the cardiac care network of Ontario. *Med Decis Making.* 1998;18:229-35.
- Rahman HAA, Wah YB, Khairudin Z, Abdullah NN. Comparison of predictive models to predict survival of cardiac surgery patients. Available at: <https://ieeexplore.ieee.org/document/6396534>. Accessed August 14, 2020.
- Celi LA, Galvin S, Davidzon G, Lee J, Scott D, Mark R. A database-driven decision support system: customized mortality prediction. *J Pers Med.* 2012;2:138-48.
- Macrina F, Puddu PE, Sciangula A, Trigilia F, Totaro M, Miraldi F, et al. Artificial neural networks versus multiple logistic regression to predict 30-day mortality after operations for type a ascending aortic dissection. *Open Cardiovasc Med J.* 2009;3:81-95.
- Peng S-Y, Peng S-K. Predicting adverse outcomes of cardiac surgery with the application of artificial neural networks. *Anaesthesia.* 2008;63:705-13.
- Mejia OAV, Antunes MJ, Goncharov M, Dallan LRP, Veronese E, Lapenna GA, et al. Predictive performance of six mortality risk scores and the development of a novel model in a prospective cohort of patients undergoing valve surgery secondary to rheumatic fever. *PLoS One.* 2018;13:e0199277.
- Chong C-F, Li Y-C, Wang T-L, Chang H. Stratification of adverse outcomes by preoperative risk factors in coronary artery bypass graft patients: an artificial neural network prediction model. *AMIA Annu Symp Proc.* 2003;2003:160-4.

25. Nouei MT, Kamyad AV, Sarzaeem M, Ghazalbash S. Fuzzy risk assessment of mortality after coronary surgery using combination of adaptive neuro-fuzzy inference system and K-means clustering. *Expert Syst.* 2016;33:230-8.
26. Nouei MT, Kamyad AV, Sarzaeem M, Ghazalbash S. Developing a genetic fuzzy system for risk assessment of mortality after cardiac surgery. *J Med Syst.* 2014; 38:102.
27. Lippmann RP, Shahian DM. Coronary artery bypass risk prediction using neural networks. *Ann Thorac Surg.* 1997;63:1635-43.
28. Mendes RG, de Souza CR, Machado MN, Correa PR, Di Thommazo-Luporini L, Arena R, et al. Predicting reintubation, prolonged mechanical ventilation and death in post-coronary artery bypass graft surgery: a comparison between artificial neural networks and logistic regression models. *Arch Med Sci.* 2015;11:756-63.
29. Kieser TM, Rose MS, Head SJ. Comparison of logistic EuroSCORE and EuroSCORE II in predicting operative mortality of 1125 total arterial operations. *Eur J Cardiothorac Surg.* 2016;50:509-18.
30. Lee H-C, Yoon H-K, Nam K, Cho YJ, Kim TK, Kim WH, et al. Derivation and validation of machine learning approaches to predict acute kidney injury after cardiac surgery. *J Clin Med.* 2018;7:322.
31. Meyer A, Zverinski D, Pfahringer B, Kempfert J, Kuehne T, Sündermann SH, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. *Lancet Respir Med.* 2018;6:905-14.
32. Wise ES, Stonko DP, Glaser ZA, Garcia KL, Huang JJ, Kim JS, et al. Prediction of prolonged ventilation after coronary artery bypass grafting: data from an artificial neural network. *Heart Surg Forum.* 2017;20:E007-14.
33. Gupta A, Lam MS. Estimating missing values using neural networks. *J Oper Res Soc.* 1996;47:229-38.
34. Gomez D, Rojas A. An empirical overview of the no free lunch theorem and its effect on real-world machine learning classification. *Neural Comput.* 2016;28: 216-28.
35. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol.* 2014;14:137.
36. Valverde-Albacete FJ, Peláez-Moreno C. 100% classification accuracy considered harmful: the normalized information transfer factor explains the accuracy paradox. *PLoS One.* 2014;9:e84217.
37. Kaur H, Pannu HS, Malhi AK. A systematic review on imbalanced data challenges in machine learning: applications and solutions. *ACM Comput Surv.* 2019;52:1-36.
38. Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A. Firth's logistic regression with rare events: accurate effect estimates and predictions? *Stat Med.* 2017;36: 2302-17.
39. Thottakkara P, Ozrazgat-Baslanti T, Hupf BB, Rashidi P, Pardalos P, Momcilovic P, et al. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PLoS One.* 2016;11:e0155705.
40. Van Calster B, Wynants L, Verbeek JFM, Verbakel JY, Christodoulou E, Vickers AJ, et al. Reporting and interpreting decision curve analysis: a guide for investigators. *Eur Urol.* 2018;74:796-804.

Key Words: risk model, prediction, mortality, machine learning, logistic regression, meta-analysis

APPENDIX E1. METHODS

Search Strategy

The search strategy was the following: (*cardiac surgery* or *heart surgery*) and (*risk OR prediction OR mortality*) and (*machine learning OR artificial intelligence OR deep learning OR neural network OR random forest OR decision tree OR support vector machine*). An AND statement was used to connect 3 searches: 1 capturing artificial intelligence, 1 cardiac surgery topic, and the last capturing risk prediction and mortality. There was no restriction on the publication date, but only articles written in English were included.

Derivation of C-statistic When Not Reported

We derived C-statistics from Gini coefficient (G) with the following formula^{E1}):

$$c - statistic = \frac{G+1}{2}$$

When only sensitivity and specificity were reported, we first calculated the diagnostic odds ratio (DOR)^{E2}:

$$DOR = \frac{sensitivity \times specificity}{(1 - sensitivity) \times (1 - specificity)}$$

and then derived the C-statistic^{E3}:

$$c - statistic = \frac{DOR}{(DOR - 1)^2} [(DOR - 1) - \ln(DOR)]$$

Details of Bayesian Estimation Framework

Bayesian methods use formal probability models to express uncertainty about parameter values. This is particularly relevant when confronting sparse data (ie, case-mix variation) or multiple comparisons. Bayesian inference consists of repeatedly sampling from a posterior distribution to get parameter estimates and their variance. Just Another Gibbs Sampler^{E4} and Markov Chain Monte Carlo simulation were used for sampling (details in Tables E1 to E3). The Markov Chain Monte Carlo sampling procedure was based on the following parameters: number of interactions = 10,000, burn-in period = 5000, number of chains = 4, noninformative normal prior for the mean equal to 0 and a uniform prior for the between study variance of the pooled effect size bounded between 0 and 100. The convergence of all estimated Bayesian meta-analysis models was verified by calculating the potential scale reduction factor of the Gelman-Rubin statistic autocorrelation of the sample (>1.05 indicative of nonconvergence).^{E4} Pooled C-statistics and 95% credibility interval were directly obtained from the corresponding posterior quantiles for machine learning and logistic regression;

models from the same validation sample. We also calculated a 95% prediction interval to depict the extent of between-study heterogeneity. This interval provides a range that likely contains a future prediction when the model is applied to a new dataset.

E-References

- E1. Hand DJ. Assessing the performance of classification methods. *Int Stat Rev.* 2012;80:400-14.
- E2. Glas AS, Lijmer JG, Prins MH, Bossel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol.* 2003;56:1129-35.
- E3. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med.* 2002;21:1237-56.
- E4. Peng S-Y, Peng S-K. Predicting adverse outcomes of cardiac surgery with the application of artificial neural networks. *Anaesthesia.* 2008;63:705-13.
- E5. Nilsson J, Ohlsson M, Thulin L, Hoglund P, Nashef SAM, Brandt J. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. *J Thorac Cardiovasc Surg.* 2006;132:12-9.
- E6. Ghavidel AA, Javadikasgari H, Maleki M, Karbassi A, Omrani G, Noohi F. Two new mathematical models for prediction of early mortality risk in coronary artery bypass graft surgery. *J Thorac Cardiovasc Surg.* 2014;148:1291-8.e1.
- E7. Allyn J, Allou N, Augustin P, Philip I, Martinet O, Belghiti M, et al. A comparison of a machine learning model with EuroSCORE II in predicting mortality after elective cardiac surgery: a decision curve analysis. *PLoS One.* 2017;12:e0169772.
- E8. Mejia OAV, Antunes MJ, Goncharov M, Dallan LRP, Veronese E, Lapenna GA, et al. Predictive performance of six mortality risk scores and the development of a novel model in a prospective cohort of patients undergoing valve surgery secondary to rheumatic fever. *PLoS One.* 2018;13:e0199277.
- E9. Chong C-F, Li Y-C, Wang T-L, Chang H. Stratification of adverse outcomes by preoperative risk factors in coronary artery bypass graft patients: an artificial neural network prediction model. *AMIA Annu Symp Proc.* 2003;2003:160-4.
- E10. Nouei MT, Kamyad AV, Sarzaem M, Ghazalbash S. Fuzzy risk assessment of mortality after coronary surgery using combination of adaptive neuro-fuzzy inference system and K-means clustering. *Expert Syst.* 2016;33:230-8.
- E11. Nouei MT, Kamyad AV, Sarzaem M, Ghazalbash S. Developing a genetic fuzzy system for risk assessment of mortality after cardiac surgery. *J Med Syst.* 2014;38:102.
- E12. Lippmann RP, Shahian DM. Coronary artery bypass risk prediction using neural networks. *Ann Thorac Surg.* 1997;63:1635-43.
- E13. Mendes RG, de Souza CR, Machado MN, Correa PR, Di Thommazzo-Luporini L, Arena R, et al. Predicting reintubation, prolonged mechanical ventilation and death in post-coronary artery bypass graft surgery: a comparison between artificial neural networks and logistic regression models. *Arch Med Sci.* 2015;11:756-63.
- E14. Tu JV, Weinstein MC, McNeil BJ, Naylor CD. Predicting mortality after coronary artery bypass surgery: what do artificial neural networks learn? The Steering Committee of the Cardiac Care Network of Ontario. *Med Decis Making.* 1998;18:229-35.
- E15. Jamaati H, Najafi A, Kahe F, Karimi Z, Ahmadi Z, Bolursaz M, et al. Assessment of the EuroSCORE risk scoring system for patients undergoing coronary artery bypass graft surgery in a group of Iranian patients. *Indian J Crit Care Med.* 2015;19:576-9.
- E16. Macrina F, Puddu PE, Sciangula A, Trigilia F, Totaro M, Miraldi F, et al. Artificial neural networks versus multiple logistic regression to predict 30-day mortality after operations for type a ascending aortic dissection. *Open Cardiovasc Med J.* 2009;3:81-95.
- E17. Celi LA, Galvin S, Davidzon G, Lee J, Scott D, Mark R. A database-driven decision support system: customized mortality prediction. *J Pers Med.* 2012;2:138-48.
- E18. Rahman HAA, Wah YB, Khairudin Z, Abdullah NN. Comparison of predictive models to predict survival of cardiac surgery patients. Available at: <https://ieeexplore.ieee.org/document/6396534>. Accessed August 14, 2020.
- E19. Drew PJ, Monson JRT. Artificial neural networks. *Surgery.* 2000;127:3-11.
- E20. Kingsford C, Salzberg SL. What are decision trees? *Nat Biotechnol.* 2008;26:1011-3.

- E21. Natekin A, Knoll A. Gradient boosting machines, a tutorial. *Front Neurobot.* 2013;7:21.
- E22. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genom Proteom.* 2018;15:41-51.
- E23. Zhang Z. Naïve Bayes classification in R. *Ann Transl Med.* 2016;4:241.
- E24. Sarica A, Cerasa A, Quattrone A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. *Front Aging Neurosci.* 2017;9:329.

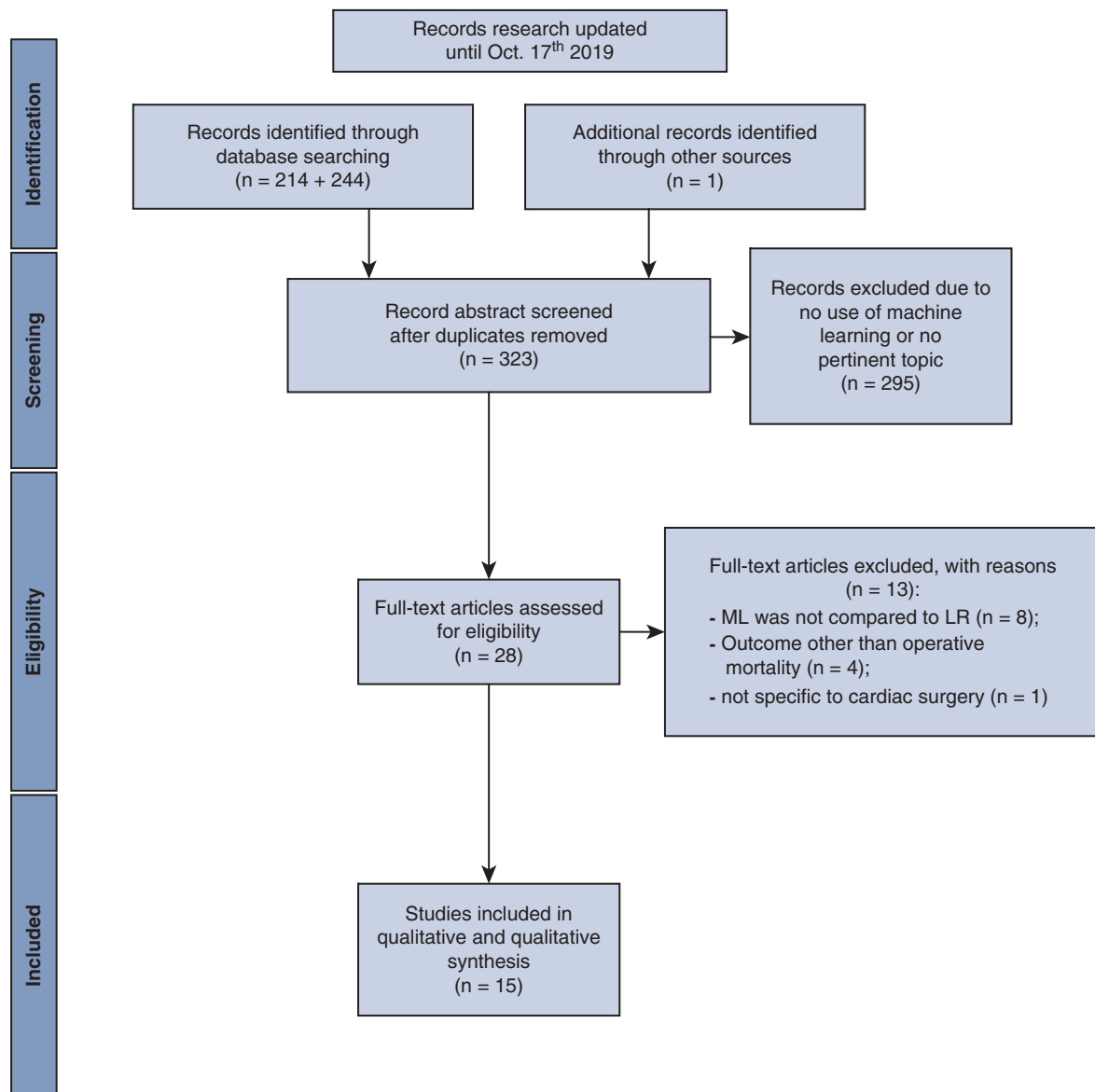


FIGURE E1. Preferred Reporting Items for Systematic Reviews and Meta-Analysis flow chart of search strategy. *ML*, Machine learning; *LR*, logistic regression.

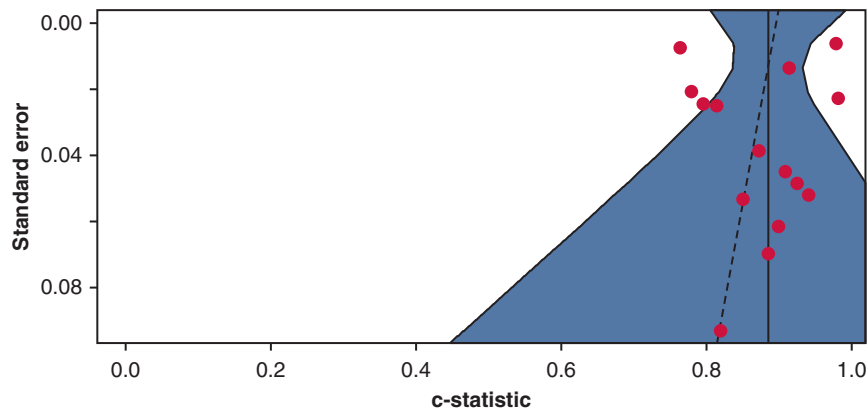


FIGURE E2. Funnel plot for assessment of small-study effect, obtained by plotting the C-statistics and the standard error for each study included.

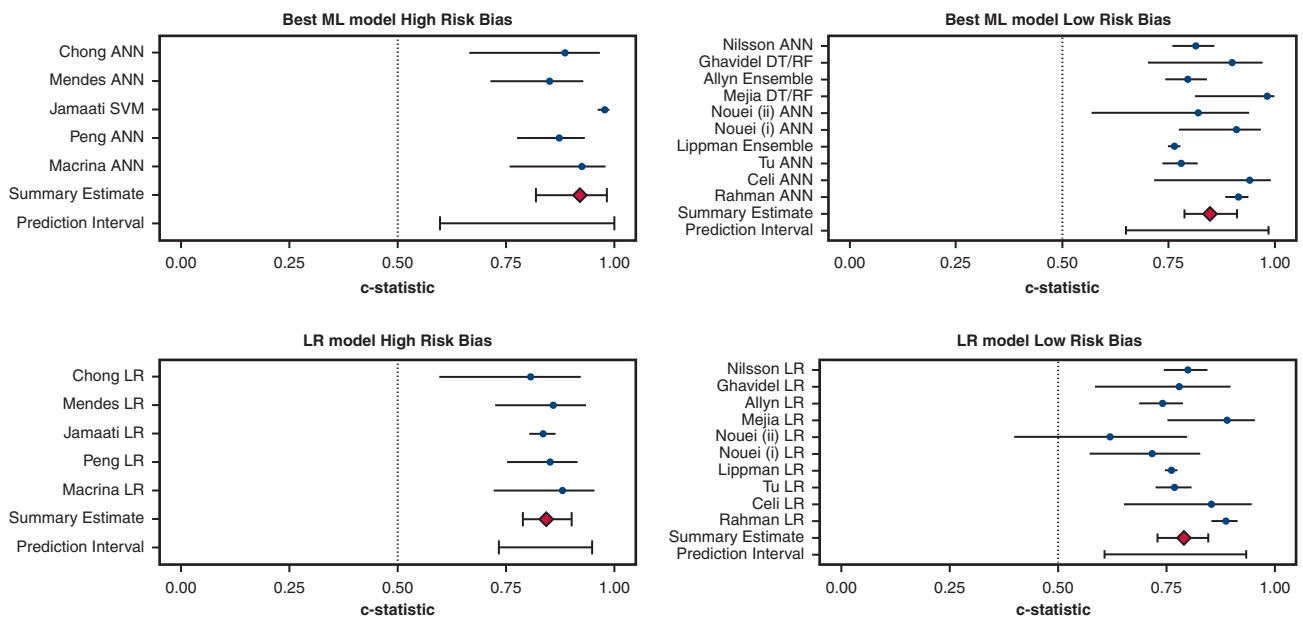


FIGURE E3. Forest plot comparing discrimination accuracy (ie, C-statistic) in mortality prediction by selecting machine learning (ML) models (top) with best performance versus logistic regression (LR) in studies with high (left) and low risk of bias (right). ANN, Artificial neural networks; SVM, support vector machine; DT, decision tree; RF, random forests.

MIS

TABLE E1. Assessment of study risk of bias*

Author, y	Item 1	Item 2	Item 3	Item 4	Item 5	Risk of bias
Nilsson, 2006 ^{E5}	NO	NO	NO	NO	NO	LOW
Ghavidel, 2014 ^{E6}	NO	NO	NO	NO	NO	LOW
Allyn, 2017 ^{E7}	NO	NO	NO	NO	NO	LOW
Mejia, 2018 ^{E8}	NO	NO	NO	NO	NO	LOW
Chong, 2003 ^{E9}	NO	NO	NO	UNCLEAR	NO	HIGH
Nouei, 2016 ^{E10}	NO	NO	NO	NO	NO	LOW
Nouei, 2014 ^{E11}	NO	NO	NO	NO	NO	LOW
Lippman, 1997 ^{E12}	NO	NO	NO	NO	NO	LOW
Mendes, 2015 ^{E13}	NO	NO	UNCLEAR	NO	UNCLEAR	HIGH
Tu, 1998 ^{E14}	NO	NO	NO	NO	NO	LOW
Jamaati, 2015 ^{E15}	YES	NO	NO	NO	NO	HIGH
Peng, 2008 ^{E4}	NO	YES	NO	NO	NO	HIGH
Macrina, 2009 ^{E16}	NO	YES	NO	NO	NO	HIGH
Celi, 2012 ^{E17}	NO	NO	NO	NO	NO	LOW
Rahman, 2012 ^{E18}	NO	NO	NO	NO	NO	LOW

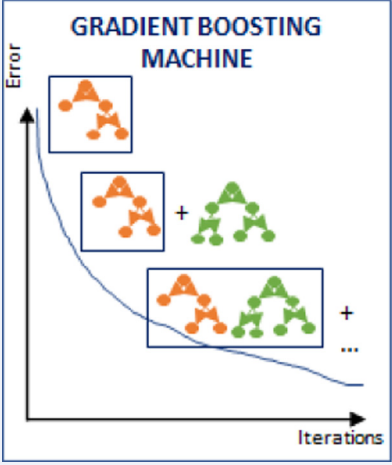
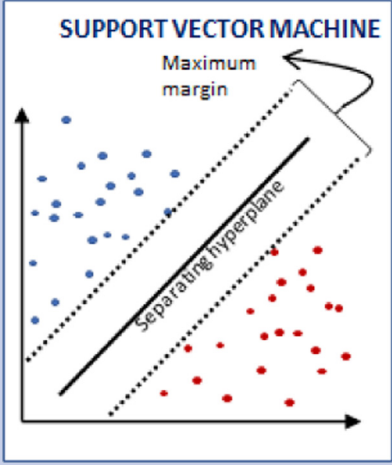
*We considered a comparison at low risk of bias if the answer was NO for all 5 signalling items. If the answer was UNCLEAR or YES for at least 1 item, we assumed high risk of bias. Item 1: Unclear or biased validation of model performance, item 2: Difference in data-driven variable selection before applying machine learning versus logistic regression, item 3: Difference in handling of continuous variables before applying machine learning versus logistic regression, item 4: Different predictors considered for logistic regression and machine learning algorithms; and item 5: Corrections for imbalanced outcomes where used only for logistic regression or only for machine learning algorithms.

TABLE E2. Description and graphical representation of machine learning (ML) model included in the studies

Model	Description	Graphical representation
Artificial neural networks	The algorithm learns from processing many labeled examples (ie, data with answers) that are supplied during training and uses this answer key to learn what characteristics of the input are needed to construct the correct output. The basic unit of computation in a neural network is the neuron, often called a node or unit. It receives input from some other nodes, or from an external source and computes an output. Each input has an associated weight, which is assigned on the basis of its relative importance to other inputs. Nodes are arranged in layers. Neural network consists of 3 types of nodes that fall within 3 corresponding layers: input layers (these nodes take input data [ie, numbers and texts]); hidden layers (responsible for number crunching [ie, mathematical operation] to detect patterns data. There can be one or multiple hidden layers), and output layer (takes input from the hidden layer[s] to generate the desired output). ^{E19}	<p>The diagram, titled "ARTIFICIAL NEURAL NETWORKS", illustrates a feedforward neural network. It consists of three layers of nodes: a first layer with five blue nodes, a hidden layer with four orange nodes, and an output layer with one green node. Every node in one layer is connected to every node in the subsequent layer by a line representing a connection or weight.</p>
Decision trees	Each decision tree is composed of nodes and branches. The topmost decision node in a tree corresponds to the best predictor called root node, which splits the records into mutually exclusive classes. After the root node, there are internal nodes, which lead to other internal nodes or to ≥ 2 terminal leaf nodes. An item is classified according to which leaf node is reached. ^{E20}	<p>The diagram, titled "DECISION TREE", shows a hierarchical structure. At the top is a blue circle labeled "Root". Two arrows point down from the root to two orange circles labeled "Internal nodes". From each internal node, two arrows point down to four yellow circles labeled "Leaf nodes".</p>

(Continued)

TABLE E2. Continued

Model	Description	Graphical representation
<p>Gradient boosting machine</p>	<p>This model represents an ensemble of learning algorithms combining multiple weak learners to build a strong predictor tool to minimize the misclassification between the predicted and the observed values. The weak learners included in this model can take any functional forms (eg, neural networks, decision tree), but most commonly are tree-based learners.^{E21}</p>	 <p>The graph is titled "GRADIENT BOOSTING MACHINE". The vertical axis is labeled "Error" and the horizontal axis is labeled "Iterations". A blue curve starts at a high error point and decreases as it moves to the right, leveling off. Three boxes illustrate the process: the first box shows a single orange tree; the second box shows two trees (one orange, one green) with a plus sign between them; the third box shows three trees (two orange, one green) with plus signs between them and an ellipsis to the right, indicating further additions.</p>
<p>Support vector machine</p>	<p>This model works by creating a decision boundary between 2 classes and this enables classification prediction. Each item from the training test is plotted as a dot and the decision boundary, also called "separating hyperplane" is identified. This line is orientated so that it is as far as possible from the closest data points of each class. Then, the lines passing for the closest data to the hyperplane define the maximum-margin hyperplane. The separating hyperplane acts as the classifier and the testing data is classified according to on which side of the line it lands.^{E22}</p>	 <p>The graph is titled "SUPPORT VECTOR MACHINE". It shows a 2D coordinate system with blue dots in the upper-left region and red dots in the lower-right region. A solid black line, labeled "Separating hyperplane", runs diagonally between the two groups. Two dashed lines are drawn parallel to the solid line, one on each side, passing through the closest data points of each class. The space between these two dashed lines is labeled "Maximum margin".</p>

(Continued)

TABLE E2. Continued

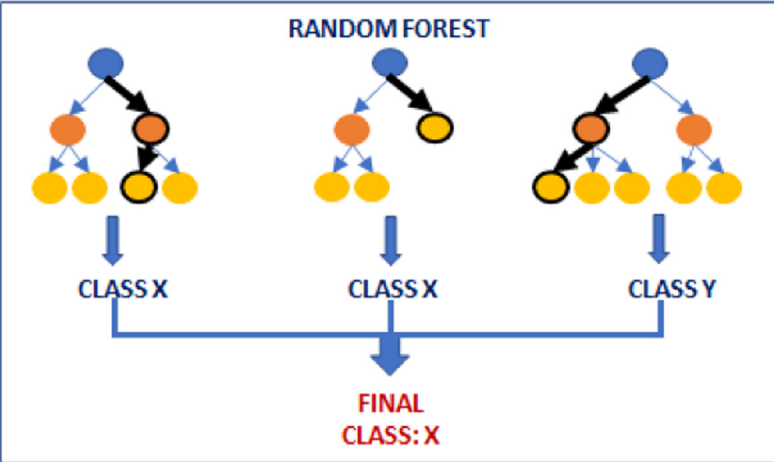
Model	Description	Graphical representation
Naïve Bayes	This model is based on the Bayes theorem. It is called naïve because it assumes each feature contributes independently to the probability of classification. The final prediction of the model is the a priori probability modified by the likelihood of each predictor. ^{E23}	
Random forest	This model represents an ensemble learning method that aggregates a large number of decision trees. When new input data are presented, each tree votes for a category and the forest prediction is based on the category that obtains the majority of the votes. ^{E24}	

TABLE E3. Description of calibration analysis

Study, y	Calibration method	Comment
Lippman, 1997 ^{E12}	χ^2 test	All machine learning and logistic regression models achieved a good calibration with the exception in high risk patients, which was less represented
Jamaati, 2015 ^{E15}	Hosmer-Lemeshow test of goodness of fit	Both machine learning and logistic regression models proved to have a good calibration
Celi, 2012 ^{E17}	Hosmer-Lemeshow test of goodness of fit	Both machine learning and logistic regression models proved to have a good calibration
Peng, 2008 ^{E4}	Hosmer-Lemeshow test of goodness of fit	Both machine learning and logistic regression models proved to have a good calibration

**000 Machine learning improves mortality risk prediction after cardiac surgery:
Systematic review and meta-analysis**

Umberto Benedetto, MD, PhD, Arnaldo Dimagli, MD, Shubhra Sinha, MD, Lucia Cocomello, MD, Ben Gibbison, MD, Massimo Caputo, MD, Tom Gaunt, PhD, Matt Lyon, MSc, Chris Holmes, PhD, and Gianni D. Angelini, MD, London and Oxford, United Kingdom

When compared to logistic regression models, machine learning appears able to provide better discrimination power in mortality prediction after cardiac surgery.