Benedetto, U., Dimagli, A., Sinha, S., Cocomello, L., Gibbison, B., Caputo, M., Gaunt, T., Lyon, M., Holmes, C., & Angelini, G. D. (2020). Machine learning improves mortality risk prediction after cardiac surgery: systematic review and meta-analysis. *Journal of Thoracic and Cardiovascular Surgery*. https://doi.org/10.1016/j.jtcvs.2020.07.105

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

Umberto Benedetto[1]*, Arnaldo Dimagli[1]*, Shubhra Sinha[1], Lucia Cocomello[1], Ben Gibbison[1], Massimo Caputo, [1] Tom Gaunt[2], Matt Lyon[2], Chris Holmes[3], Gianni D Angelini[1]

*first co-author

[1]Bristol Heart Institute, Translational Health Sciences, University of Bristol

[2]Population Health Sciences, University of Bristol

[3]Department of Statistics, University of Oxford

Corresponding author
Umberto Benedetto
Office Room 84
Level 7,
Bristol Royal Infirmary, Upper Maudlin Street BS2 8HW
Tel. +44 (0) 117 3428854
umberto.benedetto@bristol.ac.uk

Conflict of interest: none

The present study was supported by the National Institute for Health Research Bristol Biomedical Research Centre (NIHR Bristol BRC) and the British Heart Foundation.

Total word count: 3498

Glossary of abbreviations

AUC: Area Under the Receiver Operating Characteristic Curve

CABG: Coronary Artery Bypass Graft

LR: logistic regression

MHR: medical health record

ML: machine learning

SE: standard error

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

Perspective statement: 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 on 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 a significantly better prediction accuracy.

Abstract

Background: Interest on 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 (PRISMA) statement. Discrimination ability was assessed using c-statistic. Pooled c-statistics and its 95% credibility interval for ML models and LR were obtained were obtained using a Bayesian framework. Pooled estimates for ML models and LR were compared to inform on difference between the two 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%CrI 0.83-0.93 vs LR 0.81; 95%CrI 0.77-0.85; P=0.03). When individual ML algorithm were instead selected, we found a non-significant trend toward better prediction with each of ML algorithms. We found no evidence of publication bias (P=0.70).

Conclusions: The present findings suggest that when compared to LR, ML models provide better discrimination in mortality prediction after cardiac surgery . However, the magnitude and clinical impact of such an improvement remains uncertain.

Graphical abstract: Machine learning vs logistic regression model for mortality prediction in cardiac surgery.

Introduction

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 of surgical performance, counselling 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 1) inappropriate risk adverse practise that denies surgery to patients who would benefit from surgery, 2) falsely reassuring conclusions about surgeon and centre performance, 3) 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 modeller 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 modeller 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 Supplementary material.

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

Selection of studies

All abstracts were independently screened by two reviewers (A.D. and S.S); conflicts were resolved by a third reviewer (U.B). The full text of selected abstracts was independently assessed for eligibility by three reviewers (A.D., L.C., U.B.), and conflicts were resolved by consensus.

Inclusion and exclusion criteria

Studies were eligible if: 1) 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; 2) the article compared prediction models based on ML versus LR model. Studies were excluded if 1) a new modelling approach was introduced (i.e. dynamic modelling); 2) no validation was carried out; 3) the models made predictions for individual images or signals rather than participants; 4) models were developed based on high-dimensional data modalities; 5) the primary interest was assessing risk factors rather than prediction modelling; 6) they were reviews of the literature; 7) 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 (A.D. and U.B.) independently abstracted qualitative and quantitative data from selected studies. The list of extraction items was based on the CHARMS checklist (7) and the QUADAS risk of bias tool (8). 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 ROC 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 (i.e. validation sample was not used for model training) or from internal validation analysis (i.e. k-fold cross-validation or bootstrapping). The standard error (SE) of c-statistic (c) 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:

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

Based on the extracted data, we classified ML algorithms into five broad groups (11): 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 (i.e. EuroSCORE).

As proposed by Christodoulou et al. (6), from each article, we defined five signalling items to indicate potential bias (Supplementary Table 1): 1) unclear or biased validation of model performance: 2) difference in whether data-driven variable selection was performed (yes/no) before applying LR and ML algorithms; 3) difference in handling of continuous variables before applying LR and ML algorithms; 4) different predictors considered for LR and ML algorithms; 5) whether corrections for imbalanced outcomes where 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 five signalling items. If the answer was ‘‘unclear’’ or ‘‘yes’’ for at least one item, we assumed high risk of bias.

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 calculated meta-analytic estimates (details in Supplementary material). (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 et al (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 ($\geq$ 1000 vs < 1000 patients). Conventional risk scoring systems (e.g. 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 two-sided, with statistical significance set at $p < 0.05$.

Results

Our search identified 458 citations published between 6/1997 and 7/2018, of which 295 studies were excluded based on title or abstract (Supplementary Figure 1). Thirteen studies were excluded during full-text screening, and 15 studies (14,15,24–28,16–23) 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 vs LR in the cardiac surgery setting was published in 1997 (21) and it used the STS database for training and testing. However, most studies were published from 2014 and 2018. Study geographic areas were Europe (n=3)(14,16,27), Asia (n=7) (15,18–20,23,25,28), North America (n=2) (21,24), South America (n=2) (17,22) and New Zealand (n=1) (26). A total of 5 studies included unselected cardiac procedures (14,16,25,26,28), 7 studies focused on patients undergoing coronary artery bypass (CABG) only (18–24), the remaining three articles included only patients undergoing heart valve surgery for rheumatic heart valve disease (17), a combination of CABG and valve surgery (15) and Type A ascending aorta dissection surgery (27), respectively.

In 10 studies, data were retrospectively obtained from medical health records (14–16,18–20,25,27,28) or international surgical databases (EuroSCORE or STS) (14,21). One study used

data from the Cardiac Care Network of Ontario (24) and another the Registry of Cardiac Surgery Patients in Dunedin Hospital (26). Data were prospectively collected only in three studies (17,22,23). Sample size ranged from 165 to 80606 patients and operative mortality from 3.0% to 25.5%. ML models developed were artificial neural network (n=12) (14,17,27,28,18–22,24–26), decision tree analysis (n=2) (15,25), random forest (n=2) (16,17), support vector machine (n=3) (16,17,23), naïve Bayes (n=3) (16,17,26), gradient boost machine (n=1) (16) and ensemble of models (n=2) (16,21). ML models are described in Supplementary Table 2. With the exception of one study (18), 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) (16,17). C-statistic was the performance measure used in 13 studies (14,15,24,26,28,16–23), sensitivity and specificity (25) and Gini coefficient (27) were used in the remaining two studies, and c-statistic was derived using conversion equations (details in Supplementary material).

For ML models, the c-statistic ranged from 0.736 to 0.982 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 (15,16,18,21,24,26) and sample splitting only in 5 (19,20,22,25,28) studies. Other validation methods adopted were k-fold cross validation only (n=1) (17), combination of sample splitting and k-fold cross validation and external validation (n=1) (14) and external validation only (n=1) (27). In one study, the validation method was not reported (23). Calibration was reported only in 4 studies (21,23,26,28)(details regarding calibration assessment are presented in Supplementary Table 3).

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)

(15,17,19,20), single imputation (n=2) (18,21) and a combination of complete case for mandatory variables and single imputation for other variables (n=1) (14). Statistical software used for ML modelling was reported in all but one study.

Methodological Quality

Ten (67%) studies were at low risk of bias (14–17,19–21,24–26), while the remaining 5 (33%) were classified as at high risk of bias (Supplementary Table 1). In the study by Chong et al (18), although the original number of input variables included in ML and LR models were 21, it was unclear why the final number of input variable predictors in the ML model was 18. In the study by Mendes et al (22), 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 et al (23) it was unclear to assess any validation methodology used. Peng et al (28) ran a data-driven variable selection for LR model but not for ML and similarly, the study by Macrina et al (27).

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 0.77-0.85; P=0.03; 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=0.15), decision trees/random forest (0.89; 95%CrI 0.76-0.98 vs LR 0.80; 95%CrI 0.63-0.90; P=0.30), support vector machine (0.92; 95%CrI 0.75-1.00 vs LR 0.82; 95%CrI 0.65-0.96; P=0.27), and naïve Bayes (0.81; 95%CrI 0.69-0.96 vs LR 0.78; 95%CrI

0.68-0.91; P=0.8) 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=0.15) 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=0.10) (Supplementary Figure 1). Furthermore, when compared to LR, ML models achieved a better discrimination accuracy in studies published from 2010 onwards (ML 0.9; 95%CrI 0.84-0.97 vs LR 0.81; 95%CrI 0.74-0.88; P=0.02) 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=0.5). We found a trend towards 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=0.70; Supplementary Figure 2). 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 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 Supplementary Table 3.

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 5, Graphical abstract). 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 towards 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 the European System for Cardiac Operative Risk Evaluation version II (EuroSCORE II) (2) and the STS-PROM 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 that these models require modeller input as 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 co-existing comorbidities and surgical procedures. While simplified models are associated with lower variance, they may also result in mis-calibrated 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 non-linearity and the interactions among features without the need for the modeller to manually specify all interactions, as needed with LR. Moreover, compared to traditional statistical methods, ML algorithms can handle missing data more efficiently as 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 to 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 one model that works best for every problem or every dataset. The assumptions of a good model for one problem may not hold for another problem, so it is common in ML to try multiple models and find one 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 impact of improvement using ML remains uncertain. ML modelling 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 dataset which 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 upon 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 one class is significantly lower than those belonging to the other classes. This is because ML algorithms are designed to maximize accuracy (i.e. 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 under-sampling or algorithm-centered approaches which modify the algorithm to favour its prediction towards the less represented class (37). The problem of class imbalance may be particularly relevant when ML is applied for prediction in cardiac surgery as 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 (i.e. EuroSCORE or STS score) (2,3). These databases contain only a restricted number of prespecified variables limiting the capability of ML which 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, modelling 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

In the present study, 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 suffered from poor methodology and reporting (18,22,23,27,28). 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) (21,23,26,28) and only one 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.

Conclusion

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

Acknowledgment: We would like to thank Dr. Giovanni Morlino for his support with statistical analysis.

References

1. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). Eur J Cardiothorac Surg. 1999 Jul;16(1):9–13.
2. Nashef SAM, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. EuroSCORE II. Eur J Cardiothorac Surg. 2012 Apr;41(4):734–5.
3. 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 Jul;88(1 Suppl):S2-22.
4. 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 Jun;98(6):363–9.
5. 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 Aug;102(2):573–9.
6. 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 [Internet]. 2019;110:12–22. Available from: http://www.sciencedirect.com/science/article/pii/S0895435618310813
7. 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 [Internet]. 2014 Oct 14;11(10):e1001744. Available from: https://doi.org/10.1371/journal.pmed.1001744
8. 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 Nov;3:25.

9. 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 [Internet]. 2010 Jan;21(1):128–38. Available from: https://pubmed.ncbi.nlm.nih.gov/20010215

10. NCSS Statistical Software. Confidence Intervals for the Area Under an ROC Curve. Chapter 26. Available from: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/PASS/Confidence_Intervals_for_the_Area_Under_an_ROC_Curve.pdf

11. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol. 2019 Mar;19(1):64.

12. 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 Sep;28(9):2768–86.

13. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983 Sep;148(3):839–43.

14. 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 Jul;132(1):12–9.

15. 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 Oct;148(4):1291-1298.e1.

16. 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(1):e0169772.

17. Mejia OA V, 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(7):e0199277.

18. 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 proceedings AMIA Symp. 2003;160–4.

19. 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 [Internet]. 2016 Jun 1;33(3):230–8. Available from: https://doi.org/10.1111/exsy.12145

20. 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 Oct;38(10):102.

21. Lippmann RP, Shahian DM. Coronary artery bypass risk prediction using neural networks. Ann Thorac Surg. 1997 Jun;63(6):1635–43.

22. 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 Aug;11(4):756–63.

23. 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 Oct;19(10):576–9.

24. Tu J V, 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(2):229–35.

25. Rahman HAA, Wah YB, Khairudin Z, Abdullah NN. Comparison of predictive models to predict survival of cardiac surgery patients. In: 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE). 2012. p. 1–5.

26. 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 Sep;2(4):138–48.

27. 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 Jul;3:81–95.

28. Peng S-Y, Peng S-K. Predicting adverse outcomes of cardiac surgery with the application of artificial neural networks. Anaesthesia. 2008 Jul;63(7):705–13.

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 cardio-thoracic Surg Off J Eur Assoc Cardio-thoracic Surg. 2016 Sep;50(3):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 [Internet]. 2018 Oct 3;7(10):322. Available from: https://pubmed.ncbi.nlm.nih.gov/30282956

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 [Internet]. 2018 Dec 1;6(12):905–14. Available from: https://doi.org/10.1016/S2213-2600(18)30300-X

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 [Internet]. Vol. 20, The heart surgery forum. Department of Surgery, Vanderbilt University, Nashville, TN, USA.; 2017. p. E007-E014. Available from: http://europepmc.org/abstract/MED/28263144

33. Gupta A, Lam MS. Estimating Missing Values Using Neural Networks. J Oper Res Soc [Internet]. 1996 May 6;47(2):229–38. Available from: http://www.jstor.org/stable/2584344

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 Jan;28(1):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 [Internet]. 2014;14(1):137. Available from: https://doi.org/10.1186/1471-2288-14-137

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(1):e84217.

37. Kaur H, Pannu HS, Malhi AK. A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. ACM Comput Surv [Internet]. 2019 Aug;52(4). Available from: https://doi.org/10.1145/3343440

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 Jun;36(14):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(5):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 Dec;74(6):796–804.

Figure legend

Figure 1. Forest plot comparing discrimination accuracy (i.e. c-statistic) in mortality prediction by selecting machine learning (ML) models (top) with best performance vs. logistic regression (LR) (bottom). ANN: artificial neural networks, DT: decision tree, RF: random forests, SVM: support vector machine.

Figure 2. Forest plot comparing discrimination accuracy (i.e. c-statistic) in mortality prediction by selecting machine learning (ML) models based on the same algorithm vs. logistic regression (LR). ANN: artificial neural networks, DT: decision tree, RF: random forests, SVM: support vector machine.

Figure 3. Graphical abstract. Machine Learning (ML) algorithms (i.e. random forest, neural networks, support vector machine, etc.) 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. Machine learning models achieved a better prediction than logistic regression**.**

Supplemental Figure 1. PRISMA flow chart of search strategy.

Supplementary Figure 2. Forest plot comparing discrimination accuracy (i.e. c-statistic) in mortality prediction by selecting machine learning (ML) models (top) with best performance vs. logistic regression (LR) in studies with high (left) and low risk of bias (right).

Supplementary Figure 3. Funnel plot for assessment of small-study effect, obtained by plotting the c-statistics and the standard error for each study included

Central picture legend: Pooled c-statistic for mortality prediction by the best machine learning (ML) models.

522 **Table 1. Study characteristics**

Author, year (ref.)	Geographic area	Population	Age	Male sex, %	Source of data	Sample size	Operative mortality %
Nilsson, 2006 (14)	Europe	Unselected	62.6 ± 10.7	72	Retrospective (EuroSCORE database and MHR)	18362	4.9
Ghavidel, 2014 (15)	Asia	CABG and valve surgery	45-60 y 58.1 % 61-75 y 39.5% >76 y 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 (17)	South America	Rheumatic valve disease	51.2 ± 14.9 for survivors 54.4 ± 17.4 for non survivors	NR	Prospective	2919	3.5
Chong, 2003 (18)	Asia	CABG	64.0 (10.4) for training set 63.3 (9.7) for testing set	70	Retrospective (MHR)	563	7.5
Nouei, 2016 (19)	Asia	CABG	58.24 ±9.74 for survivors 62.07 ±9.47 for non survivors	70	Retrospective (MHR)	824	3.5
Nouei, 2014 (20)	Asia	CABG	58.62±10.18 for survivors 61.82±10.72 for non survivors	NR	Retrospective (MHR)	1811	3.3
Lippman, 1997 (21)	North America	CABG	NR	73.4% survivors 62.3% non-survivors	Retrospective (STS database)	80606	3.4
Mendes, 2015 (22)	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 (23)	Asia	CABG	57	51	Prospective	2220	12.2
Tu, 1998 (24)	North America	CABG	NR	NR	Retrospective (Cardiac Care Network of Ontario)	15608	3.0

Rahman, 2012 (25)	Asia	Unselected	18 – 40 y 9.4% 40 – 60 y 53.2% >60 y 37.3%	77	Retrospective (MHR)	1209	17.3
Celi, 2012 (26)	Oceania	Unselected	>80 y 100%	NR	Retrospective (Registry of Cardiac Surgery Patients in Dunedin Hospital)	165	7.4
Macrina, 2009 (27)	Europe	Acute Aortic dissection	61±12 for survivors 66±10 for non survivors	63 % for survivors 66% for non survivors	Retrospective (MHR)	208	25.5
Peng, 2008 (28)	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

523 CABG: Coronary Artery Bypass Graft; MHR: Medical Health Record; NR: not reported;
524 STS: Society of Thoracic Surgeons.

525

526 **Table 2. Study methodological characteristics**

Author, year [ref.]	Model tested	No. Predictors	Handling of missing data	Type of validation	Split ratio Training: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 (17)	RF ANN SVM NB LR	10	Missing negligible, imputation not performed	K-fold cross validation	-	NR	R package caret

	EuroSCORE II						
Chong, 2003 (18)	ANN LR	18	Coded as "missing" for categorial 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 (19)	ANN LR	40	Missing excluded for the analysis	Sample splitting	70:30	NR	NR
Nouei, 2014 (20)	ANN LR	40	Missing excluded for the analysis	Sample splitting	70:30	NR	MATLAB
Lippman, 1997 (21)	ANN Ensemble LR	36	Imputation for variables (statistical mode or mean substitution)	Sample splitting and k-fold cross validation	50:50	Performed using X^2 for comparison	LNKnet software
Mendes, 2015 (22)	ANN LR	12	NR	Sample splitting	80:20	NR	Accord. NET Framework
Jamaati, 2015 (23)	SVM LR	17	NR	NR	-	Hosmer–Lemeshow goodness-of-fit statistic	SPSS
Tu, 1998 (24)	ANN LR	17	NR	Sample splitting and k-fold cross validation	65:35	NR	Stata
Rahman, 2012 (25)	ANN DT LR	12	NR	Sample Splitting	NR	NR	SPSS PASW Modeler 13
Celi, 2012 (26)	ANN BN LR	6	NR	Sample splitting and k-fold cross validation	70:30	Hosmer–Lemeshow	Weka and R

						goodness-of-fit statistic	
Macrina, 2009 (27)	ANN LR	22	NR	External validation	-	NR	NCSS and MedCalc)
Peng, 2008 (28)	ANN LR	16	NR	Sample splitting	70:30	Hosmer–Lemeshow goodness-of-fit statistic	STATISTICA from StatSoft Inc.

ANN: Artificial Neural Networks; EEFDT: Entropy Error Fuzzy Decision Tree; EECDT: Entropy Error Crisp Decision Tree; GBM: Gradient Boosting Machine; LR: Logistic Regression; ML: Machine Learning; NB: Naive Bayesian; NR: not reported; RF: Random Forest; SVM: Support Vector Machine

530 **Table 3. Model performance characteristics**

Study	Year	ML model	Testing all	Testing deaths	c-statistic	SE c-statistic
Nilsson	2006	ANN	1246	112	0.81	0.03
		LR	1246	112	0.80	0.03
		EuroSCORE	1246	112	0.79	0.03
Ghavidel	2014	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	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	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	ANN	140	11	0.89	0.07
		LR	140	11	0.81	0.08
Nouei (ii)	2016	ANN	247	8	0.82	0.09
		LR	247	8	0.62	0.11
Nouei (i)	2014	ANN	543	20	0.91	0.05
		LR	543	20	0.72	0.07
Lippman	1997	ANN	40126	1374	0.76	0.01
		Naïve Bayes	40126	1374	0.75	0.01
		Ensemble	40126	1374	0.76	0.01
		LR	40126	1374	0.76	0.01
Mendes	2015	ANN	262	22	0.85	0.05
		LR	262	22	0.86	0.05
Tu	1998	ANN	5517	173	0.78	0.02
		LR	5517	173	0.77	0.02
Jamaati	2015	SVM	2220	270	0.98	0.01
		LR	2220	270	0.84	0.02
Peng	2008	ANN	315	37	0.87	0.04
		LR	315	37	0.85	0.04
Macrina	2009	ANN	87	20	0.93	0.05
		LR	87	20	0.88	0.06
Celi	2012	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	ANN	1209	209	0.91	0.01
		DT/RF	1209	209	0.91	0.01
		LR	1209	209	0.89	0.02

531 * derived from c-statistic as reported in original articles; ǂ derived from Gini coefficient as
532 reported in original article; ǂ derived from sensitivity and specificity as reported in original
533 article; ANN: artificial neural networks; DT/RF: decision tree/random forests; GBM:
534 gradient boosting machine; ML: Machine Learning; SE (c-statistic): standard error of the c-
535 statistic; SVM: support vector machine.

536 **Table 4. Meta-analytic estimates**

	n studies	**ML** Pooled **c**-statistic 95% credible interval	**LR** Pooled **c**-statistic 95% credible Interval	Net Benefit	P-val
Best ML model (Overall)	N=15	0.88[0.83-0.93]	0.81[0.77-0.85]	+7%	0.03
Artificial Neural Network	N=12	0.86[0.81-0.91]	0.81[0.76-0.86]	+5%	0.15
Decision Trees/Random Forest	N=4	0.89[0.76-0.98]	0.80[0.63-0.90]	+9%	0.30
Support Vector Machine	N=3	0.92[0.75-1.00]	0.82[0.65-0.96]	+10%	0.27
Naïve Bayes	N=4	0.81[0.69-0.96]	0.78[0.68-0.91]	+3%	0.8
Other	N=2	0.77[0.70-0.87]	0.76[0.54-0.96]	+1%	0.92
Best ML model-low risk of bias	N=10	0.85[0.79-0.91]	0.79[0.73-0.85]	+6%	0.15
Best ML model-high risk of bias	N=5	0.92[0.82-0.98]	0.84[0.79-0.90]	+8%	0.10
Best ML model >=2010	N=9	0.91[0.84-0.97]	0.81[0.74-0.88]	+10%	0.02
Best ML model <2010	N=6	0.81[0.75-0.89]	0.78[0.74-0.85]	+3%	0.5
Best ML model- EV	N=2	0.85 [0.64-0.99]	0.82 [0.61-0.99]	+3%	0.8
Best ML model- IV	N=13	0.88 [0.82-0.94]	0.80 [0.76-0.85]	+8%	0.04
Best ML model- ≥1000 patients	N=9	0.89 [0.79-0.95]	0.80 [0.75-0.86]	+9%	0.07
Best ML model- < 1000 patients	N=6	0.88 [0.82-0.94]	0.82[0.70-0.90]	+6%	0.13

537 LR: logistic regression; ML: machine learning; EV: external validation; IV: internal validation.