European Journal of Cardio-Thoracic Surgery 58 (2020) 1130–1136
doi:10.1093/ejcts/ezaa229 Advance Access publication 18 August 2020

ORIGINAL ARTICLE

Cite this article as: Benedetto U, Sinha S, Lyon M, Dimagli A, Gaunt TR, Angelini G *et al.* Can machine learning improve mortality prediction following cardiac surgery? Eur J Cardiothorac Surg 2020;58:1130–6.

Can machine learning improve mortality prediction following cardiac surgery?

Umberto Benedetto[a,b,*], Shubhra Sinha[a], Matt Lyon [b,c,d], Arnaldo Dimagli [a], Tom R. Gaunt[b,c,d], Gianni Angelini[a,b] and Jonathan Sterne [b,c,d]

[a] Translational Health Sciences, Bristol Heart Institute, University of Bristol, Bristol, UK
[b] NIHR Bristol Biomedical Research Centre, University of Bristol, University Hospitals Bristol NHS Foundation Trust, Bristol, UK
[c] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
[d] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

* Corresponding author. Office Room 84 Level 7, Bristol Royal Infirmary, Upper Maudlin Street, Bristol BS2 8HW, UK. Tel. +44-117-3428854; e-mail: umberto.benedetto@bristol.ac.uk (U. Benedetto).

Received 25 March 2020; received in revised form 20 May 2020; accepted 26 May 2020

Key question

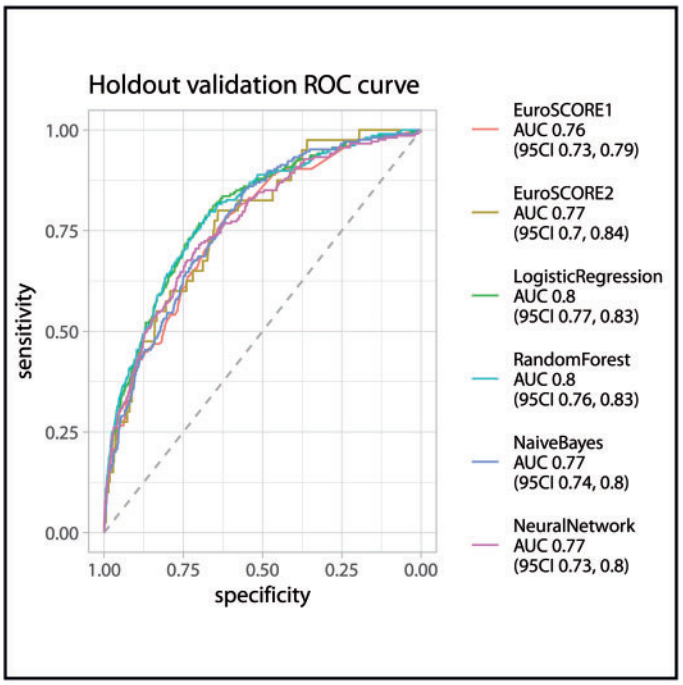
Do modern machine learning models improve the prediction of in-hospital mortality after cardiac surgery?

Key finding(s)

Machine learning models performed similarly to logistic regression models.

Take-home message

Prediction of in-hospital mortality is not improved by machine learning relative to traditional methods based on logistic regression.



Abstract

OBJECTIVES: Interest in the clinical usefulness of machine learning for risk prediction has bloomed recently. Cardiac surgery patients are at high risk of complications and therefore presurgical risk assessment is of crucial relevance. We aimed to compare the performance of machine learning algorithms over traditional logistic regression (LR) model to predict in-hospital mortality following cardiac surgery.

METHODS: A single-centre data set of prospectively collected information from patients undergoing adult cardiac surgery from 1996 to 2017 was split into 70% training set and 30% testing set. Prediction models were developed using neural network, random forest, naive Bayes and retrained LR based on features included in the EuroSCORE. Discrimination was assessed using area under the receiver operating characteristic curve, and calibration analysis was undertaken using the calibration belt method. Model calibration drift was assessed by comparing Goodness of fit χ^2 statistics observed in 2 equal bins from the testing sample ordered by procedure date.

RESULTS: A total of 28 761 cardiac procedures were performed during the study period. The in-hospital mortality rate was 2.7%. Retrained LR [area under the receiver operating characteristic curve 0.80; 95% confidence interval (CI) 0.77–0.83] and random forest model (0.80; 95% CI 0.76–0.83) showed the best discrimination. All models showed significant miscalibration. Retrained LR proved to have the weakest calibration drift.

CONCLUSIONS: Our findings do not support the hypothesis that machine learning methods provide advantage over LR model in predicting operative mortality after cardiac surgery.

Keywords: Machine learning • Mortality prediction • Neural network • Random forest • Naive Bayes

ABBREVIATIONS

AUC	Area under the receiver operating characteristic curve
CI	Confidence interval
LR	Logistic regression
ML	Machine learning
STS	Society of Thoracic Surgeons

INTRODUCTION

Preoperative assessment of surgical risk is of crucial importance in cardiac surgery due to the high risk of intraoperative and postoperative complications. Risk models can help health professionals to advise patients during the decision-making process, as well as in monitoring surgical performance and cost–benefit analyses.

Several risk stratification models have been developed to predict in-hospital mortality following cardiac surgery, for example as the European System for Cardiac Operative Risk Evaluation, EuroSCORE [1, 2] and the North American Society of Thoracic Surgeons (STS) score [3]. However, a main limitation of these scores is overestimation of risk in high-risk patient subgroups [4, 5]. This can potentially translate into risk-averse practice, falsely reassuring conclusions about surgeon and centre performance, and impaired decision-making.

Current risk scoring systems are based on logistic regression (LR). Development of LR models requires input from the modeller to address complex interaction among features and non-linear relationships of features with the outcome. For instance, the contribution of advanced age to mortality risk may not be constant across the spectrum of comorbidities. If features' interactions are overlooked in an LR model, its prediction ability will be negatively affected. In contrast, machine learning (ML) algorithms require less input from the modeller and interactions among features and non-linear relationships can be learnt automatically from the data [6]. However, the extra flexibility of ML algorithm requires larger sample to train the model.

Despite research on the utility of ML methods to improve prediction in health care has exponentially increased, ML methods have not been widely adopted in the clinical practice. Moreover, recent reports have challenged the additional value of ML in the development of clinical prediction models in a variety of clinical conditions [6].

The objective of this study was to compare ML algorithms with LR model in the prediction of in-hospital mortality after cardiac surgery, based on the set of features included in the EuroSCORE [1].

METHODS

The present study was approved by Health Research Authority and Health and Care Research Wales. Data were obtained from the National Adult Cardiac Surgery Audit (NACSA) data set, which prospectively collects clinical information for all major heart operations carried out in the UK. In the present analysis, we used a subset of patients who underwent cardiac surgery at University Hospitals Bristol NHS Trust between 1 April 1996 and 30 December 2017.

Missing or conflicting data for in-hospital mortality were obtained via record linkage to the Office for National Statistics census database. For records where data required to calculate a EuroSCORE variable were missing, it was assumed that the risk factor was not present (equal to the reference level). Missing patient age at the time of surgery was imputed as the median patient age for the corresponding financial year.

Statistical analysis and models

The primary end point was in-hospital mortality following cardiac surgery. Numerical variables were summarized as mean and standard deviation or median and interquartile range and compared using *t*-tests or Mann–Whitney tests. Categorical variables were tabulated as frequencies and percentages and compared using χ^2 test.

Procedures were ordered chronologically, the first 70% of records (1 April 1996–27 September 2011) were used for training and hyperparameter selection through five-fold cross-validation. Final model performance was evaluated using the remaining 30% (27 September 2011–30 December 2017). All prediction models were developed using the 17 features included in the original EuroSCORE [1], which include information prior to surgery on a range of patient, cardiac and operative factors. The features are age, gender, chronic obstructive pulmonary disease, extracardiac arteriopathy, neurological dysfunction, previous cardiac surgery, creatinine >200 μmol/l, active endocarditis, critical preoperative state, unstable angina, left ventricular function, recent myocardial infarction, pulmonary hypertension, emergency surgery, combined surgery other than coronary artery bypass graft, surgery on thoracic aorta and postinfarct septal rupture.

We fitted an LR (retrained LR) model to the EuroSCORE risk factors. We used the following ML approaches:

- Neural network is a computational learning system that uses a network of functions to understand and translate a data input of one form into a desired output. ML algorithms including neural networks generally do not need to be programmed with specific rules that define what to expect from the input. The neural network algorithm instead learns from processing many labelled examples (i.e. 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. Once a sufficient number of examples have been processed, the neural network can begin to process new, unseen inputs and successfully return accurate results. The more examples and variety of inputs provided during training, the more accurate the results typically become because the algorithm learns with experience. 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 (w), which is assigned on the basis of its relative importance to other inputs. The node applies a function f to the weighted sum of its inputs [i.e. f (w1 + w2 + w3 ...], which can introduce non-linearity into the output of a neuron (depending on the function chosen). Nodes are arranged in layers. Nodes from adjacent layers have connections or edges between them. All these connections have weights associated with them. Neural network consists of 3 types of nodes, which fall within 3 corresponding layers: (i) input layers: these nodes take input data (i.e. numbers, texts, etc.); (ii) hidden layers: are responsible for number crunching, i.e. mathematical operation, to detect patterns data. There can be one or multiple hidden layers; (iii) output layer: takes input from the hidden layer(s) to generate the desired output [7, 8]. In common with many ML approaches, neural networks were not specifically designed for time-related events, but as research rapidly moved forward new methods have been introduced for this purpose [9]. In our model, the number of hidden layers and nodes per hidden layer was configured manually in response to model discrimination [area under the receiver operating characteristic curve (AUC)] evaluated with cross-validation. The final model configuration used for evaluation was: input layer $n = 18$ nodes, hidden layer one $n = 90$ nodes, hidden layer two $n = 36$ nodes and output layer one node.
- Random forest represents an ensemble of several decision trees. Decision trees build classification or regression models in the form of a tree structure. This approach breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has 2 or more branches, while a leaf node is a terminal node that represents a classification or decision. 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 that lead to other internal nodes or to 2 or more terminal leaf nodes. An item is classified according to which leaf node is reached. Each item can be trained using resampling methods (i.e. bootstrapping) [10, 11]. Random forest has several parameters that have to be set by the user, e.g. number of trees in the forest (estimator), maximum number of levels (depth) in each decision tree, minimum number of data points placed in a node before the node is split and minimum samples of leaf. When new data are presented, each tree of the random forest votes for a class and the final prediction is based on the class receiving the majority of the votes. In our model, we manually tuned parameters in response to model discrimination (AUC) evaluated with cross-validation (estimators $n = 700$, maximum depth $n = 10$, minimum samples split $n = 5$, minimum samples leaf $n = 20$).
- Naive Bayes is based on the Byes theorem. It is called 'naive' 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 [12]. In our model, we used default parameters.

Full model configurations and discrimination are provided in Supplementary Material, Table S1. Models were developed and evaluated using scikit-learn v0.21.2 and TensorFlow v1.14.0 through Anaconda Python 3 v2019.07.

Discrimination was assessed by calculating model AUC with its relative 95% confidence interval (CI) using bootstrapping (2000 repetitions) (pROC R-package v1.15.3). The assessment of calibration, i.e. the model's ability to provide reliable predictions, is crucial to test risk models. Statistical techniques, such as the Hosmer–Lemeshow statistics and the Cox calibration test, are all non-informative with respect to calibration across risk classes. To better characterize the calibration of new models, we used the calibration belt model [13]. In this new approach, the relation between the logits of the probability predicted by a model and of the event rates observed in a sample is represented by a polynomial function, whose coefficients are fitted and its degree is fixed by a series of likelihood-ratio tests. This method also enables CIs to be computed for the curve, which can be plotted [13] (R-package givitiR v1.3) (R-package ResourceSelection v0.3.5). The calibration belt produces a trend with the 95% CI containing the line of equality. Open-source code is available from: https://github.com/MRCIEU/cvd-mortality-ml.

We also reported the performance of the original EuroSCORE I and EuroSCORE II for completeness. We were able to calculate the EuroSCORE II [2] only in 1889 (21.9%) patients for whom exact values of serum creatinine were available.

RESULTS

Participants

A total of 28 761 cardiac procedures were included in the final data set (Supplementary Material, Fig. S1). Patients younger than 18 years at the time of surgery were excluded ($n = 41$) to avoid the inclusion of congenital abnormalities. The outcome and full set of features were available for all records after imputation. The

Table 1: Distribution of features included in the EuroSCORE stratified for in-hospital mortality in patients who underwent adult cardiac surgery from 1996 to 2017

	Alive (N = 27 934)	Dead (N = 786)	P-value
Age (years), mean (SD)	65.29 (12.10)	69.38 (11.85)	<0.001
Female gender, *n* (%)	7149 (25.59)	286 (36.39)	<0.001
Serum creatinine ≥200 μmol/l, *n* (%)	332 (1.19)	56 (7.12)	<0.001
Extracardiac arteriopathy, *n* (%)	2346 (8.40)	131 (16.67)	<0.001
Pulmonary disease, *n* (%)	3370 (12.06)	146 (18.58)	<0.001
Neurological dysfunction, *n* (%)	593 (2.12)	27 (3.44)	0.018
Previous cardiac surgery, *n* (%)	1734 (6.21)	128 (16.28)	<0.001
Recent myocardial infarct, *n* (%)	6665 (23.86)	226 (28.75)	0.002
LVEF 30–50%, *n* (%)	5539 (19.83)	226 (28.75)	<0.001
LVEF <30%, *n* (%)	1391 (4.98)	129 (16.41)	<0.001
Systolic pulmonary pressure >60 mmHg, *n* (%)	836 (2.99)	28 (3.56)	0.414
Active endocarditis, *n* (%)	285 (1.02)	23 (2.93)	<0.001
Unstable angina, *n* (%)	2554 (9.14)	155 (19.72)	<0.001
Emergency operation, *n* (%)	884 (3.16)	208 (26.46)	<0.001
Critical preoperative state, *n* (%)	417 (1.49)	128 (16.28)	<0.001
Ventricular septal rupture, *n* (%)	53 (0.19)	32 (4.07)	<0.001
Other than isolated coronary surgery, *n* (%)	10 461 (37.45)	464 (59.03)	<0.001
Thoracic aortic surgery, *n* (%)	1363 (4.88)	148 (18.83)	<0.001

LVEF: left-ventricle ejection fraction; SD: standard deviation.

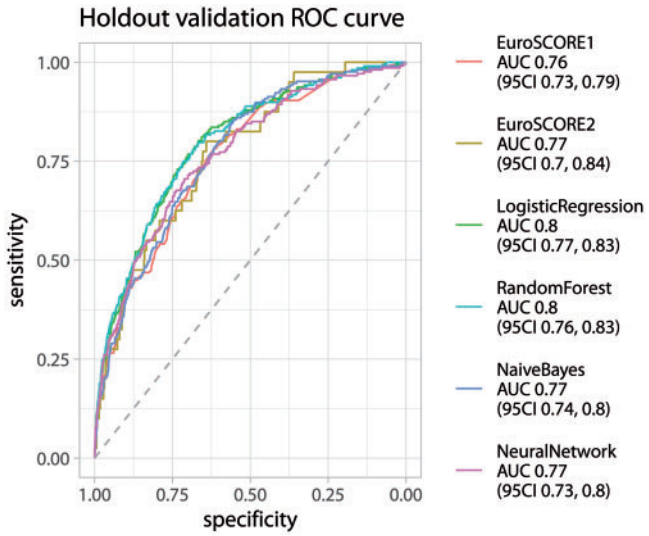


Figure 1: ROC curve of EuroSCORE I and II, logistic regression and machine learning classifiers: neural network, naive Bayes and random forest using EuroSCORE I features. The axes are true positive rate against 1−false positive rate. The area under the curve provides a measure of discrimination accuracy. The dashed line represents no classification discrimination ability. AUC: area under the receiver operating characteristic; CI: confidence interval; ROC: receiver operating characteristic.

overall percentage of missing data in the EuroSCORE variables was very low (1.7%) and records of age were missing in 86 patients. Patient characteristics are presented in Table 1. All features included in EuroSCORE I were robustly associated with the outcome in univariable analyses, except of elevated systolic pulmonary pressure. In-hospital mortality rate was 2.7% (n = 786).

Model discrimination

Results of model selection and hyperparameter tuning using the training set are reported in Supplementary Material, Table S1. Discrimination ability of models selected in the testing set is presented in Fig. 1. Retrained LR showed good discrimination (AUC 0.80; 95% CI 0.77–0.83). Among the ML classifiers, random forest showed the best discrimination ability (0.80; 95% CI 0.76–0.83), which was comparable to retrained LR model. Neural network and naive Bayes AUC were 0.77 (95% CI 0.73–0.80) and 0.77 (95% CI 0.74–0.80), respectively. Original EuroSCORE I and II AUC were 0.76 (95% CI 0.73–0.79) and 0.77 (0.70–0.84), respectively.

Probability calibration

Retrained LR had strong evidence against the null hypothesis of well-calibrated probabilities when applied to our data (Fig. 2A). Among the contemporary classifiers, neural network and random forest also showed poor calibration (Fig. 2B and C), although the latter produced probabilities that did not depart far from the line of equality. Naive Bayes produced probabilities that suggest very poor calibration. EuroSCORE I showed poor calibration (Fig. 3A) while EuroSCORE II was well calibrated, although the sample size and event number were smaller increasing the possibility of a type II error (Fig. 3B). To evaluate calibration drift in the retrained LR and ML models, the test data set was divided into 2 equal bins ordered by procedure date with approximately equal number of events (n = 102 vs n = 105). Hosmer–Lemeshow goodness of fit χ^2 statistics was calculated for the first and second quantiles (Table 2). Retrained LR had the weakest change in test statistic between quantiles (+15.9%) and therefore weakest calibration drift. Random forest had the second smallest effect (+21.2%). EuroSCORE II had too few events and could not be reliably evaluated.

DISCUSSION

The main finding of the present study is that when trained on the same set of variables, ML algorithms do not improve prediction

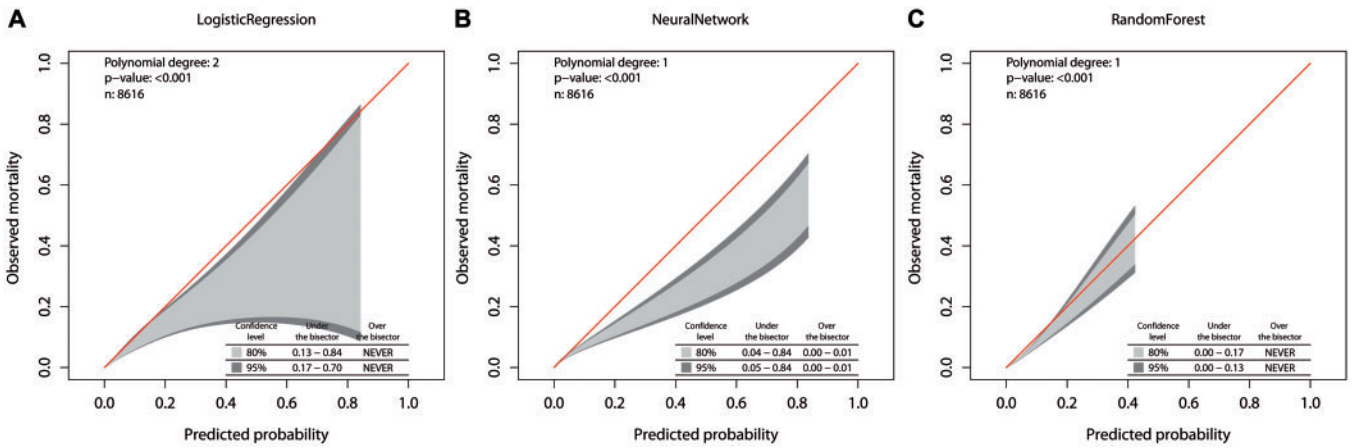


Figure 2: External probability calibration of logistic regression (**A**), neural network (**B**) and random forest (**C**) using the calibration belt method. The method regresses true mortality on classifier probability of mortality (via logit function) using polynomial logistic regression. All models showed significant miscalibration ($P < 0.001$).

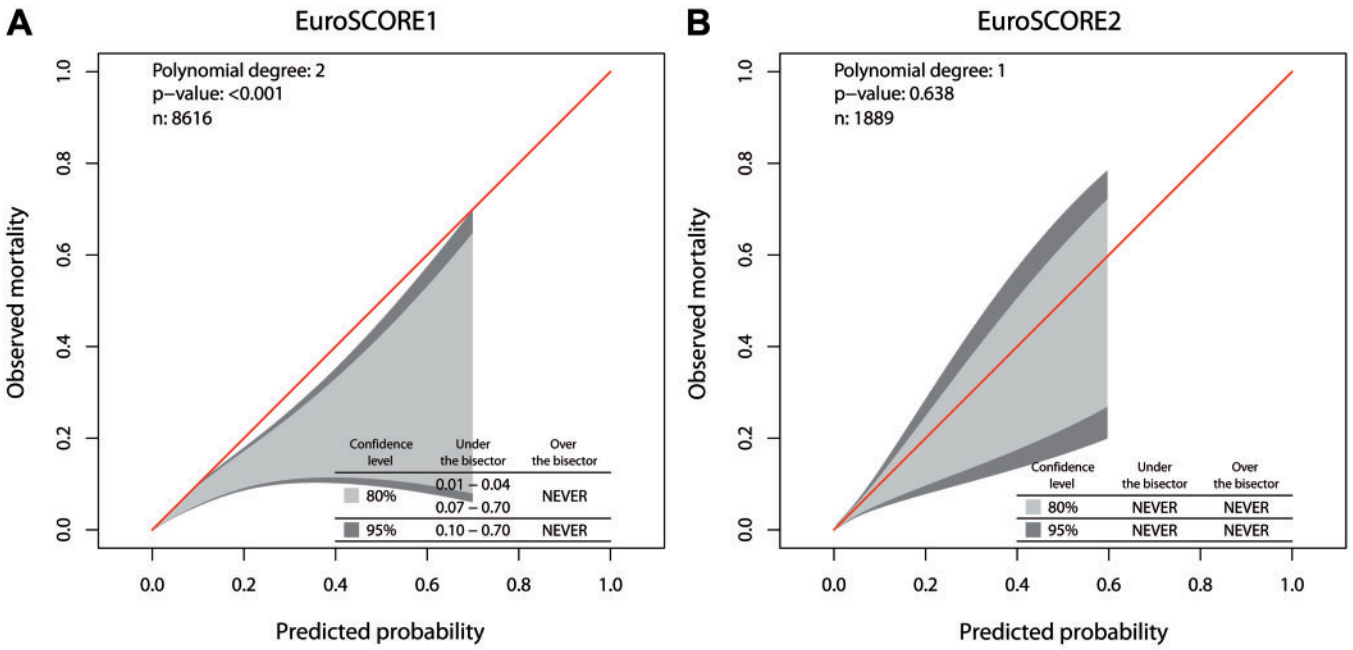


Figure 3: External probability calibration of EuroSCORE I (**A**) and EuroSCORE II (**B**) using the calibration belt method. The method regresses true mortality on classifier probability of mortality (via logit function) using polynomial logistic regression. EuroSCORE I ($P < 0.001$) but not EuroSCORE II ($P = 0.64$) showed significant model miscalibration.

Table 2: Evaluation of calibration drift

Model	χ^2 (G1)	χ^2 (G2)	Change (%)
Logistic regression (retrained)	12.45	14.81	15.9
Naive Bayes	1242.96	2126.79	41.6
Neural network	2.51	7.00	64.2
Random forest	15.53	19.70	21.2
EuroSCORE I	15.94	26.93	40.8

The test data set was divided into 2 equal bins ordered by procedure date with approximately equal number of events ($n = 102$ vs $n = 105$). Goodness of fit χ^2 statistics were calculated for the first (G1) and second (G2) groups.

over LR model. Both LR and random forest models proved to be associated with good discrimination ability but substantial miscalibration. However, these 2 models showed the least calibration drift.

Interest in risk prediction models has bloomed in clinical use to aid in multidisciplinary shared decision-making. They are also used for benchmarking outcomes and both monitoring innovations. All this applies especially in an era of expanding

multimodal therapy for coronary artery and valve disease where risk prediction plays an important role in determining which patients would benefit most from surgery or percutaneous therapy. Moreover, national cardiac surgical registries have been established in many countries and they are used to develop risk prediction model with improved performance for local populations. Two of the most used risk stratification models in cardiac surgery the European System for Cardiac Operative Risk Evaluation version (EuroSCORE and EuroSCORE II) [1, 2] and the STS-PROM Score [3] were developed based on LR. The EuroSCORE I and II have been extensively criticized [14] including poor performance in external validation particularly for high-risk subgroups [15, 16]. This has been partially attributed to the small proportion (10%) of patients aged 75 years and above in the reference data set [17]. On the other hand, STS provides superior discrimination when compared to EuroSCORE II, but it shows suboptimal calibration, especially in the high-risk subgroup [18, 19].

Calibration drift can be attributed to improvement in perioperative management of patients; however, it is possible that poor calibration of EuroSCORE II and STS score can be partially attributed to the fact that these LR-based models overlook complex interactions among features and non-linear relationship. ML methods can capture interaction among features and non-linearity without input from the modeller, and this can potentially result in improved prediction. A recent systematic review [20] on the application of ML methods in cardiovascular diseases acknowledged the potential premise of ML in certain applications such as automated imaging interpretation. However, the advantage of ML methods over traditional risk stratification tools remains unclear. Mendes *et al.* [21] found that neural networks did not outperform LR when predicting mortality in patients after coronary artery bypass grafting. Other studies have suggested an advantage from ML methods over LR. Random forest has been shown to provide better discrimination when compared to LR, EuroSCORE and EuroSCORE II [22, 23]. Ghavidel *et al.* [24] found that decision trees achieved better discrimination power when compared to EuroSCORE and retrained LR. Nilsson *et al.* [25] found that neural networks using 34 features determined a small improvement in accuracy in mortality risk prediction when compared to LR and EuroSCORE. Recently, Kilic *et al.* [26] reported that a new ML method (i.e. extreme gradient boosting) may improve prediction in cardiac surgery when compared to the STS risk models. These discordant results can partially be explained by the fact that ML methods and in particular neural network need far more events per variable to be trained and therefore their application should only be considered if very large data sets are available [27]. An important limitation of available studies is that they focused on model discrimination while calibration has been inconsistently reported. Discrimination does not assess the model accuracy in individual risk predictions (calibration), which is crucial when using a predictive model to inform decisions about individual patient. Thus, a model might perform well based on discrimination measures while suffering substantial miscalibration [28].

The present study was designed to get insights into the usefulness of ML methods to improve individual risk prediction in cardiac surgery. We used a large data set collecting information on the set of features included in the EuroSCORE, and we assessed both model discrimination and calibration. We failed to show any significant advantage from ML methods over traditional LR model based on the same set of features included in the original EuroSCORE.

There are possible explanations for the lack of advantage from ML model over LR observed in the present study. We had a limited number of events (hospital deaths) to train and test prediction models despite the large original sample. This may have limited our ability to exploit the superiority of ML methods in identifying patterns of features related to the outcome. Moreover, automatic ML model hyper-tuning could not be performed as dedicated technology required was not available. Age at the time of surgery was the only continuous variable included in the models and this may have limited the ability of ML models to capture non-linear interaction for continuous variables. We did not train models using features included in the EuroSCORE II because preoperative creatinine value was reported as dichotomous variable (<200 or ≥200 mmol/l) while the actual value, which is part of the EuroSCORE II, was available only for a minority of patients. Similarly, we could not use the set of features of the STS-PROM score because our data set did not include some of the items needed for its calculation. The present analysis aimed to compare the performance of different algorithms based on the same set of features. Therefore, data-driven variable selection to improve model performance was not performed. Finally, we limited our analysis to in-hospital mortality to be consistent with current prediction models [2, 3], but we cannot exclude that ML algorithms can improve the prediction of long-term outcomes [29].

CONCLUSION

In conclusion, the present findings suggest that the application of ML algorithms alone is unlikely to determine a substantial gain in prediction of in-hospital mortality following cardiac surgery if a small set of structured clinical data is available. A precise estimation of individual risk is likely to be achieved only by the identification of new powerful predictors that can explain more of the variance observed.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *EJCTS* online.

Funding

This study was funded by the Bristol Biomedical Research Centre (Bristol BRC). T.R.G. receives support from the UK Medical Research Council [MC_UU_00011/4].

Conflict of interest: none declared.

Author contributions

Umberto Benedetto: Conceptualization; Funding acquisition; Methodology; Supervision; Writing–original draft; Writing–review & editing. **Shubhra Sinha:** Data curation; Writing–original draft; Writing–review & editing. **Matt Lyon:** Data curation; Formal analysis; Methodology; Writing–original draft; Writing–review & editing. **Arnaldo Dimagli:** Data curation; Writing–original draft; Writing–review & editing. **Tom R. Gaunt:** Methodology; Writing–review & editing. **Gianni Angelini:** Supervision; Writing–original draft; Writing–review & editing. **Jonathan Sterne:** Funding acquisition; Methodology; Supervision; Writing–review & editing.

Reviewer information

European Journal of Cardio-Thoracic Surgery thanks Milan Milojevic, Paul Sergeant, Alexander Wahba and the other, anonymous reviewer(s) for their contribution to the peer review process of this article.

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;16:9–13.
[2] Nashef SAM, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR *et al.* EuroSCORE II. Eur J Cardiothorac Surg 2012;41:734–45.
[3] Edwards FH, Clark RE, Schwartz M. Coronary artery bypass grafting: the Society of Thoracic Surgeons National Database experience. Ann Thorac Surg 1994;57:12–19.
[4] Provenchère S, Chevalier A, Ghodbane W, Bouleti C, Montravers P, Longrois D *et al.* Is the EuroSCORE II reliable to estimate operative mortality among octogenarians? PLoS One 2017;12:e0187056.
[5] Guida P, Mastro F, Scrascia G, Whitlock R, Paparella D. Performance of the European System for Cardiac Operative Risk Evaluation II: a meta-analysis of 22 studies involving 145,592 cardiac surgery procedures. J Thorac Cardiovasc Surg 2014;148:3049–57.e1.
[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 2019;110:12–22.
[7] Drew PJ, Monson J. Artificial neural networks. Surgery 2000;127:3–11.
[8] Kingma DP, Ba J. Adam: a method for stochastic optimization. CoRR, abs/1412.6980. 2014.
[9] Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and Cox regression. ArXiv, abs/1907.00825. 2019.
[10] Kingsford C, Salzberg SL. What are decision trees? Nat Biotechnol 2008; 26:1011–13.
[11] Sarica A, Cerasa A, Quattrone A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. Front Ageing Neurosci 2017;9:329.
[12] Zhang Z. Naïve Bayes classification in R. Ann Transl Med 2016;4:241.
[13] Nattino G, Finazzi S, Bertolini G. A new calibration test and a reappraisal of the calibration belt for the assessment of prediction models based on dichotomous outcomes. Statist Med 2014;33:2390–407.
[14] Sergeant P, Meuris B, Pettinari M. EuroSCORE II, illum qui est gravitates magni observe. Eur J Cardiothorac Surg 2012;41:729–31.
[15] 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.
[16] 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.
[17] 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.
[18] Osnabrugge RL, Speir AM, Head SJ, Fonner CE, Fonner E, Kappetein AP *et al.* Performance of EuroSCORE II in a large US database: implications for transcatheter aortic valve implantation. Eur J Cardiothorac Surg 2014;46:400–8.
[19] Kirmani BH, Mazhar K, Fabri BM, Pullan DM. Comparison of the EuroSCORE II and Society of Thoracic Surgeons 2008 risk tools. Eur J Cardiothorac Surg 2013;44:999–1005.
[20] Kilic A. Artificial intelligence and machine learning in cardiovascular health care. Ann Thorac Surg 2020;109:1323–9.
[21] Mendes RG, de Souza CR, Machado MN, Correa PR, Thommazo-Luporini LD, 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;4:756–63.
[22] 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.
[23] 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.
[24] 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.
[25] Nilsson J, Ohlsson M, Thulin L, Höglund P, Nashef SA, Brandt J. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. J Thorac Cardiovasc Surg 2006;132: 12–19.
[26] Kilic A, Goyal A, Miller JK, Gjekmarkaj E, Tam WL, Gleason TG *et al.* Predictive utility of a machine learning algorithm in estimating mortality risk in cardiac surgery. Ann Thorac Surg 2020;109:1811–19.
[27] 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.
[28] Omar RZ, Ambler G, Royston P, Eliahoo J, Taylor KM. Cardiac surgery risk modeling for mortality: a review of current practice and suggestions for improvement. Ann Thorac Surg 2004;77:2232–7.
[29] Linden A, Yarnold PR. Modeling time-to-event (survival) data using classification tree analysis. J Eval Clin Pract 2017;23:1299–308.