ESC European Society of Cardiology

European Heart Journal - Digital Health (2023) **4**, 433–443
https://doi.org/10.1093/ehjdh/ztad051

ORIGINAL ARTICLE

Predicting target lesion failure following percutaneous coronary intervention through machine learning risk assessment models

Mamas A. Mamas [1,*], Marco Roffi [2], Ole Fröbert[3], Alaide Chieffo[4], Alessandro Beneduce [4], Andrija Matetic [1,5], Pim A.L. Tonino[6], Dragica Paunovic[7], Lotte Jacobs[8], Roxane Debrus[9], Jérémy El Aissaoui[10], Frank van Leeuwen[8], and Evangelos Kontopantelis [11]

[1]Keele Cardiovascular Research Group, Centre for Prognosis Research, Institutes of Applied Clinical Science and Primary Care and Health Sciences, Keele University, Keele ST5 5BG, Newcastle, UK; [2]Department of Cardiology, University Hospitals Geneva, Geneva 1205, Switzerland; [3]Faculty of Health, Örebro University, Örebro 701 82, Sweden; [4]Interventional Cardiology Unit, San Raffaele Scientific Institute, Milan 20132, Italy; [5]Department of Cardiology, University Hospital of Split, Split 21000, Croatia; [6]Department of Cardiology, Catharina Hospital, Eindhoven 5623, The Netherlands; [7]Board of Directors, European Cardiovascular Research Centre (CERC), Massy 91300, France; [8]Medical and Clinical Division, Terumo Europe NV, Leuven 3001, Belgium; [9]Biostatistics Division, Genmab A/S, Copenhagen 1560, Denmark; [10]Artificial Intelligence Division, Business and Decision, Woluwe St Lambert, Brusells 1200, Belgium; and [11]Division of Informatics, Imaging and Data Sciences, School of Health Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre (MAHSC), University of Manchester, Manchester M13 9PL, UK

Received 8 May 2023; revised 22 August 2023; online publish-ahead-of-print 31 August 2023

Aims Central to the practice of precision medicine in percutaneous coronary intervention (PCI) is a risk-stratification tool to predict outcomes following the procedure. This study is intended to assess machine learning (ML)-based risk models to predict clinically relevant outcomes in PCI and to support individualized clinical decision-making in this setting.

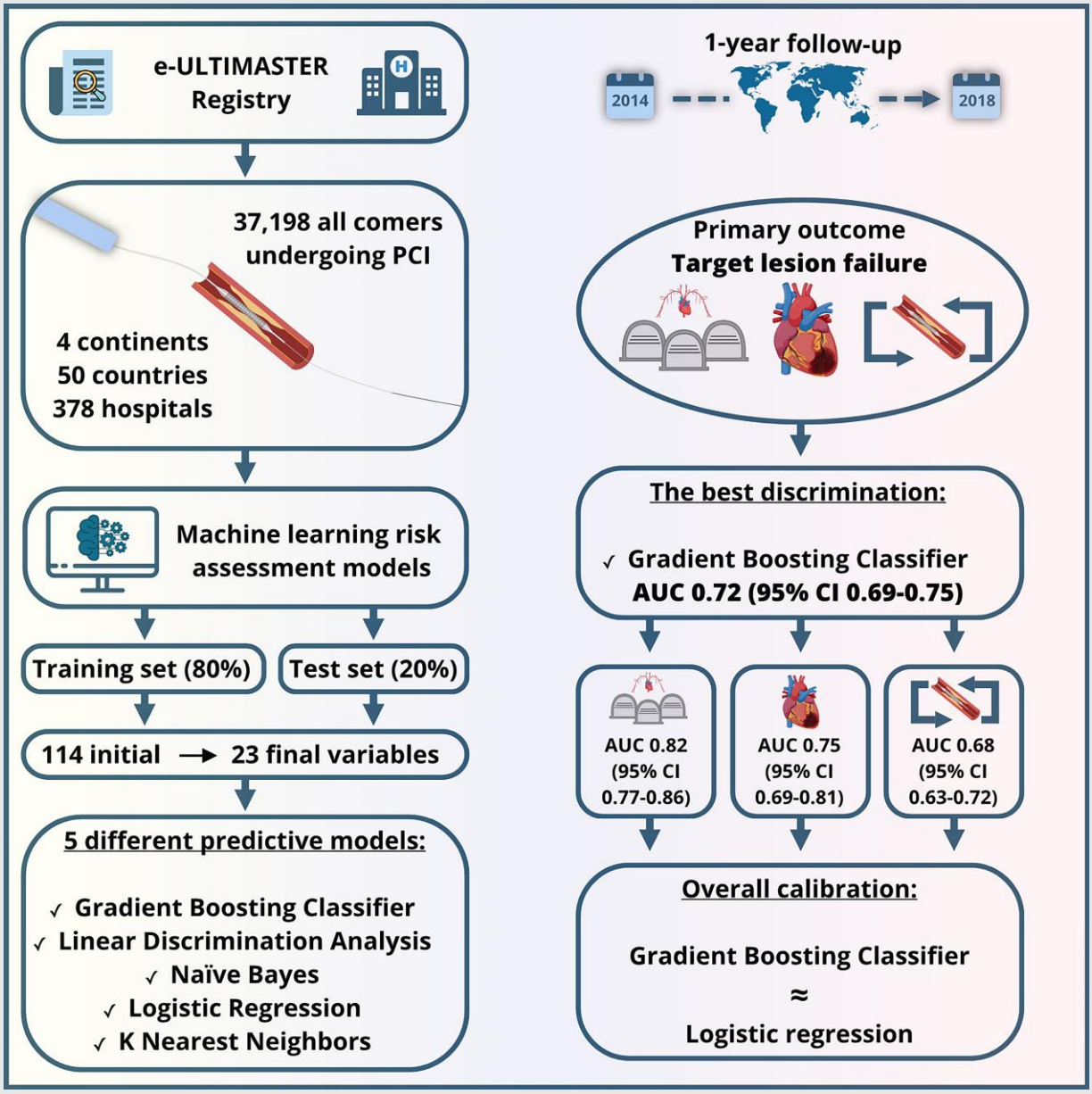
Methods and results Five different ML models [gradient boosting classifier (GBC), linear discrimination analysis, Naïve Bayes, logistic regression, and K-nearest neighbours algorithm) for the prediction of 1-year target lesion failure (TLF) were trained on an extensive data set of 35 389 patients undergoing PCI and enrolled in the global, all-comers e-ULTIMASTER registry. The data set was split into a training (80%) and a test set (20%). Twenty-three patient and procedural characteristics were used as predictive variables. The models were compared for discrimination according to the area under the receiver operating characteristic curve (AUC) and for calibration. The GBC model showed the best discriminative ability with an AUC of 0.72 (95% confidence interval 0.69–0.75) for 1-year TLF on the test set. The discriminative ability of the GBC model for the components of TLF was highest for cardiac death with an AUC of 0.82, followed by target vessel myocardial infarction with an AUC of 0.75 and clinically driven target lesion revascularization with an AUC of 0.68. The calibration was fair until the highest risk deciles showed an underestimation of the risk.

Conclusion Machine learning–derived predictive models provide a reasonably accurate prediction of 1-year TLF in patients undergoing PCI. A prospective evaluation of the predictive score is warranted.

Registration Clinicaltrial.gov identifier is NCT02188355.

* Corresponding author. Tel: +44 1782 671654, Fax: +44 1782 734719, Email: mamasmamas1@yahoo.co.uk

Graphical Abstract



Keywords Percutaneous coronary intervention • Machine learning • Drug-eluting stent • Target lesion failure • Outcomes

Introduction

Precision medicine is personalization of treatment based on an individual patient's characteristics and risk factor profile. Central to the practice of precision medicine is a risk-stratification tool that can predict the prognosis or outcome of a patient following different treatments.[1] Percutaneous coronary intervention (PCI) is the commonest means of achieving coronary revascularization[2] and outcomes are influenced by co-morbidities, disease extent, procedure complexity, and techniques. While many previous risk-stratification tools have focused on mortality, major bleeding complications, or the risk of future major adverse cardiovascular events (MACEs), there are no widely used risk scores predicting target lesion failure (TLF), a composite of cardiac death (CD), target vessel myocardial infarction (TV-MI), or clinically driven target lesion revascularization (TLR).[3–12] Target lesion failure is a frequent primary endpoint of current clinical trials as well as a clinically relevant outcome in practice.

Timely recognition of patients at risk for future TLF could identify patients who would benefit from intravascular imaging during their PCI procedure or the use of more potent antiplatelet regimes with a longer

duration. Furthermore, such models can be used for benchmarking clinical services at the centre and individual operator levels.

Machine learning (ML) enables computer algorithms to learn and perform certain tasks by automatically adapting internal parameters based on the input data without the need for human-written rules and, in the case of non-parametric ML models, without the need for strong hypotheses about the data. These features offer flexibility and, in specific cases, better performances than traditional statistical approaches. Machine learning prediction models have been developed in multiple areas of cardiovascular disease, including PCI. With some exceptions, most models used electronic hospital records with data from limited geographic areas, single hospitals, or narrow clinical presentations.[13–24]

Given the potential of ML to analyse large data sets with many variables and grasp numerous non-linear interactions among prognostic factors, we sought to generate a risk score for TLF using the data from a large, global, all-comers, PCI registry, e-ULTIMASTER.

Methods

Data set

We used the e-ULTIMASTER registry to develop the ML models. The e-ULTIMASTER registry (NCT02188355) was an all-comer, single-arm, prospective, multicentre study with clinical follow-up at 3 months and 1 year evaluating the safety and performance of the Ultimaster drug-eluting coronary stent system (Terumo Corporation, Tokyo, Japan) in daily clinical practice. Apart from general eligibility for PCI, there were no additional exclusion criteria for patient participation in this registry. The investigation was conducted worldwide, and 37 198 patients were enrolled between October 2014 and June 2018 in 378 hospitals from 50 countries across 4 continents/regions (Europe, Southeast Asia, South America/Mexico, and Africa/Middle East). To account for the wide geographic area and its potential impact on the outcomes following PCI, the variable 'region' corresponding to the above-mentioned regions has been included in the models.

The registry followed the Declaration of Helsinki (ISO 14155) and country-specific regulatory requirements. All patients signed the informed consent form that was reviewed and approved by the Institutional Review Board/Ethics Committee in each participating centre. Extensive online and risk-based on-site monitoring ensured that the collected data were of high quality.[25]

Outcomes and definitions

The primary outcome measure of the e-ULTIMASTER registry was TLF, defined as a composite of CD, with MI not attributable to a vessel other than the TV-MI and TLR at 1-year follow-up. The adverse events throughout the study were reported via an electronic web-based database. An independent clinical events committee adjudicated deaths, MIs, TLR and TV revascularization, and stent thromboses.

Training and validation process

The population was divided into a training set (80%) and a test set (20%) by stratified random sampling preserving the TLF rate, and five-fold cross-validation was used on the training set to evaluate the models. This standard procedure in ML allowed us to have a more robust estimation of the models' performance and fine-tune the algorithms at the validation stage without the risk of creating a feedback loop between the training set and the test set, which would have resulted in biased test set performance metrics.

Each classification algorithm was trained on four of the five folds and evaluated on the fifth, and the whole process was repeated five times, changing the validation fold at each iteration (see Supplementary material online, *Figure S1*). Hence, all the validation metrics are averages over the five folds, and the confidence intervals (CIs; for validation only) are computed based on the standard deviation (SD) of the performance metrics through the five iterations. We present the study following the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis reporting checklist (https://www.tripod-statement.org/).

Table 1 Variable (Gini) importance based on a gradient boosting classifier for target lesion failure at 1-year follow-up prediction

	Variable	Label	Importance
0	HBR	High bleeding risk	0.12222
1	AGE	Subject age at baseline	0.09703
2	BBMI	Body mass index at baseline	0.08485
3	PLEFN	Left ventricular ejection fraction (%)	0.08266
4	DIAB	Diabetes melitus	0.05530
5	LM	Left main vessel treated	0.04954
6	REGION	Geographical region	0.03267
7	PRNLI	Number of lesions identified	0.03050
8	PRIMDFL	Previous renal impairment disease	0.02963
9	GRF	Graft vessel treated	0.02960
10	CREAT	Creatinine kinase at pre procedure	0.02895
11	TROPO	Troponin pre procedure	0.02482
12	ACSK	Killip class for a patient with acute coronary syndrome	0.01948
13	PCMALFL	Current malignancies	0.01889
14	PPTCAFL	Previous PTCA	0.01849
15	STEMITRH	Haemodynamic support treatment of STEMI before PCI	0.01643
16	PRCTCR	Complete revascularization of the coronary tree	0.01608
17	STEMILOC	Location of the STEMI	0.01588
18	PMIFL	Previous MI	0.01567
19	NTVESTR	Number of target vessel treated	0.01208
20	LTBIF	Bifurcation lesion type treated	0.01181
21	PRETIMI	Pre-angiographic—TIMI flow	0.01095
22	ITLBIFT	True bifurcation lesion type treated	0.01079

Variable selection

Out of the initial 144 database variables, the variables for which the value is typically unknown at the time of the procedure were removed, namely post-procedural/discharge data (including follow-up). To limit the risk of over-fitting and keep a number of variables manageable for potential future inference scenarios, we selected only the most important variables based on the Gini impurity reduction criterion used to construct a gradient boosting classifier (GBC) model.

Gradient boosting classifier is an ensemble method combining many weak learners, typically small decision trees, to produce a more robust predictive model. The variables (and variable values) used to create the splits at each tree node are chosen based on their ability to separate the different classes (event/non-event). The purity of the split is evaluated by using the Gini index:

$$G = p(1 - p)$$

where p is the observed event frequency in the tree node.[26] A perfect classification gives a frequency of 1 (p) in one node and 0 in the other ($1 - p$), giving an impurity of 0, while the maximum impurity is reached for a useless classifier ($p = 1/2$). Each tree is trained to predict the residuals of the previous one and gets assigned a weight depending on the overall impurity reduction it achieves. The final prediction is obtained by a weighted vote of all the weak learners.

This algorithm offers a very natural method of assessing the importance of the predictors for the classification problem. The Gini impurity reduction achieved by a given variable weighted by the proportion of the population reaching that considered node in the classifier is used as a variable

Table 2 Variable importance (raw and normalized score) for pre-selected variables for target lesion failure at 1-year follow-up prediction

Variable	Normalized importance score (TLF1Y)					Raw importance score (TLF1Y)				
	GBC	LDA	LR	NB	KNN	GBC	LDA	LR	NB	KNN
HBR	0.1133	0.1055	0.0932	0.0943	0.0851	0.1133	0.2996	1.2086	0.3152	0.026
AGE	0.1051	0.0307	0.0523	0.0819	0.0153	0.1051	0.0873	1.117	0.2738	0.0047
BBMI	0.0503	0.0088	0.0103	0.0156	0.0191	0.0503	−0.0248	0.9769	−0.0521	0.0058
PLEFN	0.0718	0.0378	0.0353	0.0424	0.029	0.0718	−0.1072	0.921	−0.1419	0.0089
DIAB	0.0584	0.0017	0.0026	0.0027	0.0075	0.0584	−0.0048	0.9943	0.0092	0.0023
LM	0.0959	0.1149	0.0785	0.0655	0.0855	0.0959	0.3261	1.1757	0.219	0.0261
REGION	0.04	0.0074	0.0057	0.0195	0.0364	0.04	0.021	1.0128	0.0651	0.0111
PRNLI	0.0603	0.0844	0.0954	0.0855	0.0711	0.0603	0.2397	1.2134	0.2858	0.0217
PRIMDFL	0.0887	0.0568	0.0395	0.0653	0.0586	0.0887	−0.1612	0.9116	−0.2184	0.0179
GRF	0.0508	0.0674	0.0415	0.0415	0.0617	0.0508	0.1913	1.093	0.1389	0.0188
CREAT	0.0312	0.0541	0.0567	0.0502	0.0314	0.0312	−0.1536	0.8731	−0.168	0.0096
TROPO	0.0283	0.0358	0.0418	0.054	0.053	0.0283	−0.1017	0.9065	−0.1807	0.0162
ACSK	0.0132	0.0463	0.0615	0.0021	0.0493	0.0132	0.1315	1.1376	−0.0072	0.0151
PCMALFL	0.0019	0.0103	0.0083	0.0045	0.0367	0.0019	−0.0293	0.9813	0.015	0.0112
PPTCAFL	0.0487	0.0732	0.0995	0.0636	0.0424	0.0487	0.2079	1.2227	0.2128	0.0129
STEMITRH	0.0254	0.0495	0.0427	0.0295	0.0254	0.0254	0.1407	1.0955	0.0986	0.0078
PRCTCR	0.0216	0.0249	0.0292	0.0507	0.0017	0.0216	−0.0707	0.9347	−0.1695	0.0005
STEMILOC	0.039	0.0662	0.0765	0.0269	0.0703	0.039	−0.1879	0.8288	−0.0901	0.0215
PMIFL	0.0179	0.039	0.0479	0.0558	0.0506	0.0179	0.1108	1.1072	0.1864	0.0154
NTVESTR	0.0037	0.0147	0.0118	0.0372	0.0274	0.0037	−0.0416	0.9735	0.1243	0.0084
LTBIF	0.0132	0.0377	0.0432	0.0503	0.0477	0.0132	0.1071	1.0967	0.1683	0.0146
PRETIMI	0.0143	0.0256	0.0254	0.0266	0.0451	0.0143	0.0728	1.0569	0.0889	0.0138
ITLBIFT	0.0069	0.0073	0.0013	0.0343	0.0496	0.0069	0.0207	1.0028	0.1149	0.0151
INTERCEPT							−3.7518	0.0263		

GBC, gradient boosting classifier; LDA, linear discriminant analysis; LR, logistic regression; NB, Naïve Bayes; KNN, K-nearest neighbours.

Table 3 *P*-values of the variable importance scores correlation

	GBC	LDA	LR	NB	KNN
GBC	—	0.005	0.028	0.002	0.172
LDA		—	0.000	0.001	0.000
LR			—	0.001	0.002
NB				—	0.046
KNN					—

GBC, gradient boosting classifier; LDA, linear discriminant analysis; LR, logistic regression; NB, Naïve Bayes; KNN, K-nearest neighbours.

importance measure. The Gini impurity reduction achieved by all the variables is then normalized (the importance of all the variables sums to 1). Our analysis used this metric to select the predictors to be kept for modelling. For all 5 models, we chose to keep all 23 variables with a normalized importance of above 1% (*Table 1*).

Predictive algorithms and evaluation metrics

Five different predictive models were tested for the ability to discriminate between outcome classes: GBC, linear discriminant analysis (LDA), logistic regression (LR), Naïve Bayes (NB), and K-nearest neighbours (KNN).

- The GBC model is an ensemble tree-based classifier that has been briefly described in the Variable Selection section.
- The LDA classifier uses a linear decision surface to separate classes. It is a simple model with no hyperparameters and a closed-form solution.
- The LR model used in this analysis is the conventional LR.
- The NB algorithm considered here is the Gaussian NB, assuming Gaussian distributions for the variables with strong independence.
- The KNN classifier attributes a class to each data point by implementing a vote of the K-nearest points (Euclidian distance).

The models were assessed on their discriminatory ability as reflected by the area under the receiver operating characteristic (ROC) curve (AUC). The sensitivity and specificity were also computed based on a score threshold maximizing the Youden's J statistics.[27] The CIs were obtained via bootstrapping of the inference score on the training and the test sets and from the SD of the metrics over the cross-validation folds for the validation set. The AUC values were then compared pairwise with the DeLong's method.[28,29] Finally, the calibration of the models was evaluated by computing the calibration curves. To assess the calibration, the test set patients were categorized into deciles of the predicted risk score (e.g. Decile 1 = 10% of the population with the lowest predicted risk score, Decile 10 = 10% of the population with the highest predicted risk score). For each decile, we computed the average predicted risk score (x axis) and the observed event rate (y axis). In the calibration plots, the diagonal line represents perfect calibration with a perfect correlation of predicted estimates with observed event rates. Deviations above the diagonal line represent a model that underestimates risk, and deviations below the diagonal line represent a model that overestimates risk.

Table 4 Baseline patient characteristics of patients with 1-year follow-up

Patient characteristics	All patients ($n = 35\,389$)	Patient without TLF event ($n = 34\,254$)	Patient with TLF event ($n = 1135$)
Age, years	64.3 ± 11.2	64.2 ± 11.2	67.7 ± 11.5
Octogenarians (≥80 years)	8.9	8.7	15.7
Gender, male	75.9	75.9	76.0
Body mass index, kg/m^2	27.8 ± 4.6	27.8 ± 4.6	27.6 ± 4.9
≤18.5	0.7	0.7	1.8
18.5–24.9	27.7	27.6	28.2
25–29.9	44.5	44.5	43.9
≥30	27.1	27.2	26.2
Diabetes mellitus	28.3	28.0	38.7
Insulin dependent	20.6	20.3	27.2
Non-insulin dependent	79.3	79.6	72.6
Unknown	0.1	0.1	0.2
Smoking			
Never	41.5	41.6	38.8
Previous	32.6	32.4	38.3
Current	25.9	26.0	23.0
Hypertension	67.6	67.5	71.5
Hypercholesterolaemia	59.6	59.5	62.6
Family history of heart disease	35.5	35.7	30.7
Previous MI	22.9	22.6	32.3
Left ventricular ejection fraction (%)	53.8 ± 11.7	53.9 ± 11.6	50.4 ± 14.2
Previous revascularization			
Previous PCI	26.1	25.8	36.9
Previous CABG	5.7	5.5	12.8
Atrial fibrillation on OAC	5.7	5.5	10.1
Previous stroke	5.5	5.4	9.5
Peripheral vascular disease	6.7	6.5	13.0
Congestive heart failure	11.3	11.1	16.5
Renal impairment	7.0	6.7	16.6
Clinical presentation			
Silent ischaemia	9.2	9.3	8.4
Stable angina	35.8	35.9	33.3
Unstable angina	11.8	11.8	11.5
NSTEMI	23.2	23.1	26.2
STEMI	20.0	19.9	20.6

Data are mean ± SD for continuous variables or % for categorical variables. Renal impairment: estimated glomerular filtration rate <60 mL/min/1.73 m^2.
CABG, coronary artery bypass graft; MI, myocardial infarction; (N)STEMI, (non)ST-segment elevation myocardial infarction; OAC, oral anticoagulants; PCI, percutaneous coronary intervention.

To assess the content of the models, we computed the variable importance (out of the pre-selected variables) with methods suited for each of the algorithms (LR: odds ratio, GBC: Gini impurity reduction, LDA: linear coefficients, NB: inverse coefficient of variation, KNN: permutation importance). As some methods assess only the discriminatory importance and not the direction of the contribution, the absolute value of the scores has been taken, the scores have been normalized, and the variable rank computed to allow for a more robust comparison across models. Finally, the pairwise correlation of the variable importance has been computed for each pair of models along with the associated *P*-value. The results are displayed in *Tables 2* and *3*.

Individual endpoints

As the 1-year TLF is a composite of three distinct endpoints (CD, TV-MI, and TLR), there was the risk of attempting to identify a heterogeneous group of patients (with mixed profiles). We expect some models built on individual endpoints to perform better, as the targeted patient profile might be more specific in such a scenario.

In order to investigate the above question, we re-ran the analysis, focusing on one endpoint (CD, TV-MI, and TLR) at a time. Only the GBC model was considered, while the rest of the methodology was unchanged. The predictive results of each model for CD, TV-MI, and TLR are given in Supplementary material online, *Table S1*.

Results

The e-ULTIMASTER registry enrolled 37 198 patients. The 1-year follow-up rate was 95.1%, with 35 389 patients included in the analysis.

Table 5 Performance metrics for gradient boosting classifier, linear discriminant analysis, logistic regression, Naïve Bayes, and K-nearest neighbours on training, validation, and test sets

	Model	AUC (95% CI)	Specificity (95% CI)	Sensitivity (CI 95% CI)	Threshold
Training	GBC	0.73 (0.72–0.75)	0.8 (0.79–0.8)	0.56 (0.53–0.59)	0.0409
	LDA	0.69 (0.67–0.71)	0.75 (0.74–0.75)	0.53 (0.50–0.57)	0.0319
	LR	0.69 (0.67–0.71)	0.75 (0.75–0.76)	0.53 (0.49–0.56)	0.0356
	NB	0.68 (0.66–0.69)	0.65 (0.64–0.66)	0.61 (0.57–0.64)	0.0041
	KNN	0.71 (0.69–0.72)	0.64 (0.63–0.64)	0.65 (0.62–0.68)	0.0250
Validation	GBC	0.70 (0.69–0.71)	0.72 (0.66–0.77)	0.60 (0.55–0.65)	0.0409
	LDA	0.68 (0.67–0.69)	0.69 (0.61–0.77)	0.59 (0.51–0.67)	0.0319
	LR	0.68 (0.67–0.69)	0.73 (0.69–0.78)	0.55 (0.49–0.61)	0.0356
	NB	0.67 (0.66–0.68)	0.63 (0.55–0.72)	0.64 (0.54–0.74)	0.0041
	KNN	0.65 (0.65–0.66)	0.74 (0.67–0.82)	0.48 (0.41–0.56)	0.0250
Test	GBC	0.72 (0.69–0.75)	0.80 (0.79–0.81)	0.53 (0.47–0.6)	0.0409
	LDA	0.70 (0.66–0.73)	0.75 (0.74–0.76)	0.55 (0.49–0.61)	0.0319
	LR	0.70 (0.66–0.73)	0.76 (0.75–0.77)	0.54 (0.48–0.61)	0.0356
	NB	0.70 (0.67–0.73)	0.66 (0.65–0.67)	0.64 (0.57–0.69)	0.0041
	KNN	0.70 (0.66–0.73)	0.64 (0.63–0.65)	0.66 (0.60–0.72)	0.0250

AUC, area under the curve; CI, confidence interval; GBC, gradient boosting classifier; LDA, linear discriminant analysis; LR, logistic regression; NB, Naïve Bayes; KNN, K-nearest neighbours.

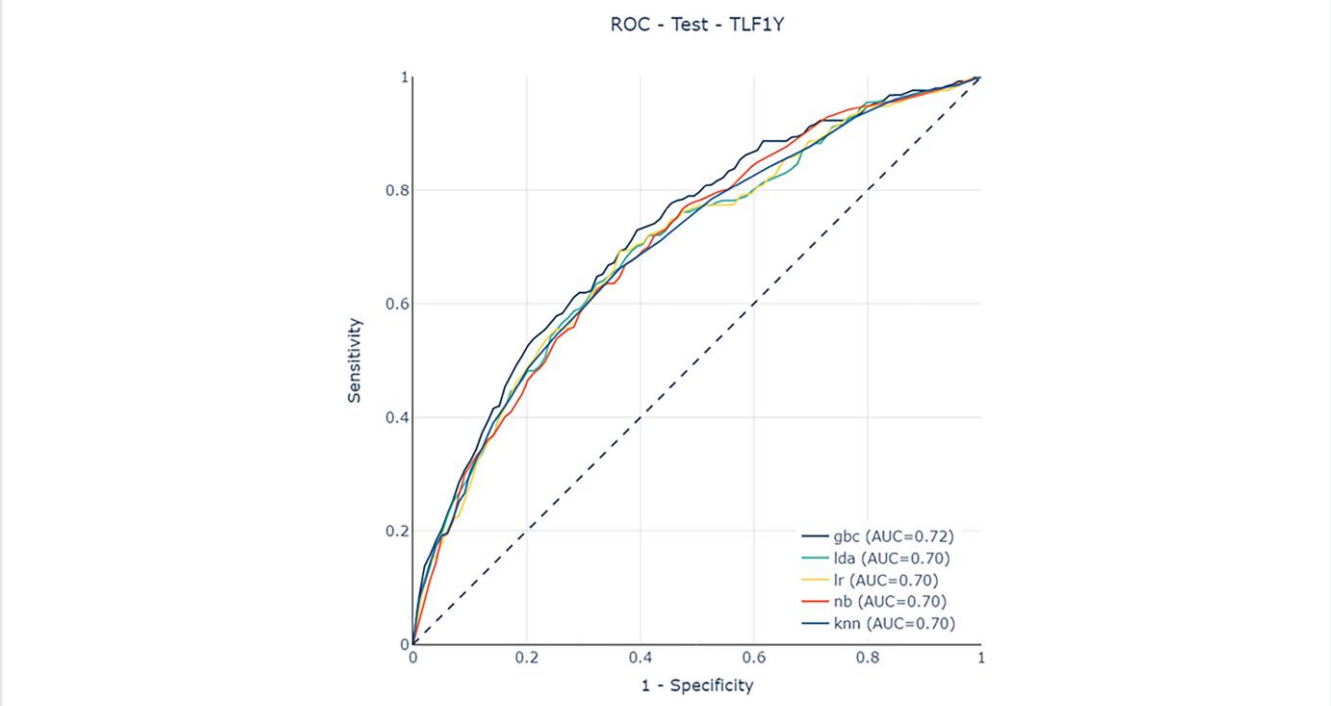


Figure 1 Receiver operating characteristic curves and performance metrics on test sets. AUC, area under the curve; GBC, gradient boosting classifier; LDA, linear discriminant analysis; LR, logistic regression; NB, Naïve Bayes; KNN, K-nearest neighbours; ROC, receiver operating characteristic; TLF1Y, target lesion failure at 1-year follow-up.

The mean age of patients was 64.3 (SD = 11.2) years, and the majority were males (76.0%), 28.3% had diabetes (of which 20.6% insulin dependent), 67.6% had hypertension, and 59.6% hypercholesterolaemia. A history of MI, PCI, coronary artery bypass grafting, peripheral artery disease, and stroke were present in 22.9, 26.1, 5.7, 6.7, and 5.5% of the patients, respectively. Patients had on average 1.8 (SD = 1.1) lesions, 20.5% classified as the American College of Cardiology/American Heart Association Class C, and 46.1% had multivessel disease. More

Table 6 Comparison of the areas under the receiver operating characteristic curves by the fast DeLong's algorithm on the test set

	GBC	LDA	LR	NB	KNN
GBC	—	<0.001	<0.001	0.02	0.01
LDA		—	0.98	0.71	0.94
LR			—	0.73	0.94
NB				—	0.82
KNN					—

GBC, gradient boosting classifier; LDA, linear discriminant analysis; LR, logistic regression; NB, Naïve Bayes; KNN, K-nearest neighbours.

than half (56.8%) of patients presented with acute coronary syndrome (*Table 4*). More than three-quarters of the procedures (82.8%) were performed via radial access, while intravascular imaging was rarely used (6.2%). A 1-year TLF occurred in 1135 patients (3.2%), 455 (1.3%) had CD, 316 (0.9%) suffered a TV-MI, and 591 (1.7%) underwent TLR.

Predictive models for the primary composite endpoint

The discriminatory ability of different models for the primary composite endpoint of 1-year TLF is presented in *Table 5* and *Figure 1*. Gradient boosting classifier showed the best AUC values in both the training, 0.73 (95% CI 0.72–0.75) and test 0.72 (95% CI 0.69–0.75) sets, outperforming other ML models in which AUC values were in the range of 0.68–0.70 in training and 0.70, for all, in the test set. The LR showed similar AUC, the training set 0.69 (95% CI 0.67–0.71) and the test set 0.70 (95% CI 0.66–0.73). A comparison of the AUC values by the fast DeLong's algorithm on the test set showed a significantly better discriminatory ability for GBC (*Table 6*). The overall accuracy was the highest for the GBC model (0.79) with slightly lower scores for LR (0.75) and LDA (0.74) (see Supplementary material online, *Table S2*).

The calibration test showed large differences in the different ML models (*Figures 2* and *3*; Supplementary material online, *Figure S2*). The GBC model showed good calibration until the last two deciles, indicating an underestimation of the risk by 1.5–2% for patients at higher risk. Conventional LR showed a similar pattern, although with better calibration in the highest-risk decile.

The variable importance scores given by the different models are significantly correlated ($P < 0.05$; *Table 6*), except for GBC–KNN. Gradient boosting classifier shows the strongest differences in variable importance, giving more weight to DIAB, BBMI, PLEFN, PRIMDFL, and less to ACSK than LR. The top 3 most agreed-upon variables are HBR, LM, and PRNLI, with an average importance rank of 1.8, 2.6, and 3.2, respectively.

Predictive models for individual endpoints

The GBC models on CD and TV-MI showed better performances despite the limited number of events to learn from (AUC 0.82 and 0.75, respectively). The TLR model is less performant than the composite endpoint model, with an AUC of 0.68 in the test set (see Supplementary material online, *Table S1*).

Discussion

The present findings represent the first report of ML models to develop a predictive score for 1-year TLF in contemporary PCI practice derived from a global population of patients across Europe, South East Asia, South America/Mexico, and Africa/Middle East. The main findings of the study are: (i) the GBC demonstrated accurate discriminative capability in predicting 1-year TLF with improved performance when compared with the other models; (ii) the GBC prediction ability was improved for CD and TV-MI, while the performance for TLR was lower than the composite endpoint (TLF); (iii) the calibration of the GBC model was comparable with LR, and both models showed good calibration until the last two risk deciles.

Using a contemporary database of >35 000 patients from almost 400 hospitals worldwide, we generated 5 different ML risk models to predict TLF events at 1-year post PCI and identified the most robust algorithm. This analysis suggests that ML is a valuable tool to generate a prediction model for TLF after PCI and offers the potential to include imaging data and physiological parameters in future iterations. The results of the present study are comparable with previously reported predictive scores in PCI derived by ML,[15–21] but with the potential advantage of the inclusion of a general PCI population and wide geographic distribution from 50 countries across 4 continents.[15–21] Therefore, these findings may have a broader applicability to different healthcare systems in patients treated within them.

Apart from GBC, other models in the test set had an AUC of 0.70, a conventionally used threshold for acceptable performance. Several factors could influence performance. First, the TLF event rate was low in this study and even with large population sizes and advanced modelling algorithms, predicting rare events is fundamentally complex. Second, TLF is a composite endpoint of CD, TV-MI, and TLR, all of which might have different underlying risk factors making the development of a single model to predict a composite of these outcomes complex. Also, the data originate from a global registry with different clinical practices around revascularization, intravascular imaging, and post-intervention pharmacological management across different heterogeneous populations. Nevertheless, this is a strength of our analysis as it represents real-world practice, applicable to a wider geographical area and potentially provides greater utility in the real world.

The calibration of the GBC model showed a good agreement between predicted and observed events in the first eight deciles, with an underestimation in the highest risk deciles. In clinical practice, both physician and patient are primarily concerned about the accuracy of a prognostic estimate of a risk model. Therefore, calibration plays an equally important role as discrimination. A poorly calibrated model may over or underestimate the risk. Our model's calibration was well aligned in lower risk categories but the accuracy decreased in the highest two deciles. Whether the calibration performance in the highest risk categories represents an obstacle for risk score use requires additional consideration. Nevertheless, similar phenomena have been observed in other national PCI risk scores.[11,12]

Previous studies aiming to generate a predictive score for composite endpoints such as MACEs, which include death, MI, and TLR, using conventional statistical methods have shown moderate discriminatory ability. A study comparing 6 conventional risk scores used frequently in daily clinical practice found that AUC for MACE varied between 0.53 and 0.63, and the best AUC for mortality only was 0.76.[30] At the same time, the risk for TLR was not reliably predicted by any of the scores. The authors concluded that clinical factors are of reduced utility for predicting MI and TLR, although they are critical in predicting mortality.

Given these competing aspects, when outcomes are combined to derive TLF, creating a well-performing risk score may be challenging. Other studies reported similar results when assessing the performance of risk scores for composite PCI-related endpoints. The study by Garg *et al.*[31] validated well-established clinical [age, creatinine, and ejection fraction (ACEF)], anatomical [syntax score (SX)], and combined clinical/anatomical [clinical syntax score (CSX)] scores on an all-comers patient population. The authors report that purely clinical ACEF score had a low discriminatory ability for predicting TLF with an AUC of 0.59. In contrast, the SX and CSX scores had slightly better

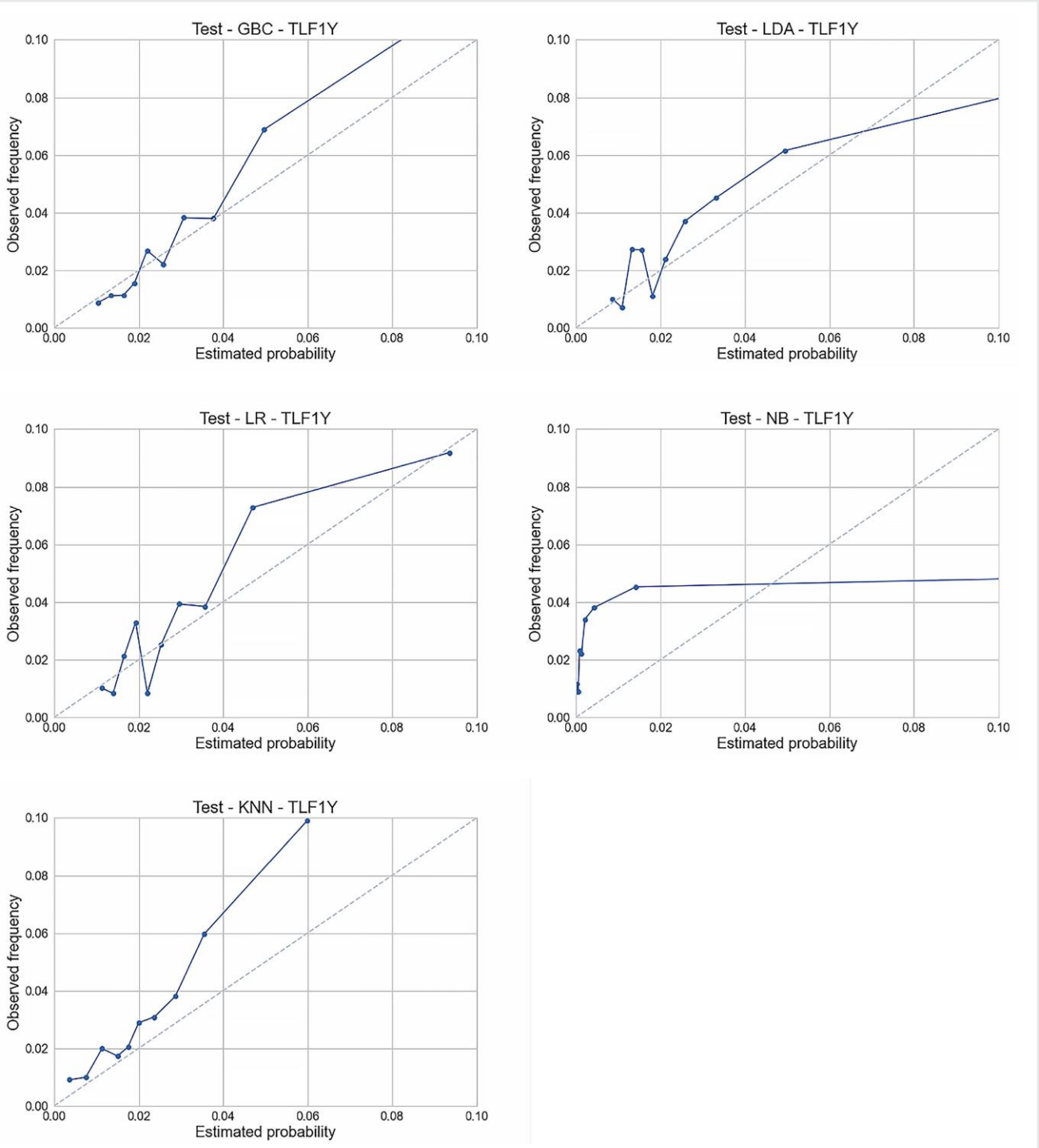


Figure 2 Probability calibration curves on the test set. GBC, gradient boosting classifier; LDA, linear discriminant analysis; LR, logistic regression; NB, Naïve Bayes; KNN, K-nearest neighbours; TLF1Y, target lesion failure at 1-year follow-up.

discriminative power with an AUC of 0.62 and 0.63, respectively. Both scores with clinical predictors had high discrimination in predicting CD (0.84 ACEF and 0.71 CSX), while the CX score only modestly predicted TLR with an AUC of 0.63. Compared with those results, the ML model created in the current analysis could be more helpful in identifying the patients at higher risk of developing TLF.

More than half of the patients with a TLF in the present study had TLR (1.7%), while 1.3 and 0.9% of patients had CD or suffered TV-MI. Considering that different, often competing, factors influence TLR and CD and that some patients experienced multiple events, ML obtained scores' discriminatory ability appears acceptable. Although the model performance was moderate, similar models are lacking and

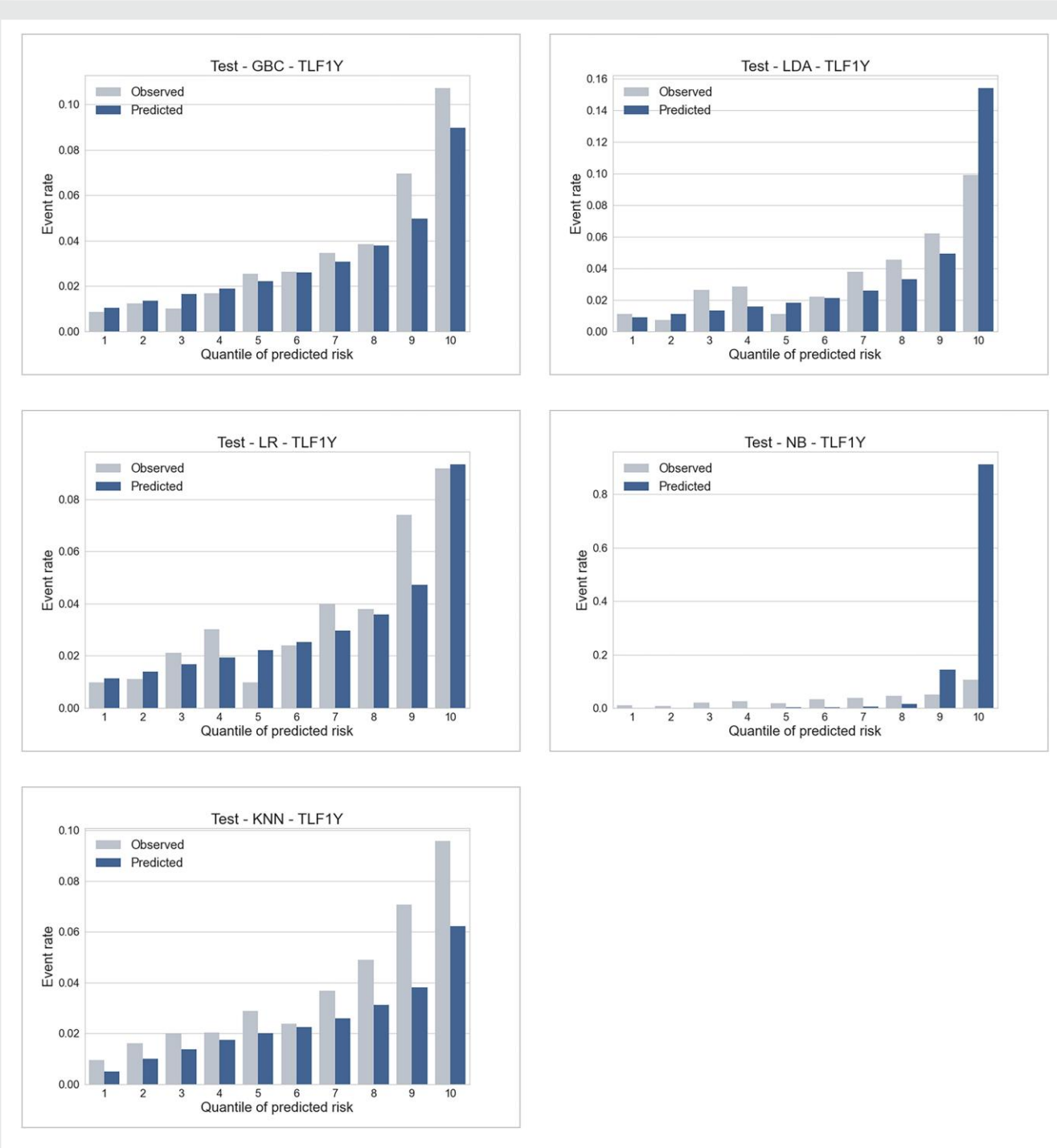


Figure 3 Predicted vs. observed frequency of events per risk deciles. Decile 1 represents 10% of the population with the lowest predicted risk score, Decile 10 represents 10% of the population with the highest predicted risk score. GBC, gradient boosting classifier; LDA, linear discriminant analysis; LR, logistic regression; NB, Naïve Bayes; KNN, K-nearest neighbours; TLF1Y, target lesion failure at 1-year follow-up.

this study offers other potential advantages (large population size, wide geographic region, and advanced modelling algorithms) that should allow its applicability to real-world population undergoing PCI. For the individual endpoint analysis with the GBC model in the test set, we produced the highest performance for CD (AUC 0.82), TV-MI (AUC 0.75), and the lowest for TLR (AUC 0.67). Those results point towards the different predictors for different events and perhaps the high impact that procedural factors (excluded from the current analysis) play in developing TLR and TV-MI events.

The present analysis identified patients with the highest risk for TLF. Those patients experienced a 10–15 times higher event rate when compared with patients within the lowest risk category. Recognizing patients at a higher risk of adverse events would facilitate informed discussion between the patient and physician, a joint decision about the

most appropriate treatment option, and better counselling for those patients.

Timely recognition of patients at a higher risk of adverse events could be actioned, highlighting patients who may benefit from intravascular imaging use, use of more potent antiplatelet regimes, or treatment with dual antiplatelet therapy for more prolonged durations. Machine learning risk models could support individualized clinical decision-making in this setting. Although many efforts are invested in developing, implementing, and refining ML models in medicine, there is still a lack of standards to evaluate those tools and safeguard patients against unintended consequences.[32] Nevertheless, ML models can process many variables, assimilate new data in real-time when linked with electronic hospital records, and continuously improve their predictive accuracy. Furthermore, imaging modalities, unstructured text, and other data carriers could be incorporated, something not currently possible with conventional methods.

Limitations

The current study has several limitations. First is the absence of data outside the e-ULTIMASTER study for external validation. However, a large number of patients, five-fold cross-validation, and random data split into derivation and test cohorts may partly address this limitation. Second, the e-ULTIMASTER registry did not capture data on optimal stent deployment or intravascular imaging that would provide important information about the tendency towards the development of TLF. The analysis did not include antithrombotic management of patients post PCI and secondary prevention, with the potential impact on survival and stent thrombosis with a consequent MI and TLR. However, these are modifiable factors that are not known when treatment decisions are taken. The selection of predictors using only a GBC is another limitation as it could bias other models. However, the features of GBC are highly advantageous for this purpose, and the probability of bias is minimal. The LR produced better results than in many prior scores, indicating that the key could be the selection of parameters. Furthermore, the analysis did not include other potentially advantageous models such as the 'random forest' model, but this was carefully outweighed by the authors during the initiation phase of the study. Finally, this study encompassed patients with only one specific type of bioresorbable polymer drug-eluting stent, and therefore, a wider application of these results to other stent platforms becomes questionable.

Conclusions

The ML-derived predictive models provided a reasonably accurate prediction of 1-year TLF in an all-comer patient population undergoing PCI. A prospective evaluation of the predictive score is warranted, including its external validation, to better understand the clinical implications of these findings.

Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health*.

Funding

The e-ULTIMASTER study was funded by Terumo Europe NV, Leuven, Belgium.

Conflict of interest: M.A.M. receives unrestricted educational grant and consulting fees from Terumo; M.R. receives unrestricted educational grant from Terumo, Cordis/Cardinal Health, Medtronic, and Biotronik; O.F. receives speaker fees from Sanofi Pasteur; A.C. receives speaker fees from Abbott Vascular, Biosensor, and Boston Scientific, as well as consulting fees from Abiomed and Shock Wave Medical; D.P. is a former employee of Terumo Europe NV (Belgium); L.J. is an employee of Terumo Europe NV (Belgium); R.D. is a former employee of Terumo Europe NV (Belgium).

Data availability

The data underlying this article will be shared on reasonable request with the corresponding author.

Consent

All patients signed the informed consent form that was reviewed and approved by the Institutional Review Board/Ethics Committee in each participating centre.

References

1. Garratt KN, Schneider MA. Thinking machines and risk assessment: on the path to precision medicine. *J Am Heart Assoc* 2019;**8**:e011969.
2. Neumann FJ, Sousa-Uva M, Ahlsson A, Alfonso F, Banning AP, Benedetto U, *et al.* 2018 ESC/EACTS guidelines on myocardial revascularization. *Eur Heart J* 2019;**40**:87–16.
3. Cutlip DE, Windecker S, Mehran R, Boam A, Cohen DJ, van Es GA, *et al.* Clinical endpoints in coronary stent trials: a case for standardised definitions. *Circulation* 2007;**115**:2344–2351.
4. Sianos G, Morel MA, Kappetein AP, Morice MC, Colombo A, Dawkins K, *et al.* The SYNTAX score: an angiographic tool grading the complexity of coronary artery disease. *EuroIntervention* 2005;**1**:219–227.
5. Farooq V, van Klavern D, Steyerberg EW, Meliga E, Vergouwe Y, Chieffo A, *et al.* Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: development and validation of SYNTAX score II. *Lancet* 2013;**381**:639–650.
6. Farooq V, Vergouwe Y, Räber L, Vranckx P, Garcia-Garcia H, Diletti R, *et al.* Combined anatomical and clinical factors for the long-term risk stratification of patients undergoing percutaneous coronary intervention: the Logistic Clinical SYNTAX score. *Eur Heart J* 2012;**33**:3098–3104.
7. Wu C, Hannan EL, Walford G, Ambrose JA, Holmes DR Jr, King SB III, *et al.* A risk score to predict in-hospital mortality for percutaneous coronary interventions. *J Am Coll Cardiol* 2006; **47**:654–660.
8. Peterson ED, Dai D, DeLong ER, Brennan JM, Singh M, Rao SV, *et al.* Contemporary mortality risk prediction for percutaneous coronary intervention: results from 588,398 procedures in the National Cardiovascular Data Registry. *J Am Coll Cardiol* 2010;**55**:1923–1932.
9. Ranucci M, Castelvecchio S, Menicanti L, Frigiola A, Pelissero G. Risk of assessing mortality risk in elective cardiac operations: age, creatinine, ejection fraction, and the law of parsimony. *Circulation* 2009;**119**:3053–3061.
10. Garg S, Sarno G, Garcia-Garcia HM, Girasis C, Wykrzykowska J, Dawkins KD, *et al.* A new tool for the risk stratification of patients with complex coronary artery disease: the Clinical SYNTAX Score. *Circ Cardiovasc Interv* 2010;**3**:317–326.
11. McAllister KSL, Ludman PF, Hulme W, de Belder MA, Stables R, Chowdhary S, *et al.* A contemporary risk model for predicting 30-day mortality following percutaneous coronary intervention in England and Wales. *Int J Cardiol* 2016;**210**:125–132.
12. Brennan JM, Curtis JP, Dai D, Fitzgerald S, Khandelwal AK, Spertus JA, *et al.* Enhanced mortality risk prediction with a focus on high-risk percutaneous coronary intervention: results from 1,208,137 procedures in the NCDR (National Cardiovascular Data Registry). *JACC Cardiovasc Interv* 2013;**6**:790–799.
13. Westcott RJ, Tcheng JE. Artificial intelligence and machine learning in cardiology. *JACC Cardiovasc Interv* 2019;**12**:1312–1314.
14. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, *et al.* Artificial intelligence in cardiology. *J Am Coll Cardiol* 2018;**71**:2668–2679.
15. Auerbach A, Fihn SD. Discovery, learning, and experimentation with artificial intelligence–based tools at the point of care—perils and opportunity. *JAMA Netw Open* 2021;**4**:e211474.
16. Hsieh MH, Lin SY, Lin CL, Hsieh MJ, Hsu WH, Ju SW, *et al.* A fitting machine learning prediction model for short-term mortality following percutaneous catheterisation intervention: a nationwide population-based study. *Ann Transl Med* 2019;**7**:732.
17. Mortazavi BJ, Bucholz EM, Desai NR, Huang C, Curtis JP, Masoudi FA, *et al.* Comparison of machine learning methods with national cardiovascular data registry models for prediction of risk of bleeding after percutaneous coronary intervention. *JAMA Netw Open* 2019;**2**:e196835.
18. Wang Y, Zhu K, Li Y, Lv Q, Fu G, Zhang W. A machine learning-based approach for the prediction of periprocedural myocardial infarction by using routine data. *Cardiovasc Diagn Ther* 2020;**10**:1313–1324.

19. D'Ascenzo F, De Filippo O, Gallone G, Mittone G, Deriu MA, Iannaccone M, *et al.* Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets. *Lancet* 2021;**397**:199–207.
20. Zack CJ, Senecal C, Kinar Y, Metzger Y, Bar-Sinai Y, Widmer RJ, *et al.* Leveraging machine learning techniques to forecast patient prognosis after percutaneous coronary intervention. *JACC Cardiovasc Interv* 2019;**12**:1304–1311.
21. Hsieh MH, Lin SY, Lin CL, Hsieh MJ, Hsu WH, Ju SW, *et al.* A fitting machine learning prediction model for short-term mortality following percutaneous catheterization intervention: a nationwide population-based study. *Ann Transl Med* 2019;**7**:732.
22. Al'Aref SJ, Singh G, van Rosendael AR, Kolli KK, Ma X, Maliakal G, *et al.* Determinants of in-hospital mortality after percutaneous coronary intervention: a machine learning approach. *J Am Heart Assoc* 2019;**8**:e011160.
23. Deo RC. Machine learning in medicine. *Circulation* 2015;**132**:1920–1930.
24. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950;**3**:32–35.
25. Cimci M, Polad J, Mamas MA, Iniguez-Romo A, Chevalier B, Abhaichand R, *et al.* Outcomes and regional differences in practice in a worldwide coronary stent registry. *Heart* 2022;**108**:1310–1318.
26. Mann CJH. *Clinical Prediction Models: A Practical Approach to Development, Validation and Updating*, Vol. 38. Leeds: Emerald Publishing Limited; 2009.
27. 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**:128–138.
28. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;**44**:837–845.
29. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Proc Lett* 2014;**21**:1389–1393.
30. Kovacic JC, Limaye AM, Sartori S, Lee P, Patel R, Chandela S, *et al.* Comparison of six risk scores in patients with triple vessel coronary artery disease undergoing PCI: competing factors influence mortality, myocardial infarction, and target lesion revascularization. *Catheter Cardiovasc Interv* 2013;**82**:855–868.
31. Garg S, Serruys PW, Silber S, Wykrzykowska J, van Geuns RJ, Richardt G, *et al.* The prognostic utility of the SYNTAX score on 1-year outcomes after revascularization with zotarolimus- and everolimus-eluting stents: a substudy of the RESOLUTE All Comers Trial. *JACC Cardiovasc Interv* 2011;**4**:432–441.
32. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. *Circ Cardiovasc Qual Outcomes* 2020;**13**:e006556.