XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study

Clinical Medicine Insights: Cardiology
Volume 16: 1–6
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11795468221133611
SAGE

Alexander Moore[1] and Max Bell[2,3]

[1]Head of Data Science at Managed Self Limited, London, England, UK. [2]Perioperative Medicine and Intensive Care, Karolinska University Hospital, Stockholm, Sweden. [3]Section of Anaesthesiology and Intensive Care Medicine, Department of Physiology, Karolinska Institutet, Stockholm, Sweden.

ABSTRACT: We wanted to assess if "Explainable AI" in the form of extreme gradient boosting (XGBoost) could outperform traditional logistic regression in predicting myocardial infarction (MI) in a large cohort. Two machine learning methods, XGBoost and logistic regression, were compared in predicting risk of MI. The UK Biobank is a population-based prospective cohort including 502 506 volunteers with active consent, aged 40 to 69 years at recruitment from 2006 to 2010. These subjects were followed until end of 2019 and the primary outcome was myocardial infarction. Both models were trained using 90% of the cohort. The remaining 10% was used as a test set. Both models were equally precise, but the regression model classified more of the healthy class correctly. XGBoost was more accurate in identifying individuals who later suffered a myocardial infarction. Receiver operator characteristic (ROC) scores are class size invariant. In this metric XGBoost outperformed the logistic regression model, with ROC scores of 0.86 (accuracy 0.75 (CI ±0.00379) and 0.77 (accuracy 0.77 (CI ± 0.00369) respectively. Secondly, we demonstrate how SHAPley values can be used to visualize and interpret the predictions made by XGBoost models, both for the cohort test set and for individuals. The XGBoost machine learning model shows very promising results in evaluating risk of MI in a large and diverse population. This model can be used, and visualized, both for individual assessments and in larger cohorts. The predictions made by the XGBoost models, points toward a future where "Explainable AI" may help to bridge the gap between medicine and data science.

KEYWORDS: Myocardial infarction, machine learning, artificial intelligence, cohort study

RECEIVED: January 19, 2022. **ACCEPTED:** August 26, 2022.

TYPE: Original Research

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Max Bell, Perioperative Medicine and Intensive Care, Karolinska University Hospital, Norrbacka S2:05, Stockholm, 17176, Sweden. Email: max.bell@regionstockholm.se

Key Messages

- **What is already known about this subject?** Artificial Intelligence models have been used to predict myocardial infarction in rather small patient groups, with under 1000 individuals.
- **What does this study add?** In this UK Biobank prospective cohort study including over 500 000 persons, we demonstrate that extreme gradient boosting (XGBoost), a machine learning model, outperforms a logistic regression model. The XGBoost had a receiver operator characteristic score of 0.86 as compared to the logistic regression score of 0.77. Moreover, the XGBoost model is transparent, meaning we know what variables affect the output.
- **How might this impact on clinical practice:** XGBoost, a transparent "Explainable AI" model, may bridge the gap between medicine and data science: it can be used to predict risks of myocardial infarction in both groups and individuals.

Introduction

In adults, the prevalence of cardiovascular disease (CVD) is close to 50% and is a major cause of morbidity and mortality worldwide.[1] Notably, this holds true even though annual death rates attributable to coronary heart disease (CHD) declined 31.8% from 2006 to 2016.[1] Similarly, a study of over 45 000 hospitalizations for myocardial infarction (MI) showed a 24% decrease from 2000 to 2008.[2]

This substantial decline in the incidence of coronary heart disease is partly driven by modifiable risk factors; changes in cholesterol, improved blood pressure control, decreased smoking and increased physical activity.[3] Correctly focusing primary preventive efforts, such as lifestyle changes, and identifying the patients who stand to benefit the most from preventative medication requires an understanding of the risk factors for the development of CVD and MI. This was clearly shown in the ASPREE study, where over 19 000 elderly patients were randomly assigned to oral aspirin or placebo.[4] No decrease in CVD was seen in the aspirin group, on the contrary more severe bleeding events were observed, highlighting the benefits of individualized risk stratification.

Multiple risk factor models exist. Following the Framingham Risk Score, the first large study of CVD risk factors,[5] risk assessment tools from the American and European cardiology societies have been launched; the EURO Score[6] and the American College of Cardiology (ACC)/American Heart Association (AHA) Heart Risk Score.[7] These scores include classic risk factors such as age, gender, diabetes, smoking, and blood pressure. To increase accuracy and individualization for cardiovascular risk prediction, biomarkers, such as blood lipids have been added to the Framingham Risk Score and to the EURO Score.[6] Recently, high-sensitivity cardiac troponins have been shown not only to be able to detect acute MI, but to function as predictors of adverse outcomes.[8-10] The Norwegian HUNT study added high-sensitivity C-reactive protein, troponins, and cholesterol to above mentioned classic risk factors.[11] Specifically, troponins provided improved prognostic information.

With or without adding biomarkers, an alternative path to the improvement of predictive properties is opened by the development of machine learning (ML). Recently, a machine learning algorithm predicted the likelihood of acute MI, incorporating age, sex, with paired high-sensitivity cardiac troponin I outperformed the European Society of Cardiology 0/3-hour pathway.[12] This ML algorithm was trained on 3013 patients and tested on 7998 patients with suspected myocardial infarction.[12]

The present study aims to test a machine learning model, XGBoost[13] on predicting myocardial infarction in a population-based cohort of over 500 000 subjects, the UK Biobank.[14] We hypothesized that the ML algorithm would outperform logistic regression models.

Methods

Study design

This was a prospective observational study, testing machine learning and traditional logistic regression models, described in detail below. The North West Multi-Center Research Ethics Committee approved the UK Biobank study. All participants provided written informed consent to participate. This research has been conducted using the UK Biobank Resource under Application Number 54045.

Data source, the UK Biobank

UK (United Kingdom) Biobank is a large, population-based prospective study, established to allow detailed investigations of the genetic and non-genetic determinants of the diseases of middle and old age.[15,16] The 500 000 participants were assessed from 2006 to 2010 in 22 assessment centers throughout the UK, the range of settings provided socio-economic and ethnic heterogeneity and an urban–rural mix. The wide distribution of all exposures allows generalizable associations between baseline characteristics and health outcomes. The assessments consisted of electronic signed consent; a self-completed touch-screen questionnaire; a computer-assisted interview; physical and functional measures and the collection of blood, urine, and saliva.

Patient and public involvement statement

The investigation was conducted using the UK Biobank resource. Details of patient and public involvement in the UK Biobank are available online. No patients were involved in the research question or the outcomes, nor were they involved in design the study. No patients advised or interpreted results. There are no specific plans to disseminate the results of the present project to UK Biobank participants, but the UK Biobank disseminates key findings from projects on its website.

Primary outcome

Our models were trained to predict myocardial infarction. This was extracted from data-field 6150 in the UK Biobank ("vascular/heart problems diagnosed by a doctor"). There have been 3 separate instances where this question was posed to the participants (initially between 2006 and 2010, the first follow up was between 2012 and 2013 and the final follow up was in 2014). Any participant who selected "Heart Attack" (11 849 participants) when answering this question in any of these instances was included in our positive class. More details on how these features are defined can be found through the UK Biobank[15] and details regarding cardiovascular outcomes are discussed in a study from 2019.[17]

Logistic regression

Logistic regression models are commonly used for risk factor analysis.[18] They use a logistic function to model a binary dependent variable. A linear combination of predictors (input features) is used to calculate the probability that a given input belongs to one of the 2 classes. As features are combined linearly, explaining the predictions made by logistic regression models is straightforward. Moreover, logistic regression models have been shown to produce classification accuracies that are comparable to state of the art machine learning techniques.[19] In this study we implement a linear model for regularized logistic regression.

XGBoost

Extreme Gradient Boost (XGBoost) is a powerful ensemble learning method, well suited to tabular datasets. Ensemble learning methods aggregate predictions of many individually trained classifiers, the combined prediction is typically more powerful than an individual classifier.[20] In the case of XGBoost the ensemble's constituent classifiers are decision trees.

Decision trees are graphical models, where distinct nodes are connected by branches, they are powerful tools for classification.[21] Each node represents a condition that is used to split the data. Data will pass along different branches dependent on whether it satisfies the condition at a particular node. Nodes at the bottom of a decision tree are known as leaf nodes. During training a tree adjusts the decision rules of its nodes in order to maximally separate the training data.[22] The most common class of the training data that terminate in a leaf node is assigned to that node after training. In order to classify new data, it must first be passed down the decision tree. Its class is then determined by the class assigned to the leaf node it terminates in.

Table 1. Model comparison.

	XGBOOST		LOGISTIC REGRESSION	
	ROC score	0.86	ROC score	0.77
	Accuracy	0.75	Accuracy	0.77
CLASS:	PRECISION	RECALL	PRECISION	RECALL
0.0 (Healthy)	0.99	0.75	0.99	0.77
1.0 (MI)	0.07	0.81	0.07	0.78

XGBoost combines decision trees using a process known as gradient boosting.[13] Boosting methods build classifiers sequentially, such that the error from one classifier is passed on to the next. By training decision trees on the gradient of the loss produced by the previous tree, XGBoost is capable of producing prediction accuracies that match many state of the art supervised learning techniques, including neural networks.[23] In machine learning hyperparameters configure various aspects of an algorithm, they must be set before training and they can have a big impact on performance.[24] There are several hyperparameters controlling XGBoost, for example the number of decision trees to include must be specified. To maximize the accuracy of XGBoost these hyperparameters must be optimized. In order to maximize the utility of the training set, hyperparameter optimization typically uses a process known as cross-validation. Cross-validation randomly partitions the training set to produce a small validation set, used to measure performance. In this way, the effect of changing a hyperparameter can be quantified and the optimal value can be selected. Multiple rounds of cross-validation, with different partitions, are common.[21]

The following hyperparameters are typically optimized before training an XGBoost model: the number of estimators (decision trees), the max depth of a given decision tree, the minimum child weight in a decision tree, the minimum loss reduction required to make a partition (gamma), the number of columns to be subsampled when constructing a tree, and a regularization parameter (alpha).

Shapely values (SHAP values) originated as a concept in 1951 from cooperative game theory.[25] More recently they have been used as a tool for interpreting ensemble tree models. They facilitate the explanation of highly non-linear models, such as XGBoost, breaking down the impact of input features on prediction.[26] The SHAP value of a feature is calculated using the change in a model's output if that feature's value was replaced with a baseline value.[27] Consequently, considering the sum off all SHAP values is equivalent to considering the overall difference between a model's prediction and the baseline.[27] In addition to breaking down the importance of an individual's input features, SHAP values can explain the global impact of features across a population. SHAP values can be visualized in a number of ways.[27]

Model training and validation

When validating supervised learning algorithms, it is essential that a portion of the data (~10%) is set aside and not used in training. This data is known as the test set. An algorithm's performance can be assessed by measuring its ability to correctly map inputs to outputs in the test set. Measuring performance on data used during training can result in overfitting or selection biases.[28]

Crude accuracy metrics, the total number of correct classifications, can be misleading when assessing the performance of a classification model. Consider a classification task where 0.01% of the population are high risk, a model could achieve 99.99% accuracy by classifying all individuals as low risk. Receiver operator characteristic (ROC) scores reflect the area underneath a curve obtained by plotting the true positive rate against the false positive rate. They are a more useful measure of performance in classification tasks with unbalanced classes.

Results

Model performance

Results comparing XGBoost with a logistic regression model are presented in Table 1. The models were trained using 90% of the cohort (~450 000 participants). The remaining 10% (~50 000 participants) was set aside for use as a test set. The scores in Table 1 reflect the performance of both models when classifying the test set.

Both models are equally precise but deviate when it comes to recall. The logistic regression model has a slightly higher recall for class 0.0 (the healthy class) whereas XGBoost has a higher recall for class 1.0 (the MI class). This means that the logistic regression model classified more of the healthy class correctly, whilst XGBoost was more accurate when it came to identifying individuals who would go on to suffer a myocardial infarction (MI).

The above explains why the logistic regression model has a slightly higher crude accuracy score, as the healthy is significantly larger than MI class. A model that is more accurately recalling participants in the larger class is likely to end up with a larger total number of correct classifications. The confidence intervals were extremely tight due to the large test set (accuracy XGBoost: 0.75 (CI ±0.00379) and Logistic regression: 0.77 (CI ±0.00369).

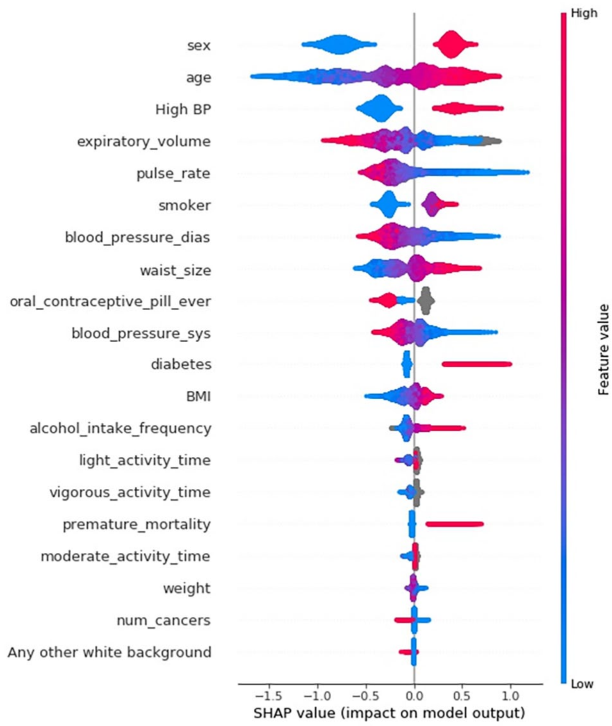


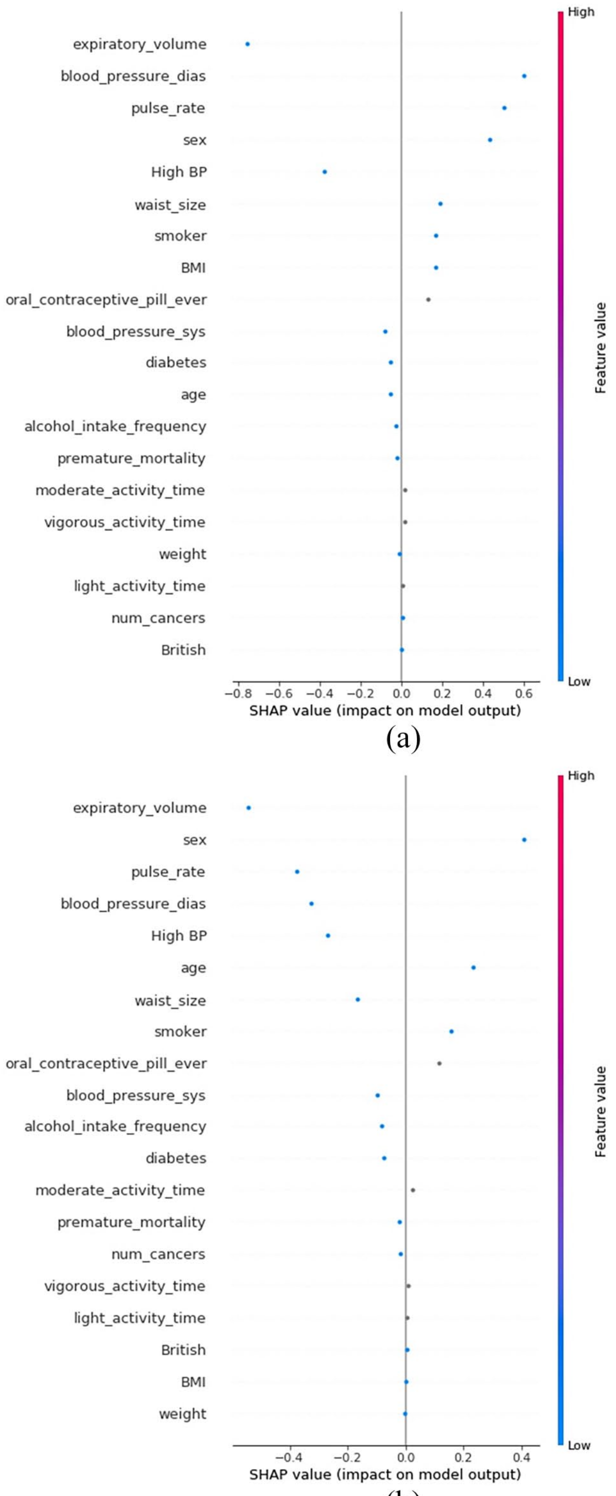
Figure 1. SHAP values extracted from a XGBoost model trained to predict MI (~50 000 participants).

ROC scores are class size invariant. In this metric XGBoost scores significantly higher than the logistic regression model (+0.09).

Model visualization

SHAP values from an XGBoost model trained to predict MI using the UK Biobank cohort are shown in Figure 1. These were calculated for the 50 000 individuals in the test set. Each of these individuals is represented by a single data point. Here we see that having a high feature value for sex (being male) has a positive SHAP value, which means that being male increased the likelihood of being classified as class 1.0 (MI class). Conversely, a low feature value for sex (being female) has a negative impact on the model, meaning being female contributed to being classified as class 0.0 (healthy class). Features in Figure 1 are ordered based on their cumulative effect on model output.

SHAP values can also be used to unpack individual predictions made by XGBoost. In Figure 2, the SHAP values for 2 individuals are displayed, one who was classified as MI class (2i) the other was classified as healthy (2ii). Large positive SHAP values for diastolic blood pressure, and pulse rate, seen in Figure 2a, contributed to one individual being classified as MI class. These features have large negative SHAP values for the other individual, seen in Figure 2b, meaning they strongly contributed to the individual being classified as healthy. Once again, features are ordered based on their impact on model output. The ordering of features is different in Figure 2a and b as the relative importance of input features can vary between individuals.



(a)

(b)

Figure 2. (a) SHAP values for an individual classified as MI class and (b) SHAP values for an individual classified as healthy class.

A SHAP dependence plot represents the impact of 2 input variables on classification. Figure 3 explores the impact of waist size and sex on MI. Each data point represents an individual in the test set, color coded by their sex. Waist size is plotted on the x-axis and the SHAP values (or model impact) are plotted on the y-axis. From inspecting Figure 3 we can see that most of the individuals with particularly small waists are

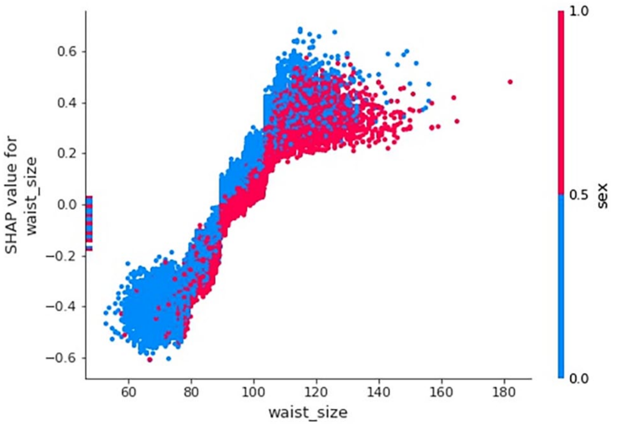


Figure 3. A SHAP dependence exploring the relationship between waist size, sex, and myocardial infarction.

female, this is associated with negative SHAP values. In these cases, an individual's waist size made it more likely that they would be classified as healthy. As waist size increases SHAP values increase, this relationship continues linearly to a large cluster of individuals with positive SHAP values. In this cluster waist size made it more likely that an individual would be classified as belonging to the MI class. The positive correlation between waist size and SHAP value does not continue beyond the large cluster.

The impact of having an extremely large waist is different for men and women. The men with the largest waists have SHAP values that are comparable with individuals in the cluster, implying their waist size had no additional impact on classification. On the other hand, women in the large cluster tend to have the highest SHAP values; waist size is likely to play a bigger role in classification for these women.

Discussion

In this large cohort, the XGBoost machine learning model outperformed a multivariable logistic regression model in assessing risk of myocardial infarction. Importantly, our XGBoost model also allows for personalized risk estimations.

Previous attempts at using ML models exist in the field of cardiology. In 2005, Green et al[29] trained artificial neural network (ANN) ensembles and logistic regression models on data from 634 patients presenting in the emergency department with chest pain. The ANN model performed slightly better than the logistic regression, but this was a single center study with limited power. Another small study, using data from 310 patients, tested ANN-algorithms for early diagnosis of acute myocardial infarction and prediction of size of infarction in patients presenting with chest pain.[30] The authors conclude that specifically designed ANN-algorithms allow early prediction of major AMI size and could be used for rapid assessment of these patients. Numerous studies outside of cardiology have tested machine learning models with varying degrees of success. Deep learning systems had high sensitivity and specificity for identifying diabetic retinopathy and related eye diseases.[31] Using machine learning technology to correctly classify indeterminate pulmonary nodules increased the reclassification performance, as compared to that of existing risk models.[32] In contrast, ML algorithms did not outperform traditional regression approaches in a low-dimensional setting for outcome prediction after traumatic brain injury.[33] In 2021 D'Ascenzo et al[34] used a machine learning-based approach in identifying predictors of events after an acute coronary syndrome, showing it to be feasible and effective.

The findings in the present study have multiple implications. First, we show that ML-algorithms allow prediction of MI-risk in a large population. In future studies, when more high-resolution data is added, it is likely that the machine learning models will outperform logistic regression by a wider margin. Secondly, the XGBoost model predicts individual risk: this allows for patients with elevated risk patterns to aim for targeted and tailor-made life-style changes and in some cases medical treatments. In concert, these personalized interventions could decrease cardiac morbidity and possibly even add life years. As seen in the results section; the logistic regression model had a slightly higher crude accuracy score, as the healthy participants were significantly more common than individuals who would go on to suffer a myocardial infarction. Any model that is more accurately recalling participants in the larger class is likely to end up with a total number of correct classifications. In contrast, ROC scores are not affected by variable class sizes. In this metric XGBoost scored significantly higher than the logistic regression model. This means the risk of false negatives (a person at high risk of MI is classified as safe) is minimized using XGBoost; from an individual and epidemiological standpoint the cost of false negatives is much higher than false positives.

Our study has strengths and limitations. We had access to a large high-resolution longitudinal dataset from the UK Biobank. We trained the XGBoost ML model on 450 000 subjects and tested its performance on 50 000 individuals; compared to previous machine learning cardiology studies with less than 1000 patients, this adds to both model performance and generalizability. Comparing traditional logistic regression with ML models adds to transparency regarding the utility of these novel techniques. Said transparency is further achieved by choosing XGBoost—this is not a "black box" ML system; we can follow how the data flows through the system. Moreover, we demonstrate how XGBoost lowers misclassification events of patients at risk of myocardial infarction. A potential limitation to any MI-study is how the event is classified and recorded. If we had full biomarker and ECG data on all UK Biobank participants, we would likely have many more cases; that would enable more accurate machine learning and logistic regression predictions. Moreover, we acknowledge that our logistic regression model could be improved by adding interaction terms and accounting for non-linearity with regards to certain inputs. It is certainly possible to create more precise logistic models, but this requires active input from both statisticians and physicians with experience in whatever disease one is modeling. A benefit of the XGBoost model approach is that very good predictive properties are made possible automatically. The generalizability is difficult to assess, studies across other datasets, preferably from multiple countries would be valuable.

In conclusion the XGBoost machine learning model shows very promising results in evaluating risk of MI in a large and diverse population. This model can be used both for individual assessments and in larger cohorts.

Acknowledgments

We wish to thank the UK Biobank for making this research possible.

Contributorship Statement

Author Contributions: Mr Moore had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: Moore and Bell.
Acquisition, analysis, or interpretation of data: Moore and Bell.
Drafting of the manuscript: Moore and Bell.
Critical revision of the manuscript for important intellectual content: Moore and Bell.
Statistical analysis: Moore
Obtained funding: N/A.
Supervision: Bell.

Data Availability Statement

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

ORCID iD

Max Bell https://orcid.org/0000-0001-7464-0324

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Benjamin EJ, Muntner P, Alonso A, et al. Heart disease and stroke statistics-2019 update: a report from the American Heart Association. 2019;139:e56-e528.
2. Yeh RW, Sidney S, Chandra M, Sorel M, Selby JV, Go AS. Population trends in the incidence and outcomes of acute myocardial infarction. *New Engl J Med*. 2010;362:2155-2165.
3. Mannsverk J, Wilsgaard T, Mathiesen EB, et al. Trends in modifiable risk factors are associated with declining incidence of hospitalized and nonhospitalized acute coronary heart disease in a population. *Circulation*. 2016; 133:74-81.
4. McNeil JJ, Wolfe R, Woods RL, et al. Effect of aspirin on cardiovascular events and bleeding in the healthy elderly. *N Engl J Med*. 2018;379:1509-1518.
5. Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J 3rd. Factors of risk in the development of coronary heart disease--six year follow-up experience. The Framingham Study. *Ann Intern Med*. 1961;55:33-50.
6. Piepoli MF, Hoes AW, Agewall S, et al. 2016 European guidelines on cardiovascular disease prevention in clinical practice: the sixth joint task force of the European Society of Cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of 10 societies and by invited experts) developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). *Eur Heart J*. 2016;37:2315-2381.
7. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on clinical practice guidelines. *J Am Coll Cardiol*. 2019;74:1376-1414.
8. Blankenberg S, Salomaa V, Makarova N, et al. BiomarCaRE Investigators. Troponin I and cardiovascular risk prediction in the general population: the BiomarCaRE consortium. *Eur Heart J*. 2016;37:2428-2437.
9. Ford I, Shah AS, Zhang R, et al. High-sensitivity cardiac troponin, statin therapy, and risk of coronary heart disease. *J Am Coll Cardiol*. 2016;68:2719-2728.
10. Neumann JT, Twerenbold R, Blankenberg S. Application of high-sensitivity troponin in suspected myocardial infarction. Reply. *New Engl J Med*. 2019;381:2483.
11. Sigurdardottir FD, Lyngbakken MN, Holmen OL, et al. Relative prognostic value of cardiac troponin I and C-reactive protein in the general population (from the nord-Trøndelag Health [HUNT] study). *Am J Cardiol*. 2018;121:949-955.
12. Than MP, Pickering JW, Sandoval Y, et al. Machine learning to predict the likelihood of acute myocardial infarction. *Circulation*. 2019;140:899-909.
13. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Paper presented at: The 22nd ACM SIGKDD International Conference; August 3, 2016. *arXiv:1603.02754*.
14. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12:e1001779.
15. Centre UBC. UK Biobank: protocol for a large-scale prospective epidemiological resource. 2020
16. Ollier W, Sprosen T, Peakman T. UK Biobank: from concept to reality. *Pharmacogenomics*. 2005;6:639-646.
17. Littlejohns TJ, Sudlow C, Allen NE, Collins R. UK Biobank: opportunities for cardiovascular research. *Eur Heart J*. 2019;40:1158-1166.
18. Reed P, Wu Y. Logistic regression for risk factor modelling in stuttering research. *J Fluency Disord*. 2013;38:88-101.
19. Tolles J, Meurer WJ. Logistic regression: relating patient characteristics to outcomes. *JAMA*. 2016;316:533-534.
20. Opitz D, Maclin R. Popular ensemble methods: an empirical study. *J Artif Intell Res*. 1999;11:169-198.
21. Bishop C. *Pattern Recognition and Machine Learning*. Springer; 2006:140-155.
22. Dev VA, Eden MR. Gradient boosted decision trees for lithology classification. In: Muñoz SG, Laird CD, Realff MJ, eds. *Computer Aided Chemical Engineering*. Elsevier; 2019;113-118.
23. Fauzan M, Murfi H. The accuracy of XGBoost for insurance claim prediction. *Int J Adv Soft Comput Appl*. 2018;10:159-171.
24. Claesen M, De Moor B. Hyperparameter search in machine learning. 2015.
25. Shapley L. A value for n-person games. In: Kuhn H, Tucker A, eds. *Contributions to the Theory of Games II*. Princeton University Press; 1953;307-317.
26. towardsdatascience.com. Interpretable machine learning with XGBoost. 2018. Accessed January 1, 2022. https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27
27. Dan Becker. Shap-values. 2022. Accessed January 1, 2022. https://www.kaggle.com/dansbecker/shap-values
28. Cawley G, Talbot N. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res*.2010;11:2079-2107.
29. Green M, Björk J, Forberg J, Ekelund U, Edenbrandt L, Ohlsson M. Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room. *Artif Intell Med*. 2006;38:305-318.
30. Eggers KM, Ellenius J, Dellborg M, et al. Artificial neural network algorithms for early diagnosis of acute myocardial infarction and prediction of infarct size in chest pain patients. *Int J Cardiol*. 2007;114:366-374.
31. Ting DSW, Cheung CYL, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318: 2211-2223.
32. Massion PP, Antic S, Ather S, et al. Assessing the accuracy of a deep learning method to risk stratify indeterminate pulmonary nodules. *Am J Respir Crit Care Med*. 2020;202:241-249.
33. Gravesteijn BY, Nieboer D, Ercole A, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *J Clin Epidemiol*. 2020;122:95-107.
34. D'Ascenzo F, De Filippo O, Gallone G, 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.