# Actionable absolute risk prediction of atherosclerotic cardiovascular disease: a behavior-management approach based on data from 464,547 UK Biobank participants

Ajay Kesar[1*], Adel Baluch[1], Omer Barber[1], Henry Hoffmann[1], Milan Jovanovic[1], Daniel Renz[1], Bernard Leon Stopak[1], Paul Wicks[1], Stephen Gilbert[1,2]

[1] Ada Health GmbH, Berlin, Germany

[2] EKFZ for Digital Health, University Hospital Carl Gustav Carus Dresden, Technische Universität Dresden, Dresden, Germany

* Corresponding author

E-mail: science@ada.com

# Abstract

Cardiovascular diseases (CVDs) are the primary cause of all global death. Timely and accurate identification of people at risk of developing an atherosclerotic CVD and its sequelae, via risk prediction model, is a central pillar of preventive cardiology. However, currently available models only consider a limited set of risk factors and outcomes, do not focus on providing actionable advice to individuals based on their holistic medical state and lifestyle, are often not interpretable, were built with small cohort sizes or are based on lifestyle data from the 1960s, e.g. the Framingham model. The risk of developing atherosclerotic CVDs is heavily lifestyle dependent, potentially making a high percentage of occurrences preventable. Providing actionable and accurate risk prediction tools to the public could assist in atherosclerotic CVD prevention. We developed a benchmarking pipeline to find the best set of data preprocessing and algorithms to predict absolute 10-year atherosclerotic CVD risk. Based on the data of 464,547 UK Biobank participants without atherosclerotic CVD at baseline, we used a comprehensive set of 203 consolidated risk factors associated with atherosclerosis and its sequelae (e.g. heart failure).

Our two best performing absolute atherosclerotic risk prediction models provided higher performance than Framingham and QRisk3. Using a subset of 25 risk factors identified with feature selection, our reduced model achieves similar performance while being less complex. Further, it is interpretable, actionable and highly generalizable. The model could be incorporated into clinical practice and could allow continuous personalized predictions with automated intervention suggestions.

# Introduction

Cardiovascular diseases (CVDs) are the number one cause of all global death (1,2). In 2016, 17.9 million people died of CVDs alone, accounting for 31% of all global deaths (1). The direct costs of CVDs in the US for 2010 were $272.5 billion whereas indirect costs were $171.7 billion and are expected to increase to $818.1 and $275.8 billion in 2030 respectively (3,4). Atherosclerosis alone is responsible for 1.3% of all hospital stays with costs of $9 billion per year, while all atherosclerosis-related diseases amount to $43.5 billion of total hospital costs annually (5). Individually, patients with CVD incur more than twice the medical costs of age- and sex-matched patients without CVD, largely because of the increased likelihood of subsequent hospitalizations. The greatest differences in total CVD costs usually occur when comparing patients with and without a secondary CVD hospitalization (6).

All current guidelines on the prevention of CVD in clinical practice recommend the assessment of total CVD risk since atherosclerosis is usually the product of a number of risk factors (7,8) and in recent years these guidelines have evolved to focus on the absolute risk of disease as opposed to relative risk (7–10). Clinician tools for CVD risk estimation must enable rapid and accurate estimation of an individual patient's absolute CVD risk (7), or for opportunistic screening of high-risk patients from relevant populations (11). Screening is the identification of unrecognized disease or risk of disease in individuals without symptoms. In addition to opportunistic screening, which is carried out without a predefined strategy (e.g. when the individual is consulting a general practitioner (GP) for some other reason), tools can be used for systematic screening, which is centrally organised strategic screening in the general population or in targeted subpopulations, such as subjects with a family history of premature CVD or familial hyperlipidaemia (7). There is ongoing debate on the role of systematic centralised population based screening in CVD (10,12), one reason for this being the tendency for increased use of burdensome diagnostic testing following the use of risk based screening

74 tools(10)(13). A relatively new area of screening is self-screening, carried out by proactive
75 individuals, using smartphone or smartwatch app based screening tools, which may use built in
76 app-linked sensors, or screening chat-bots (14–16). There is public demand for reliable,
77 actionable, explainable and usable health information tools (17), including for disease
78 screening.
79
80 The risk to build up atherosclerotic plaque varies and is determined by multiple factors such as
81 genetics, environment and lifestyle (11,18–21). With genetics being unmodifiable and the
82 environment being difficult to change, the risk of developing atherosclerotic plaque can be
83 reduced based on an individual's lifestyle which is modifiable (19,20).
84 Thus, atherosclerotic CVD is actionable and preventable by addressing behavioral risk factors,
85 such as smoking, physical activity and nutrition (1,11,19,20).
86
87 Most diseases, including atherosclerotic CVDs, have a complex pathophysiology that involves
88 multiple interacting molecular systems, making it insufficient to look only at an isolated biological
89 pathway or a subset of markers to predict disease risk (22). A precision medicine based
90 approach is required, where multiple biological layers are considered (i.e., 'multi-omics'),
91 alongside clinical and lifestyle data (22). Such an approach has the potential to capture all
92 important interactions or correlations detected between molecules in different biological layers,
93 providing a holistic understanding of an individual's current health status and enabling the
94 quantification of an individual's absolute risk of atherosclerotic CVDs (23,24).
95
96 Previous studies in this area use an outdated or very limited set of risk factors and outcomes for
97 their analysis (7,25). In recent years, the knowledge of behavioral risk factors and of the
98 pathophysiology of atherosclerotic CVDs have advanced tremendously (11,25). Current
99 absolute risk prediction models have limited predictive capability as they have not been trained

on all possible atherosclerotic CVD outcomes (26–28), or they include outcomes which are unmodifiable such as those related to pregnancy, accidents, or congenital factors (28). Both SCORE (Systematic COronary Risk Evaluation) and SCORE2 (29,30), are models for predicting relative CVD risk, whereas we focus on predicting absolute CVD risk, which is why we chose to omit those models from our analysis. Another related investigation, which also used the UK Biobank (UKB) dataset, developed multiple Cox Proportional Hazard models for 10-year CVD risk prediction, with a reduced version requiring 47 risk factors and another version disregarding all cholesterol risk factors as well as systolic blood pressure, in order to provide a simple approach for risk prediction in remote settings with limited testing resources (31). However, survival models such as the proportional hazard model, are not designed to provide absolute risk estimates for individual patients.

Machine learning (ML) based approaches have many advantages, such as superior performance, being able to identify complex non-linear patterns, the ability to encode diverse and high dimensional data types, being more stable to outliers, allowing continuous model updates, versatility for different domains and scalability (32–35). However, classic disadvantages of ML based approaches are their lack of interpretability, risk for inherent bias due to the used data, difficulty to acquire physician adoption, explaining to physicians why a new risk model might be superior to existing ones, with all of these hindering widespread adoption of ML based risk prediction models (35,36). One example for ML based CVD risk prediction is the AutoPrognosis based approach, where an ensemble of multiple ML pipelines has also been applied on the UK Biobank dataset for 5-year CVD risk prediction (28). Further, using a purely ML driven approach can lead to a model that requires too many risk factors to compute risk, which is infeasible for routine clinical check-ups. Another disadvantage of purely data-driven approaches is the inclusion of risk factors which might show strong

correlations but are unrelated to the pathophysiology of CVDs or are not actionable, making them inapplicable in a clinical setting or as an actionable self-management tool (28).

The aim of this study was to use a large-data ML approach to develop an actionable absolute risk prediction tool which takes into account the holistic health of an individual and has a focus on behavioral risk factors relating to atherosclerotic CVD outcomes. Our goal was to have a highly holistic understanding of an individual's current health status, to better quantify their risk of atherosclerotic CVDs and to provide actionable advice. We aimed to do this by taking multiple biological layers into account, which are: (i) multi-omics data from blood samples (e.g. lipidome and proteome); (ii) family history (e.g. genome), (iii) lifestyle data, (iv) clinical data and (v) environmental data; along with (vi) an extensive set of risk factors and outcomes.

We used data from 464,547 participants of the UK Biobank study who did not have atherosclerotic CVD at baseline. We created an automated pipeline to benchmark risk prediction classifier algorithms against each other, then evaluated their predictive performances in the overall population and tested the generalizability of the top-performing classifiers through retraining and testing on different sub-populations. We explored the clinical implications of the proposed classifiers, with a focus on the top-performing models. This study does not focus on the algorithmic aspects of the utilized classifiers.
Methodological details on the utilized classifiers can be found in the open-source documentation of the respective algorithms of the scikit-learn (37) and xgboost (38) libraries and in the supporting information (S4 Table).

# Materials and Methods

## Study design and participants

The UK Biobank is a long-term prospective large-scale biomedical database including over 500,000 participants aged 40-69 years (when recruited between 2006 and 2010). The database is globally accessible to approved researchers undertaking research into the most common and life-threatening diseases and continuously collects phenotypic and genotypic data about its participants, including data from questionnaires, physical measures, blood, urine and saliva samples, lifestyle data (39). This data is further linked to each participant's health-related records, accelerometry, multimodal imaging, genome-wide genotyping and longitudinal follow-up data for a wide range of health-related outcomes (39,40). The UK Biobank study protocol is available online (41).

The North West Multi-centre Research Ethics Committee approved the UK Biobank study and all participants provided written informed consent prior to study enrollment. Our research is covered by the UK Biobank's Generic Research Tissue Bank (RTB) Approval and was approved by the UK Biobank Access Management Team (42).

We excluded participants with atherosclerotic CVDs present before or during baseline, participants who chose to leave the UKB study and participants who were lost due to various reasons. The resulting cohort consisted of 464,547 participants. The last available date of participant follow-up was March 5th, 2020.

## Risk factor definition

We curated a list of all generally known risk factors and outcomes for atheroscelortic CVDs from the medical literature and from validated risk prediction models. This preliminary list of risk factors was reduced through curation to focus on those factors that were clearly involved in the pathophysiology of atherosclerosis and those that are modifiable through behavioral change. The curation was carried out by three medical doctors with experience in diagnosing or scientifically modelling cardiovascular diseases. We consolidated all relevant UKB columns into

203 risk factors and grouped them into six categories: demographics (e.g. age, biological sex, ethnicity), biomarkers (e.g. cholesterol, glucose, blood pressure, heart rate), lifestyle (e.g. alcohol consumption, smoking, physical activity, sleep, social visits), environment (e.g. exposure to tobacco smoke, work and housing and other socio-economic related factors), genetics (e.g. family history of cvd, stroke, diabetes, high cholesterol, high blood pressure) and comorbidities (e.g. heart arrhythmias, diabetes, acute & chronic kidney injury, migraines, rheumatoid arthritis, systemic lupus erythematosus, severe mental illnesses (schizophrenia, bipolar disorder, depression, psychosis), diagnosis or treatment of erectile dysfunction, atypical antipsychotic medication). A categorized list of all risk factors used in our analysis is provided in the supplementary data (S1 Table).

## Outcome definition

In the same manner as described above, an initial list of atherosclerotic CVDs was further reviewed and curated by the same team of medical doctors. All resulting CVDs of interest are associated with atherosclerotic plaque build-up, are modifiable and relate to the collected risk factors only. Thus, we disregard brain haemorrhages due to accidents and congenital and pregnancy-related CVDs, which are not actionable. The curated list of all ICD-10 and ICD-9 outcomes meeting the above criteria consists of 193 total (125 unique) CVD outcomes, e.g. coronary/ischaemic heart disease, heart attack, angina, stroke, cardiac arrest, congestive heart failure, left ventricular failure, myocardial infarction, aortic valve stenosis, cerebral artery occlusions, nontraumatic haemorrhages. A list with all outcome codes used in our analysis is provided in the supplementary data (S2 Table). An atherosclerotic CVD event was defined as the first occurrence out of the following: any of the atherosclerotic CVD outcome diagnosis codes, also as primary or secondary death cause during the 10-year follow-up period.

### Cohort Follow-up

Follow-up time was set to 10 years as commonly used in other risk models (see table 2 in (7)) and counted from the date of one's initial assessment center visit. Individuals who died from other causes during their follow-up period or had a relevant CVD event past their individual follow-up period, were marked as not having had a relevant CVD event.

## Models used in comparison

**Framingham Risk Score.** The Framingham 10-year CVD absolute risk score is based on the data of the two prospective studies, the Framingham Heart Study and the Framingham offspring study (26). The cohort consists of 8491 participants, with 4522 women and 3969 men who attended a baseline examination between 30 and 74 years of age and were free of CVD. A positive CVD outcome was defined as any of the following: coronary death, myocardial infarction, coronary insufficiency, angina, ischemic stroke, hemorrhagic stroke, transient ischemic attack, peripheral artery disease and heart failure.

Participants were followed-up for 12 years where 1174 participants developed a CVD. Two biological sex-specific risk models were derived, where Body Mass Index (BMI) substitues lipid measurements. The variables used were biological sex, age, total cholesterol, HDL cholesterol, treated and untreated systolic blood pressure, smoking status and diabetes status.

The Framingham risk calculators and model coefficients are publicly available (43). We imputed missing data using simple mean imputation.

**QRisk3.** The QRisk3 10-year CVD absolute risk score is based on a prospective open cohort study using data from general practices (GPs), mortality and hospital records in England (27). The cohort consists of 10.56 million patients between the age of 25 and 84 years, where 75% of the patients were used for training and 25% for validation. Patients with a pre-existing CVD,

226 missing Townsend score or using statins were removed from the baseline. Patients were
227 classified as having a positive CVD outcome when any of the following outcomes was present
228 during follow-up in the GP, hospital or mortality records: coronary heart disease, ischaemic
229 stroke, or transient ischaemic attack. QRisk3 used the following ICD-10 codes: G45 (transient
230 ischaemic attack and related syndromes), I20 (angina pectoris), I21 (acute myocardial
231 infarction), I22 (subsequent myocardial infarction), I23 (complications after myocardial
232 infarction), I24 (other acute ischaemic heart disease), I25 (chronic ischaemic heart disease), I63
233 (cerebral infarction), and I64 (stroke not specified as haemorrhage or infarction). The utilized
234 ICD-9 codes were: 410, 411, 412, 413, 414, 434, and 436. Participants were followed-up for 15
235 years where 363,565 participants of the training set (4,6%) developed a relevant CVD. One
236 biological sex-specific risk model was derived.
237 The risk factors used in the final model were age, ethnicity, deprivation, systolic blood pressure,
238 BMI, total cholesterol/HDL cholesterol ratio, smoking status, family history of coronary heart
239 disease, diabetes status, treated hypertension, rheumatoid arthritis, atrial fibrillation, chronic
240 kidney disease, systolic blood pressure variability, diagnosis of migraine, corticosteroid use,
241 systemic lupus erythematosus, atypical antipsychotic use, diagnosis of severe mental illnesses,
242 diagnosis or treatment of erectile dysfunction.
243 The QRisk3 risk calculator and model coefficients are publicly available (44), built into all major
244 NHS GP systems and included in the national guidelines
245 (https://www.healthcheck.nhs.uk/seecmsfile/?id=1687, accessed 10th November 2021). We
246 imputed missing data using simple mean imputation.
247
248 **Standard linear and ML models.** We compared regularized linear regression (with L1 penalty),
249 random forests and gradient boosting (xgboost implementation) for assessing the highest
250 achievable Area Under the Receiver Operating Characteristic Curve (AUROC) value, which we
251 used for assessing the trade-off between number of features and predictive performance of

several simpler *practical risk predictors,* as determined by an iterative feature elimination procedure outlined below. L1 regularization for logistic regression implements a strong penalty for non-zero feature weights, resulting in a feature selection procedure that discards features that are likely to be non-predictive. Random Forest is an ensemble method that fits many decision trees independently to a subset of the data. We implemented both methods using their scikit-learn library implementation. Finally, we evaluated Extreme Gradient Boosting: Gradient boosting is an ensemble tree-based machine learning method that combines many weak classifiers to produce a stronger one. It sequentially fits a series of classification or regression trees, with each tree created to predict the outcomes misclassified by the previous tree (45). By sequentially predicting residuals of previous trees, the gradient boosting process has a focus on predicting more difficult cases and correcting its own shortcomings. Extreme Gradient Boosting (XGB / XGBoost) is a specific implementation of the gradient boosting process, and uses memory-efficient algorithms to improve computational speed and model performance (38,46). For completeness, we evaluated a number of other standard classifiers, but discarded them due to too high computational complexity or inferior performance so we do not report their performances here: Decision Trees, Voting Classifiers, Multi-Layer Perceptrons with 2 layers and 200 and 150 neurons each (Neural Network), stochastic gradient descent implementing a support vector machine algorithm (47,48), Ada Boost (49,50), Gradient Boosting (45), K Neighbors (51), Quadratic Discriminant Analysis (52) and Gaussian Naive Bayes (37,53).

## Model development and benchmarking using pipeline

We built a benchmarking pipeline for automated and reproducible data extraction, normalization, imputation, model training, tuning of model hyperparameters, classification, documentation and reporting.

We implemented all models using their respective scikit-learn library or xgboost library implementation using the Python programming language (37,38). Details on the used Python libraries and methods are provided in the supplementary data (S3 and S4 Tables). Categorical values were one-hot encoded. Data normalization was performed by removing the mean and scaling to unit variance. Data imputation was performed for all models using a simple mean imputation. The models' hyper-parameters were determined using grid search and stratified k-fold cross validation using 3 folds to avoid overfitting.

Finally, we assessed model performance mainly using the AUROC.

## Iterative feature elimination

We employed an iterative feature elimination procedure based on the regularized logistic regression for finding the best trade-off between predictive performance and number of risk factors, with the aim of creating a risk prediction algorithm that is applicable in the clinical context. We used the standard L1 regularization (also known as Lasso) proposed by (54); it implements a strong penalty on non-zero feature weights of our logistic regression model, resulting in a sparse feature set for prediction.

A logistic regression coefficient value □ can be interpreted as the expected change in log odds of having the outcome per unit change in the feature x_. Therefore, increasing the feature by one unit multiplies the odds of having the outcome by $e^{\beta}$. This means that we can interpret the coefficients as feature importance values in the sense that the feature with the smallest coefficient has the least importance on model predictions. Importantly, this holds only true in the context of the parameters contained in the current model. Thus, we re-estimate the model after each feature elimination round.

In each iteration, we re-estimated the logistic regression model on the remaining parameters, and then discarded all parameters that were set to zero by the L1 regularization; finally, we also

discarded the parameter with the lowest non-zero absolute value.

As an additional step, we created a ranking of the relative feature importance value of each feature by dividing its absolute coefficient weight by the sum of all absolute coefficient weights.

### Statistical analysis

To reduce overfitting, we evaluated the classification performance of all our benchmarked algorithms by using 3-fold stratified cross-validation and measuring the Area Under the Receiver Operating Characteristic Curve. For the cross-validation, we used a training set with 325,182 participants to train and derive our standard linear and ML models and then assessed the AUROC performance on the held-out test set with 139,365 participants using 203 risk factors respectively. We report the AUROC and the 95% confidence intervals (Wilson score intervals) for all models.

### Generalizability

With 442,620 out of the 502,551 patients in the UK Biobank, the cohort has a high proportion (88.1%) of participants with British ethnicity. In an effort to estimate a proxy for out-of-sample generalizability, we re-trained the two best models, XGB and Logistic Regression with L1 regularization, only on whites and tested their performance on a non-white test set. The white-only training set consists of 378,836 participants (81.5%). The non-white test set consists of 85,711 participants (18.5%).

# Results

### Characteristics of the training and test populations

Of 502,551 patients in the UK Biobank, we filtered out 7.6% who already experienced a relevant CVD outcome (during or before baseline) and the participants being lost or who withdrew from

the biobank. This resulted in 464,547 participants who met the inclusion criteria. 28,561 (6.1%) of those participants developed at least one of the relevant CVD outcomes during their 10-year follow-up period. We used a common 70% of the data as a training set and 30% as a hold-out test set. Table 1 shows the overlap of our atherosclerotic CVD outcome definition with the CVD outcome definition used in the related work approach by Alaa et al. (28):

**Table 1. CVD outcomes statistics according to definition in current study and the comparator study definition by Alaa et al.** (28)**.**

| **Statistic measured** | **Number** |
|---|---|
| No. of atherosclerotic CVD outcomes that developed in 10-year follow-up according to definition in current study | 28,561 |
| No. of CVD outcomes that developed in 10-year follow-up according to comparator study definition | 28,242 |
| No. of CVD outcomes after 10-year follow-up that overlap in the current study and comparator study definition | 456,184 out of 464,547 (98%) |
| No. of CVD outcomes identified in the current study but not in comparator studies | 4,341 |
| No. of CVD outcomes included in comporator studies, but not in current study | 4,022 |

## Prediction accuracy

**Comparison of prediction models.** The resulting prediction accuracy of the benchmarked models is depicted in Table 2. We used both Framingham 10-year CVD risk versions, with and without lipids, as well as QRisk3 as baseline models to assess the performance of predicting

someone's 10-year risk of developing an atherosclerotic cardiovascular disease based on a holistic set of risk factors, with a focus on actionable risk factors and outcomes. The best performing model was XGB with an AUROC of 75.73%, only marginally higher than the Logistic Regression model with L1 regularization (75.44%) and substantially better than the Random Forest model (66.90%).

**Table 2. Performance of all tested classifiers including baseline models.**

| No. | Algorithm Name | AUROC and 95% confidence intervals |
|---|---|---|
| 1 | Extreme Gradient Boosting (XGB) | 0.7573 (0.755-0.7595) |
| 2 | Logistic Regression with L1 regularization | 0.7544 (0.755-0.7595) |
| 3 | QRisk3 | 0.725 (0.7226-0.7273) |
| 4 | Framingham Lipid & BMI | 0.680 (0.6775-0.6824) & 0.681 (0.6788-0.6837) |
| 5 | Random Forest | 0.6690 (0.6666-0.6715) |

Fig 1 shows the AUROCs of the best performing models XGB and from Logistic Regression with L1 regularization, which is the simplest model tested and amongst the top two best performing models. Logistic Regression comes with the advantages of being interpretable by providing reasoning for its classifications, and being a simple and robust method (35).

In order to better evaluate the clinical implications and significance of our results, we compared the results of our benchmarked models with our baseline models Framingham and QRisk3. Table 2 shows that both, our XGB and Logistic Regression classifiers achieved superior

353 performance compared to the baseline models. Apart from the Random Forest model, all tested
354 models had a higher AUROC than both baseline Framingham (68.0% and 68.1%) and QRisk3
355 (72.5%) models.
356 The difference in AUROC performance of the Framingham score in our experiments in Fig 1
357 and the one stated from Alaa et al. (28) in their study are explainable by the related work
358 approach using an older UK Biobank version with 40,000 fewer baseline patients and their last
359 available date of participant follow-up being February 17, 2016. Furthermore, our UK Biobank
360 version has biochemistry data which was released May 1, 2019 including cholesterol and
361 additional questionnaires data which the related approach did not have. Additionally, more
362 diagnosis data was made available over time. These dataset differences explain the difference
363 in AUROC.
364
365 **Fig 1. AUROC of Logistic Regression with L1 regularization and XGBoost**
366
367 Figs 2 and 3 show the AUROCs of all baseline models on imputed and unimputed data
368 respectively.
369
370 **Fig 2. AUROC curves of baseline models on imputed data**
371
372
373 **Fig 3. AUROC curves of baseline models on unimputed data**
374
375 Both Framingham versions perform nearly identically on imputed and unimputed data whereas
376 QRisk3 performs worse on unimputed data.
377 **Feature elimination vs. predictive performance**

Fig 4 shows how the performance of the best Logistic Regression model depends on the number of risk factors used. Stepwise discarding the risk factors leads to a relatively unchanged and stable model performance until around 170 iterations of feature elimination. This indicates that for predicting an individual's 10-year atherosclerotic CVD risk, many features provide only marginal value and a small subset of features provides substantial informative value. After around 170 iterations, there was a marked decline in model performance associated with further reductions in utilized features.

**Fig 4. Performance of best Logistic Regression model depending on number of features.** AUROC performance of best performing Logistic Regression model with L1 regularization (continuous blue line) compared to number of features utilized in each iterative feature elimination step (orange line), dotted blue horizontal line showing intersection of 25 features with iterative feature elimination step, allowing for extrapolation to model performance.

Table 3 shows in more detail the dependence of the model performance on the number of features. Utilizing only 25 (88%) out of the 203 total risk factors still leads to a reasonable AUROC performance, with a high reduction in utilized features. Compared to the model performance with an AUROC of 75.44% when using all 203 risk factors, the model still achieves 74.15% with the 25 most informative risk factors.

We also assessed the concrete performance for fewer features. To reach the same performance as QRisk3 of 72.5% AUROC, 16 features would be necessary. The two most informative features are age and biological sex. To reach a similar performance as Framingham (68.0%), two features would be necessary (68.98%). It is worth noting that both Framingham and QRisk3 were trained and tuned on other datasets and have different CVD definitions and objectives.

**Table 3. Performance of best Logistic Regression model depending on number of features.**

| Number of Features | AUROC |
|---|---|
| 203 | 75.44 |
| 40 | 75.01 |
| 25 | 74.15 |
| 20 | 73.32 |
| 17 | 72.76 |
| 10 | 70.88 |
| 2 | 68.98 |

**Generalizability results**

We assessed the generalizability of our models with the aforementioned approach of re-training the two previously best performing models only on a white cohort and testing them on a non-white cohort. Table 4 and Fig 5 show the results for Logistic Regression and XGB. The Logistic Regression model has an AUROC of 75.86% in the generalizability experiment, compared with an AUROC of 75.44% in the previous experiment. XGB has an AUROC of 76.26% in the generalizability experiment and 75.73% in the previous experiment. These results show marginal differences to the results of the previous experiments.

**Table 4. Model performance when trained on whites and tested on non-whites.**

| Model | AUROC on generalizability experiment | Previous AUROC results |
|---|---|---|
| Logistic Regression with L1 regularization | 75.86% | 75.44% |
| XGBoost | 76.26% | 75.73% |

**Fig 5. AUROC of Logistic Regression with L1 regularization and XGBoost when trained on whites and tested on non-whites.**

**Predictive ability of individual variables in UK Biobank.**

Table 5 shows the relative regression feature weights of the 25 most informative risk factors in descending order. A full list is provided in the supplementary materials (S5 Table). Based on our previous manual curation of risk factors and outcomes, we can see that the most informative risk factors are distributed across 5 categories (Table 6). The two most informative features were age and biological sex.

**Table 5. Relative regression feature weights of 25 most informative risk factors from best Logistic Regression model.**

| Feature number | Risk factor name | Relative informative value descending |
|---|---|---|

| | | |
|---|---|---|
| 1 | Age | 0.0938 |
| 2 | Biological sex | 0.0485 |
| 3 | Systolic blood pressure | 0.0284 |
| 4 | Social visits: About once a week | 0.0277 |
| 5 | Social visits: 2-4 times a week | 0.0273 |
| 6 | Walking pace: Brisk pace | 0.0268 |
| 7 | Total cholesterol HDL ratio | 0.0267 |
| 8 | Total cholesterol | 0.0239 |
| 9 | LDL cholesterol | 0.0235 |
| 10 | Familial CVD | 0.0218 |
| 11 | Social visits: About once a month | 0.0203 |
| 12 | Sleep problems: Not at all | 0.0188 |
| 13 | Alcohol with meals: Yes | 0.0184 |
| 14 | Smoking | 0.0184 |
| 15 | Social visits: Almost daily | 0.0178 |
| 16 | No. of cigarettes daily | 0.0163 |
| 17 | Hypertension | 0.0160 |
| 18 | Walking pace: Steady average | 0.0154 |

| | | |
|---|---|---|
| | pace | |
| 19 | Waist circumference | 0.0150 |
| 20 | Alcohol with meals: It varies | 0.0141 |
| 21 | Social visits: Once every few months | 0.0139 |
| 22 | Overall health rating: Excellent | 0.0134 |
| 23 | Other Heart Arrhythmias | 0.0129 |
| 24 | Overall health rating: Poor | 0.0123 |
| 25 | Sleep problems: Several days | 0.0122 |

**Table 6. Categorization of the 25 most informative risk factors into categories from the best Logistic Regression model.**

| **Category** | **Risk Factors** |
|---|---|
| Demographics | Age, Biological sex |
| Biomarkers | Waist circumference, systolic blood pressure, total cholesterol, LDL cholesterol, total cholesterol HDL ratio |
| Comorbidities | Hypertension, sleep problems: not at all, sleep problems: several days, other heart arrhythmias |
| Family History | Familial CVD |

| | |
|---|---|
| Lifestyle Factors | Social visits: about once/week, social visits: 2-4 times/week, social visits: about once/month, social visits: almost daily, social visits: once every few months, smoking, no. of cigarettes daily, alcohol with meals: yes, alcohol with meals: it varies, walking pace: steady average pace, walking pace: Brisk pace, overall health rating: excellent, overall health rating: poor |

# Discussion

Using data gathered from the large longitudinal cohort UK Biobank study, we developed a pipeline to benchmark several classification models for predicting a subject's 10-year absolute risk of developing an atherosclerotic CVD. We used an extensive set of physician curated risk factors and outcomes methodology, employing a holistic view of the subject's current health status rooted in a precision medicine approach. The models were trained and evaluated using data from 464,547 UK Biobank participants, spanning 203 CVD risk factors for each subject. Using a simple Logistic Regression model with a holistic set of risk factors significantly improved the accuracy of atherosclerotic CVD risk prediction compared to currently available, widely used and recommended models such as Framingham and QRisk3. Both of these existing models rely on a limited set of risk factors and outcomes and do not focus on modifiable lifestyle factors. Further, our best performing Logistic Regression model utilizes new CVD risk predictors showing high predictive power, which are social visits, walking pace and overall health rating. The frequency of social visits could be indicative of someone's current mental health status, which has been shown to be a relevant CVD risk factor (55,56). These and other non-laboratory risk factors could be collected by means of a questionnaire or passively deduced using data analytics from data sources such as GPS, calendar and sensors from smartphones,

452 smartwatches and fitness trackers.

453 Additionally, our best performing models, XGBoost and Logistic Regression, showed marginal
454 differences when trained and tested on particular sub-populations, which is indicative of good
455 generalizability to other ethnicities.

456 As there was little performance difference between the best performing models, we primarily
457 discuss the simplest model, Logistic Regression with L1 regularization. This model has the
458 inherent benefit of offering reasoning for its predictions, through analyzing the learned
459 coefficients for every risk factor and having feature selection performed by the L1 regularization.
460 With L1 regularization, less important risk factors' coefficients are minimised and also set to
461 zero, which then leads to entire removal of these features from the model, and fewer risk factors
462 needed for an accurate prediction.

463

464 Using iterative feature elimination, we identified a subset of the 25 most relevant risk factors
465 providing a similar performance compared to using all 203 risk factors. With the 25 most
466 relevant risk factors belonging to five different categories, suggests that different biological
467 layers contribute to the risk of atherosclerotic CVD. This result indicates that it is insufficient to
468 assess only one biological layer for accurate risk prediction, confirming the findings of other
469 studies for identifying novel biomarkers and pathways in complex diseases (57). This result
470 supports our initial model development approach: to use a holistic model for an individual's
471 health. Our approach was rooted in precision medicine and takes into account multiple
472 biological layers by using multi-omics as well as clinical and lifestyle data with the aim to capture
473 all potential interactions or correlations detected between molecules in different biological layers
474 (22). Multi-omics data generated for the same set of samples can provide useful insights into
475 the interaction of biological information at multiple layers and thus can help in understanding the
476 mechanisms underlying the complex biological condition of interest.

477

In our model, the lifestyle category contributed the most risk factors, suggesting that it is essential to include someone's daily lifestyle data and not just periodic snapshots of clinical data into an individual's risk assessment for a complex disease like CVD. The causal relationships between the risk factors considered in our model and atherosclerotic CVDs have been demonstrated by other studies (11,19,21,25). Innovative approaches are needed in order to tackle the increasing prevalence and mortality of CVD-related diseases (2), and the associated healthcare systems' financial burdens. This is especially required in low and middle income countries where CVD prevalence has also been increasing and is expected to increase as a consequence of an aging and growing population (2).

There is potential for novel disruptive approaches to affordably improve CVD outcomes. Areas where this may have an impact is in novel approaches to screening, lifestyle coaching and prevention (2). Screening will become more accessible and widespread by more (near-)medical-grade sensors being integrated into smartphones and smartwatches, enabling continuous monitoring of relevant behavioral CVD risk factors, as well as biomarkers such as heart rate, blood pressure and blood glucose. By gathering a wider spectrum of relevant risk factors for cardiovascular disease automatically and continuously, an ongoing and personalized cardiovascular disease risk prediction could be enabled. Through linking personalised information on an individual's CVD risk with app-based programmes for sustained behavioural modification, it may be possible to lower the incidence and mortality of CVDs (58). Combined with a companion smartphone-based app, an AI or healthcare provider-generated personalised intervention program could be provided, and targeted at those people who need it the most. Many studies have shown that digital health interventions are cost effective for managing CVD (for a review see (59)). One report found that a community-based prevention program could have a mean return on investment (ROI) on medical cost savings of $5.60 for every $1 spent within a 5 year timeframe by improving physical activity and nutrition and reducing tobacco

usage (60). A review of 11 in-home cardiac rehabilitation programs for the secondary prevention of CVD found that social support, goal setting, monitoring, credible instructions and literature resources are all effective behavior change techniques to reduce behavioral risk factors for CVD (61).

The improvement achieved by our models might be partially attributed to being trained and assessed on the UK Biobank dataset, whereas the baseline Framingham model was derived from a different population. The population and many of the data sources used in the QRisk3 model are similar, being the general UK population and using their GP, hospital and mortality records. However, our risk model generation approach and QRisk3's approach were designed with different aims and objectives and the modelling strategy was different. For these reasons, direct comparison between the models is limited. Notable differences between the approaches include a more limited set of risk factors included in Framingham and QRisk3's and a focused and wider range of atherosclerotic CVDs included in our approach.

The results from our generalizability subanalysis indicate that our XGB and Logistic Regression models might generalize well to other ethnicities and do not overfit to our cohort, however, this needs to be further evaluated with more data from diverse ethnicities.

Our results show that our models have improved performance over the baseline models Framingham and QRisk3 (Table 2). This is because the selection of the appropriate disease modelling approach, classifiers and careful tuning of the model's hyperparameters are crucial steps for realizing the potential benefits of ML. Our pipeline automates some of these steps which makes the tuning and discovery of new disease risk models easily accessible for clinical research. Our prospective cohort modelling approach, which is rooted in precision medicine, is

529 the first to generate an atherosclerotic CVD absolute risk prediction tool based upon a complete
530 definition of atherosclerotic CVD outcomes and a holistic set of risk factors.
531
532 **Limitations**
533 The UK Biobank only admitted participants for their initial signup from the ages 40 and up. This
534 might limit the applicability of the risk score for younger populations and further tests with data
535 from younger populations need to be conducted.
536
537 There are many missing data values related to the potential risk factors for many participants.
538 Having more unimputed data of relevant CVD risk factors could improve the predictive
539 performance of all our benchmarked classifiers and could also lead to changes in the classifier
540 ranking from Table 2 and relative risk factor importances in Table 5. However, the use of
541 imputed data is highly unlikely to have an impact on our conclusion that a holistic set of risk
542 factors and an exhaustive atherosclerotic CVD outcome definition could improve atherosclerotic
543 and actionable CVD risk prediction.
544
545 An additional limitation of our study is that the UK Biobank dataset consists of participants of
546 predominantly (88%) British ethnicity, with an even larger portion having a white background
547 (91%). Therefore, further assessments of the influence of the ethnicity predictor need to be
548 carried out to enable a generalizable tool. Previous work in this area indicates that the plaque
549 growth process seems to be independent of ethnicity (21).
550 A further limitation of this UK focused dataset is that socio-economic and other environmental
551 factors differ between countries. This is another potential bias that needs to be further evaluated
552 with datasets from other countries with different socio-economic characteristics.
553

Disease risk prediction models which include subjective non-laboratory risk factors, such as the self-reported health rating and usual walking pace, should be cautiously evaluated to minimize self-reported bias. These risk factors have been found to be good predictors of someone's overall CVD risk in another study using UK Biobank data (28).

# Conclusions

We benchmarked multiple classifiers to predict an individual's 10-year risk of developing an atherosclerotic CVD, using a holistic set of risk factors and a specific definition of atherosclerotic CVDs. Our reduced Logistic Regression with L1 regularization classifier, a simple and interpretable model, is amongst our best prediction models, includes actionable lifestyle factors, has great predictive power and requires 13 unique features. Our experiments showed that a two feature-questionnaire is as accurate as the Framingham models and a 16 feature-questionnaire is as accurate as QRisk3 for 10-year atherosclerotic CVD risk prediction. Both prediction models, XGBoost and Logistic Regression, generalize well to non-white people, which might indicate that our models generalize well to other (western) countries. Framingham and QRisk3, which are well established and validated absolute risk prediction models, do not perform as well on predicting individuals' 10-year risk of developing an atherosclerotic CVD. With our Logistic Regression model, we created a promising new interpretable, actionable and accurate risk prediction tool that could assist individuals and public health in CVD risk reduction.

# Acknowledgments

# Author Contributions

**Conceptualization.** Ajay Kesar, Stephen Gilbert, Paul Wicks, Bernard Leon Stopak

**Data Curation.** Ajay Kesar, Adel Baluch, Omer Barber, Milan Jovanovic

**Formal Analysis.** Ajay Kesar, Daniel Renz

**Funding Acquisition.** Stephen Gilbert, Bernard Leon Stopak, Henry Hoffmann

**Investigation.** Ajay Kesar

**Methodology.** Ajay Kesar, Daniel Renz

**Project Administration.** Ajay Kesar

**Resources.** Ajay Kesar, Stephen Gilbert, Bernard Leon Stopak, Henry Hoffmann

**Software.** Ajay Kesar, Daniel Renz

**Supervision.** Stephen Gilbert, Henry Hoffmann

**Validation.** Ajay Kesar, Daniel Renz

**Visualization.** Ajay Kesar

**Writing – Original Draft Preparation.** Ajay Kesar, Daniel Renz, Paul Wicks, Stephen Gilbert

**Writing – Review & Editing.** Ajay Kesar, Henry Hoffmann, Daniel Renz, Bernard Leon Stopak, Paul Wicks, Stephen Gilbert

# References

1. Cardiovascular diseases (CVDs) [Internet]. [cited 2021 Sep 28]. Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

2. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019. J Am Coll Cardiol. 2020 Dec 22;76(25):2982–3021.

3. Heidenreich PA, Trogdon JG, Khavjou OA, Butler J, Dracup K, Ezekowitz MD, et al. Forecasting the Future of Cardiovascular Disease in the United States. Circulation. 2011 Mar 1;123(8):933–44.

4. Weintraub WS, Daniels SR, Burke LE, Franklin BA, Goff DC, Hayman LL, et al. Value of Primordial and Primary Prevention for Cardiovascular Disease. Circulation. 2011 Aug 23;124(8):967–90.

5. Evsikova C, Raplee I, Lockhart J, Jaimes G, Evsikov A. The Transcriptomic Toolbox: Resources for Interpreting Large Gene Expression Data within a Precision Medicine Context for Metabolic Disease Atherosclerosis. J Pers Med. 2019 Apr 29;9:21.

6. Nichols GA, Bell TJ, Pedula KL, O'Keeffe-Rosetti M. Medical care costs among patients with established cardiovascular disease. Am J Manag Care. 2010 Mar 1;16(3):e86–93.

7. Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, 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 Aug 1;37(29):2315–81.

8. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. J Am Coll Cardiol. 2014 Jul 1;63(25 0 0):2935–59.

9. Sedgwick JEC. Absolute, attributable, and relative risk in the management of coronary heart disease. Heart. 2001 May 1;85(5):491–2.

10. Jackson R. Guidelines on preventing cardiovascular disease in clinical practice: Absolute risk rules—but raises the question of population screening. BMJ. 2000 Mar 11;320(7236):659–61.

11. Libby P, Bonow RO, Mann DL, Tomaselli GF, Zipes DP. Braunwald's Heart Disease E-Book: A Textbook of Cardiovascular Medicine. Elsevier Health Sciences; 2018. 2527 p.

12. Eriksen CU, Rotar O, Toft U, Jørgensen T. What is the effectiveness of systematic population-level screening programmes for reducing the burden of cardiovascular diseases? [Internet]. Copenhagen: WHO Regional Office for Europe; 2021 [cited 2021 Oct 12]. (WHO Health Evidence Network Synthesis Reports). Available from: http://www.ncbi.nlm.nih.gov/books/NBK567843/

13. Lim LS, Haq N, Mahmood S, Hoeksema L. Atherosclerotic Cardiovascular Disease Screening in Adults: American College of Preventive Medicine Position Statement on Preventive Practice. Am J Prev Med. 2011 Mar 1;40(3):381.e1-381.e10.

14. Espinoza J, Crown K, Kulkarni O. A Guide to Chatbots for COVID-19 Screening at Pediatric Health Care Facilities. JMIR Public Health Surveill. 2020 Apr 30;6(2):e18808.

15. Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, et al. Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation. N Engl J Med. 2019 Nov 14;381(20):1909–17.

16. Lemmen C, Simic D, Stock S. A Vision of Future Healthcare: Potential Opportunities and Risks of Systems Medicine from a Citizen and Patient Perspective—Results of a Qualitative Study. Int J Environ Res Public Health. 2021 Sep 19;18(18):9879.

17. Peeters JM, Krijgsman JW, Brabers AE, Jong JDD, Friele RD. Use and Uptake of eHealth in General Practice: A Cross-Sectional Survey and Focus Group Study Among Health Care Users and General Practitioners. JMIR Med Inform. 2016 Apr 6;4(2):e4515.

18. Bui QT, Prempeh M, Wilensky RL. Atherosclerotic plaque development. Int J Biochem Cell Biol. 2009 Nov 1;41(11):2109–13.

19. Herrington W, Lacey B, Sherliker P, Armitage J, Lewington S. Epidemiology of Atherosclerosis and the Potential to Reduce the Global Burden of Atherothrombotic Disease. Circ Res. 2016 Feb 19;118(4):535–46.

20. Bentzon JF, Otsuka F, Virmani R, Falk E. Mechanisms of Plaque Formation and Rupture. Circ Res. 2014 Jun 6;114(12):1852–66.

21. Insull W. The Pathology of Atherosclerosis: Plaque Development and Plaque Responses to Medical Treatment. Am J Med. 2009 Jan 1;122(1, Supplement):S3–14.

22. Picard M, Scott-Boyer M-P, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021 Jan 1;19:3735–46.

23. Collins FS, Varmus H. A New Initiative on Precision Medicine [Internet]. https://doi.org/10.1056/NEJMp1500523. Massachusetts Medical Society; 2015 [cited 2021 Sep 29]. Available from: https://www.nejm.org/doi/10.1056/NEJMp1500523

24. Leon-Mimila P, Wang J, Huertas-Vazquez A. Relevance of Multi-Omics Studies in Cardiovascular Diseases. Front Cardiovasc Med. 2019;6:91.

25. Fruchart J-C, Nierman MC, Stroes ESG, Kastelein JJP, Duriez P. New Risk Factors for Atherosclerosis and Patient Risk Assessment. Circulation. 2004 Jun 15;109(23_suppl_1):III–15.

26. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008 Feb 12;117(6):743–53.

27. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017 May 23;357:j2099.

28. Alaa AM, Bolton T, Angelantonio ED, Rudd JHF, Schaar M van der. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLOS ONE. 2019 May 15;14(5):e0213653.

29. Conroy RM, Pyörälä K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003 Jun 1;24(11):987–1003.

30. SCORE2 working group and ESC Cardiovascular risk collaboration. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. Eur Heart J. 2021 Jul 1;42(25):2439–54.

31. Dolezalova N, Reed AB, Despotovic A, Obika BD, Morelli D, Aral M, et al. Development of an accessible 10-year Digital CArdioVAscular (DiCAVA) risk assessment: a UK Biobank study. Eur Heart J - Digit Health. 2021 Sep 1;2(3):528–38.

32. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep. 2020 Jul 20;10(1):11981.

33. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol. 2019 May 1;20(5):e262–73.

34. Doupe P, Faghmous J, Basu S. Machine Learning for Health Services Researchers. Value Health. 2019 Jul 1;22(7):808–15.

35. Adadi A, Berrada M. Explainable AI for Healthcare: From Black Box to Interpretable Models. In: Bhateja V, Satapathy SC, Satori H, editors. Embedded Systems and Artificial Intelligence. Singapore: Springer Singapore; 2020. p. 327–37.

36. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med. 2019 Jan;25(1):30–6.

37. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12(85):2825–30.

38. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016 Aug 13;785–94.

39. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, 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 Mar 31;12(3):e1001779.

40. About us [Internet]. [cited 2021 Nov 9]. Available from: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us

41. Collins R. UK Biobank Protocol. :112.

42. Ethics [Internet]. [cited 2021 Nov 9]. Available from: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics

43. Cardiovascular Disease (10-year risk) | Framingham Heart Study [Internet]. [cited 2021 Nov 10]. Available from: https://framinghamheartstudy.org/fhs-risk-functions/cardiovascular-disease-10-year-risk/

44. QRISK3 [Internet]. [cited 2021 Nov 10]. Available from: https://qrisk.org/three/index.php

45. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. Ann Stat. 2001;29(5):1189–232.

46. XGBoost Documentation — xgboost 1.6.0-dev documentation [Internet]. [cited 2021 Nov 8]. Available from: https://xgboost.readthedocs.io/en/latest/

47. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Their Appl. 1998 Jul;13(4):18–28.

48. Zhang T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: Proceedings of the twenty-first international conference on Machine learning [Internet]. New York, NY, USA: Association for Computing Machinery; 2004 [cited 2021 Nov 12]. p. 116. (ICML '04). Available from: https://doi.org/10.1145/1015330.1015332

49. Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J Comput Syst Sci. 1997 Aug 1;55(1):119–39.

50. Hastie T, Rosset S, Zhu J, Zou H. Multi-class AdaBoost. Stat Interface. 2009;2(3):349–60.

51. Omohundro SM. Five balltree construction algorithms. International Computer Science Institute Berkeley; 1989.

52. Srivastava S, Gupta MR, Frigyik BA. Bayesian quadratic discriminant analysis. J Mach Learn Res. 2007;8(6).

53. Zhang H. The optimality of naive Bayes. AA. 2004;1(2):3.

54. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B Methodol. 1996;58(1):267–88.

55. Correll CU, Solmi M, Veronese N, Bortolato B, Rosson S, Santonastaso P, et al. Prevalence, incidence and mortality from cardiovascular disease in patients with pooled and specific severe mental illness: a large-scale meta-analysis of 3,211,768 patients and 113,383,368 controls. World Psychiatry. 2017;16(2):163–80.

56. Cunningham R, Poppe K, Peterson D, Every-Palmer S, Soosay I, Jackson R. Prediction of cardiovascular disease risk among people with severe mental illness: A cohort study. PLOS ONE. 2019 Sep 18;14(9):e0221521.

57. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017 May 5;18(1):83.

58. Gao W, Yu C. Wearable and Implantable Devices for Healthcare. Adv Healthc Mater. 2021 Sep 1;10(17):2101548.

59. Jiang X, Ming W-K, You JH. The Cost-Effectiveness of Digital Health Interventions on the Management of Cardiovascular Diseases: Systematic Review. J Med Internet Res. 2019 Jun 17;21(6):e13166.

60. Trust for America's Health. Prevention for a healthier America: Investments in disease prevention yield significant savings, stronger communities. 2008;

747 61. Heron N, Kee F, Donnelly M, Cardwell C, Tully MA, Cupples ME. Behaviour change
748 techniques in home-based cardiac rehabilitation: a systematic review. Br J Gen Pract.
749 2016 Oct;66(651):e747–57.

750

# 751 Supporting Information

752 **S1 Table. List of all risk factors used in our analysis.** (XLSX)

753 The listed risk factors were summarized into 203 risk factors for the respective UK Biobank

754 participant.

755 **S2 Table. List of all outcomes used in our analysis.** (XLSX)

756 The following outcomes were all consolidated into one final binary outcome column indicating if

757 the respective UK Biobank participant did or did not develop one the relevant atherosclerotic

758 CVDs during their individual 10-year follow-up period starting from their individual initial

759 assessment attendance date.

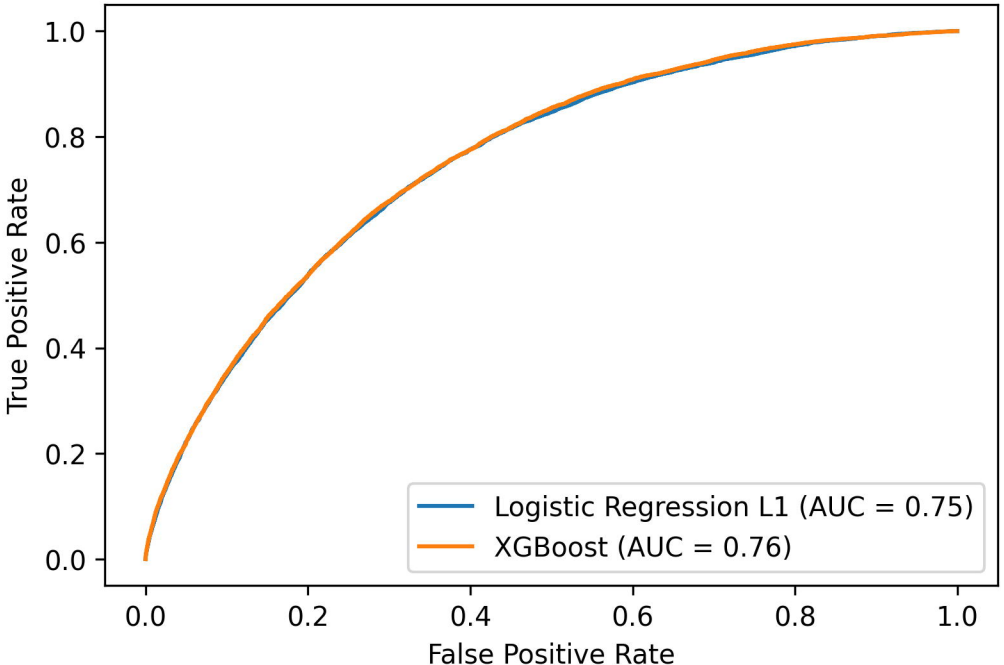
760 **S3 Table. Specifications of the python (v3.9.6) libraries and their versions used in this**

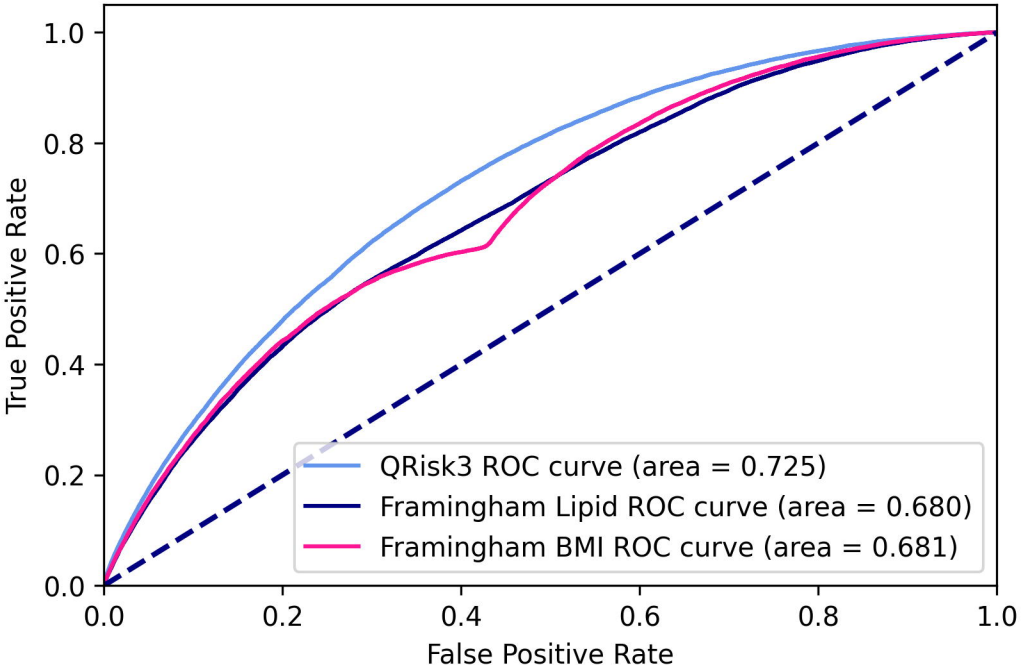
761 **study.** (PDF)

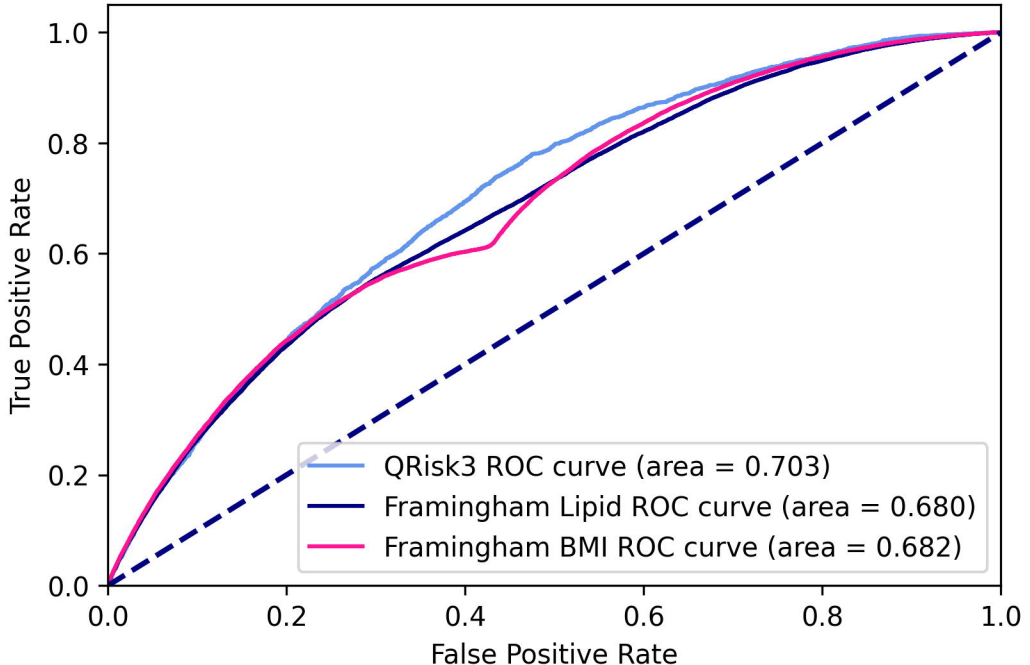
762 **S4 Table. List of utilized open-source methods, best parameters and references.** (PDF)

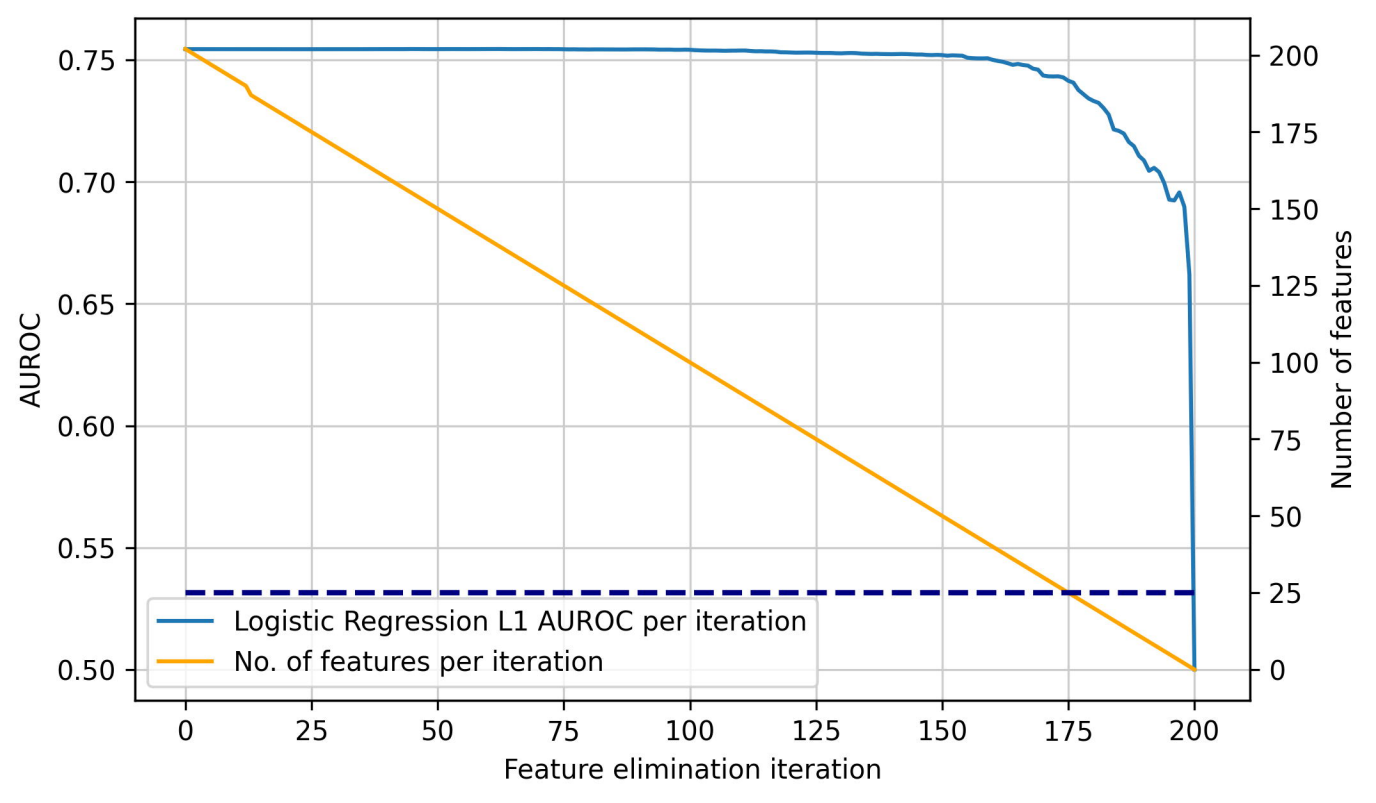
763 **S5 Table. Full list of relative informative values for each risk factor for best performing**

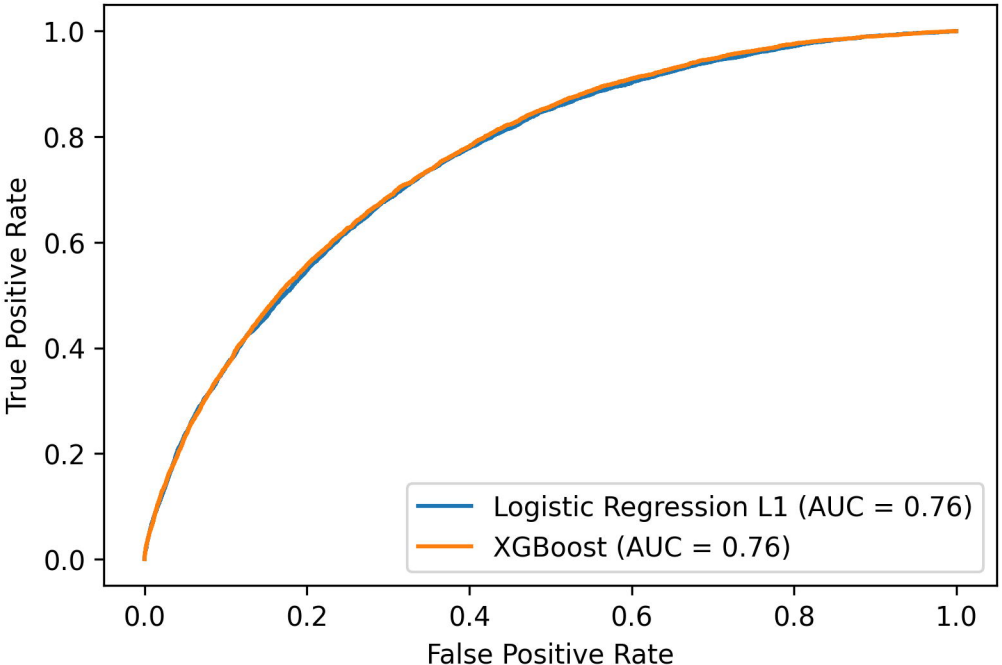
764 **Logistic Regression model.** (XLSX)


1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Logistic Regression L1 (AUC = 0.75)
XGBoost (AUC = 0.76)


1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
QRisk3 ROC curve (area = 0.725)
Framingham Lipid ROC curve (area = 0.680)
Framingham BMI ROC curve (area = 0.681)


1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
QRisk3 ROC curve (area = 0.703)
Framingham Lipid ROC curve (area = 0.680)
Framingham BMI ROC curve (area = 0.682)


AUROC
Number of features
Feature elimination iteration
Logistic Regression L1 AUROC per iteration
No. of features per iteration
0.75
0.70
0.65
0.60
0.55
0.50
200
175
150
125
100
75
50
25
0
0
25
50
75
100
125
150
175
200


1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Logistic Regression L1 (AUC = 0.76)
XGBoost (AUC = 0.76)