Medical
Research
Archives

European
Society of
Medicine

RESEARCH ARTICLE

Identifying Biomarkers of Cardiovascular Diseases with Machine Learning: Evidence from The UK Household Longitudinal Study

Vasilis Nikolaou[1,*], Sebastiano Massaro[1,2], Masoud Fakhimi[1], Wolfgang Garn[2]

[1]Surrey Business School, University of Surrey, Guildford, UK
[2]The Organizational Neuroscience Laboratory, London, UK

***Correspondent author:** v.nikolaou@surrey.ac.uk

ABSTRACT

Cardiovascular diseases are a significant global health concern, responsible for one-third of deaths worldwide and posing a substantial burden on society and national healthcare systems. To effectively address this challenge and develop targeted intervention strategies, the ability to predict cardiovascular diseases from standardized assessments, such as occupational health encounters or national surveys, is critical. This study aims to assist these efforts by identifying a set of biomarkers, which together with known risk factors, can predict cardiovascular diseases on the onset. We used a sample of 7,767 individuals from the UK household longitudinal study 'Understanding Society' to train several machine learning models able to pinpoint biomarkers and risk factors at baseline that predict cardiovascular diseases at a ten-year follow-up. A logistic regression model was trained for comparison. A gaussian naïve bayes classifier returned 82% recall in contrast to 48% of the logistic regression, allowing us to identify the most prominent biomarkers predicting cardiovascular diseases. These findings show the opportunity to use machine learning to identify a wide range of previously overlooked biomarkers associated with cardiovascular diseases onset and thus encourage the implementation of such a model in the early diagnosis and prevention of cardiovascular diseases in future research and practice.

Keywords: Machine Learning; Naïve Bayes; Logistic Regression; Biomarkers; Cardiovascular diseases

OPEN ACCESS

Published: January 31, 2024

Citation: Nikolaou V, Massaro S, et al., 2024. Identifying Biomarkers of Cardiovascular Diseases with Machine Learning: Evidence from The UK Household Longitudinal Study, Medical Research Archives, [online] 12(1).

https://doi.org/10.18103/mra.v12i1.4966

DOI
https://doi.org/10.18103/mra.v12i1.4966

ISSN: 2375-1924

1. Introduction

Cardiovascular diseases (CVDs) refer to a group of conditions that affect the heart and/or blood vessels, occurring when fatty deposits accumulate in the arteries, leading to blood clots that can damage the heart and other organs.[1] According to the World Health Organization (WHO), CVDs are the main cause of death globally. In 2019, approximately 18 million deaths were attributed to CVDs, accounting for 32% of global deaths, with heart attack and stroke accounting for 85% of those deaths.[2] In addition to the high mortality rate, CVDs substantially impact several aspects of people's life and society as a whole, spanning from sustained healthcare costs to workplace well-being, to name a few.[3] In recent years, several approaches have been taken to mitigate the burden of CVDs to reduce healthcare costs and mortality rates (e.g., providing aspirin to high-risk individuals, controlling diabetes, weight reduction in obese individuals).[4] However, despite routinely capturing acknowledged risk factors for CVDs, including smoking, high blood pressure, high cholesterol, diabetes, high blood glucose, increased blood lipids, and obesity, standardized assessments like clinical tests, occupational assessments, national surveys, and GP encounters have not been systematically used to address CVDs.[1,2] This is surprising since these assessments can provide valuable data to mitigate the burden of CVDs. For instance, in a 10-year longitudinal study of 24,558 healthy US women, Ridker et al.[5] found that age, adult haemoglobin (HbA), systolic blood pressure, current smoking, C-reactive protein (CRP), and total cholesterol were important risk factors associated with CVD events (e.g., myocardial infarction, ischemic stroke, coronary revascularization, and cardiovascular death).

Building on this knowledge, increasing attention has thus been placed on advancing research by using available biomarkers associated with CVDs.[6-9] In social sciences and medicine, attention has primarily been directed towards linking just a few selected biomarkers acting as antecedents of CVDs to issues such as stress levels and socioeconomic differences, overlooking the broader picture of how a set of biomarkers could combine with existing risk factors to predict the onset of CVDs. We argue that by leveraging available secondary data collected via routine assessments, this knowledge represents a valuable asset for health systems to envision more targeted interventions aimed at early diagnosis and prevention of CVDs, thereby reducing the economic and societal burden they impose. In other words, by identifying and analyzing multiple biomarkers together with known risk factors, we can develop a more comprehensive understanding of CVDs, leading to more effective interventions. This approach holds promise in both reducing healthcare costs and improving patient outcomes, highlighting the importance of utilizing routine assessments and secondary data to further our knowledge of CVDs.

In the remainder of this paper, we address this gap as follows. First, we review the existing literature that has associated biomarkers with CVDs and highlight that traditional statistical methodology has been predominantly employed, limiting the detection of a comprehensive set of biomarkers. Next, we present our methodology that utilizes several machine learning computational models on a large dataset from the 'UK Understanding Society longitudinal survey' to identify a set of biomarkers and individual factors predictive of CVDs at the onset. Our findings put forward a Gaussian Naïve Bayes (GNB) classifier able to deliver the highest recall as opposed to the regression model allowing us to identify the most prominent biomarkers predicting CVDs. Finally, we discuss the implications for social science and medical research and practice.

2. Biomarkers and cardiovascular diseases: Existing Knowledge and Current Gaps

In recent years, there has been a growing research interest in identifying individual physiological and lifestyle factors that can be used to predict CVDs (see Table 1 in Appendix).

Melander et al.,[10] investigated the usefulness of biomarkers in predicting cardiovascular risk along with conventional risk factors such as smoking, diabetes, hypertension, or hyperlipidemia. The authors studied a cohort of 5,067 middle-aged participants from Malmö (Sweden) without cardiovascular disease who underwent a baseline assessment (between 1991 and 1994) that included a range of biomarkers, including CRP, cystatin C, lipoprotein-associated phospholipase 2, midregional proadrenomedullin (MR-proADM), midregional proatrial natriuretic peptide, and N-terminal pro-B-type natriuretic peptide (N-BNP). The participants were followed up until 2016 for the first occurrence of cardiovascular events (i.e., myocardial infarction, stroke, coronary death). As a result, several biomarkers were assessed, and those retained for predicting cardiovascular events were CRP and N-BNP.

Shlipak et al.,[11] focused on six biomarkers (N-terminal prohormone brain natriuretic peptide (Nt-proBNP), cystatin C, albuminuria, CRP, interleukin-6, and fibrinogen) to predict cardiovascular events (stroke, myocardial infarction, and coronary heart disease death) among 979 patients with pre-existing coronary artery disease. Three of those biomarkers (Nt-proBNP, albuminuria, and CRP) reflecting hemodynamic stress, kidney damage, and inflammation, respectively, were found to be significantly associated with an increased risk of cardiovascular events.

Using data from the Uppsala Longitudinal Study of Adult Men (ULSAM), a community-based cohort of elderly men, Zethelius et al.,[12] investigated whether a combination of biomarkers reflecting myocardial cell damage, left ventricular dysfunction, renal failure, and inflammation (troponin I, N-terminalpro–brain natriuretic peptide, cystatin C, and C-reactive protein, respectively) could improve risk-stratification of a person beyond the assessment of established risk factors for cardiovascular disease, such as, age, systolic blood pressure, use or non-use of antihypertensive treatment, total cholesterol, high-density lipoprotein cholesterol, use or non-use of lipid-lowering treatment, presence or absence of diabetes, smoking status, and body-mass index. Using Cox proportional-hazards models – which assumes a constant hazard over time when modelling the association between biomarkers and the time to the first occurrence of a CVD event - they found that the four additional biomarkers improved the model's predictability for increased risk of death from CVDs.

Folsom et al.,[13] used data from the prospective Atherosclerosis Risk in Communities (ARIC) Study to analyse the association of 19 novel risk markers with incident CHD in 15,792 adults followed up from 1987-1989. The study found that in addition to the traditional risk factors (age, race, sex, total and high-density lipoprotein cholesterol levels, systolic blood pressure, antihypertensive medication use, smoking status, and diabetes), C-reactive protein level was significantly associated with increased risk of CHD (Hazard Ratio [HR] = 1.17, 95% CI:1.05 - 1.30; P = 0.005).

Wang et al., [14] focused on ten biomarkers (C-reactive protein, B-type natriuretic peptide, N-terminal pro–atrial natriuretic peptide, aldosterone, renin, fibrinogen, d-dimer, plasminogen-activator inhibitor type 1, and homocysteine; and the urinary albumin-to-creatinine ratio) in 3,209 participants of the Framingham Heart Study to evaluate the risk of CVD events. After a 7-year follow-up, the biomarkers which held the highest predictive power for CVDs death were B-type natriuretic peptide level, C-reactive protein level, the urinary albumin-to-creatinine ratio, homocysteine level, and renin level. Additionally, the biomarkers that were the strongest predictors of major cardiovascular events were B-type natriuretic peptide level and the urinary albumin-to-creatinine ratio.

Finally, in a multicenter clinical trial, Blankenberg et al.,[15] evaluated nine inflammatory biomarkers, microalbuminuria, and N-terminal pro-brain natriuretic peptide (Nt-proBNP) in 3,199 participants in the Heart Outcomes Prevention Evaluation (HOPE) Study. They aimed to improve cardiovascular (myocardial infarction, stroke, or cardiovascular death) risk prediction beyond that obtained from traditional risk factors in a secondary-prevention population. Nt-proBNP (HR = 1.72 per increment SD, 95% CI 1.39 to 2.12; P<0.0001), soluble intercellular adhesion molecule-1 (HR = 1.46, 95% CI 1.19 to 1.80; P=0.0003), microalbuminuria (HR = 1.55, 95% CI 1.22 to 1.98; P=0.0004), soluble interleukin-1 receptor antagonist (HR 1.30, 95% CI 1.05 to 1.61; P=0.02), and fibrinogen (HR = 1.31, 95% CI 1.05 to 1.62; P=0.02) remained significantly related to the primary outcome.

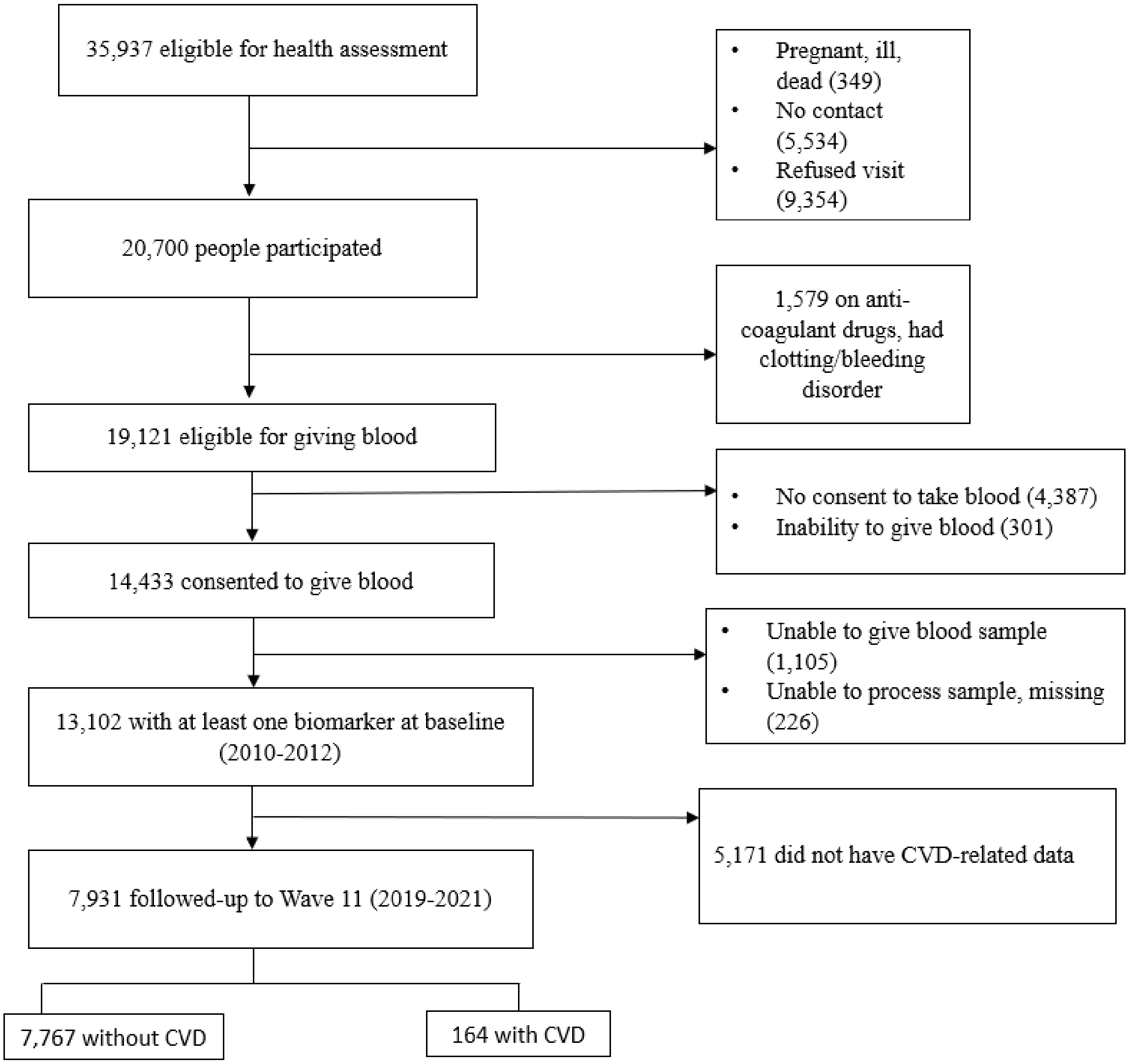
Notwithstanding the valuable insights offered in the studies above, a shared characteristic is the use of Cox-proportional hazard models – that assume constant hazard over time, which may not always be a realistic assumption as the risk of CVD often changes over time – to assess the association of potential biomarkers and occurrence of CVD events after adjusting for known risk factors.

Surprisingly, however, there is little evidence of research going beyond traditional statistical methods and using machine learning tools. Here, we argue that Machine Learning (ML) can offer a more efficient means to identify a wide range of biomarkers, along with already known risk factors, able accurately predict the occurrence of CVD events. Several ML models have been developed in recent years in cognate domains and were successful to predict acute myocardial ischemia;[16] improve CVDs risk prediction by using an automatic algorithm tool that selects and tunes ensembled of ML models on data from the Biobank database;[17] and predict cardiovascular comorbidities in patients with COPD.[18] Thus, we are confident that using ML to identify key biomarkers, whose association with CVD may have not yet been fully explored, may not only represent a promising way to validate findings obtained with traditional statistical methods but also aid healthcare systems to leverage large-scale data available from routine assessments toward the implementation of effective interventions against the impact of CVDs.

3. Methods

This study utilized data from the UK Household Longitudinal survey "Understanding Society", [19] which began its first wave of data collection from January to March 2011 and has since gathered a wealth of information on household and individual characteristics. During the second wave (January 2010 to March 2012), all adults aged 16 and above underwent a nurse health assessment interview to collect data on a range of physical measures and biomarkers. The baseline period for this study is defined as the time between 2010 and 2012, during which none of the participants reported any CVDs. The sample included 7,931 participants with available CVD data, who were followed up to the eleventh wave (January 2019 to March 2021), during which 164 individuals (2.1%) reported experiencing at least one of the following CVD events: congestive heart failure, coronary heart disease, angina, heart attack or myocardial infarction, and/or stroke (Figure 1).

Figure 1. Study flow chart


35,937 eligible for health assessment
• Pregnant, ill, dead (349)
• No contact (5,534)
• Refused visit (9,354)
20,700 people participated
1,579 on anti-coagulant drugs, had clotting/bleeding disorder
19,121 eligible for giving blood
• No consent to take blood (4,387)
• Inability to give blood (301)
14,433 consented to give blood
• Unable to give blood sample (1,105)
• Unable to process sample, missing (226)
13,102 with at least one biomarker at baseline (2010-2012)
5,171 did not have CVD-related data
7,931 followed-up to Wave 11 (2019-2021)
7,767 without CVD
164 with CVD

Baseline demographic characteristics included age, sex, height, weight, body mass index (BMI), waist circumference, smoking and waist-to-height ratio (WHtR). The latter has been found to be a more significant risk factor than waist circumference alone or BMI for cardiometabolic risk.[20] It has also been reported as an accurate anthropometric index to identify individuals with cardiovascular risk factors, both children and adults[21] and to be used as a cardiovascular stratification factor among obese youths.[22] Baseline clinical characteristics included both target biomarkers (i.e., total cholesterol and high-density lipoprotein (HDL) cholesterol, triglycerides, glycated haemoglobin, high sensitivity C-reactive protein, clauss fibrinogen, haemoglobin, ferritin, albumin, alkaline phosphatase, alanine transaminase, aspartate transaminase, gamma-glutamyl transferase, creatinine, urea, testosterone, insulin-like growth factor 1, and dehydroepiandrosterone sulphate) and common risk factors (i.e., systolic blood pressure, blood pressure medications such as diuretics, beta-blockers, ace inhibitors, and calcium blockers, as well lipid lower medication).

To avoid losing information, we imputed the missing values for biomarkers (ranging from 2% to 42%), systolic blood pressure (17%), and demographic characteristics (ranging between 1% to 2%) with a multivariate imputation of chained equations – an imputation technique that assumes that the missing data are missing at random, i.e., the probability that a value is missing depends only on the observed values. [23] We then standardized the continuous parameters (i.e., biomarkers, systolic blood pressure, age, weight, height, BMI, waist circumference, and waist-to-height ratio) in the same scale and used one-hot encoding for the categorical variables.

We used 29 predictors, including 18 biomarkers (total cholesterol and high-density lipoprotein (HDL) cholesterol, triglycerides, glycated haemoglobin, high sensitivity C-reactive protein, clauss fibrinogen, haemoglobin, ferritin, albumin, alkaline phosphatase, alanine transaminase, aspartate transaminase, gamma-glutamyl transferase, creatinine, urea, testosterone, insulin-like growth factor 1, and dehydroepiandrosterone sulphate) and 11 risk factors (age, systolic blood pressure, sex, body mass index, waist-to-height ratio, smoking, diuretics, beta-blockers, ace inhibitors, calcium blockers, and lipid lower medication) to train several machine learning models. The chosen models are logistic regression, decision tree,[24] random forest,[25] extreme gradient boosting, [26] and Gaussian naïve Bayes.[27] They were trained on a random split of 80% of the data (i.e., training dataset) and tested on the remaining 20% (i.e., test dataset). Due to a severe unbalanced outcome (98% without CVD vs 2% with CVD), we stratified our sample – during the random split – by CVD status (with, without) to ensure that both training and test datasets have the same proportion of CVD cases (2%).

We then used synthetic minority oversampling (SMOTE) for the training dataset to adjust for class imbalance. This method generates additional samples of CVD cases that resemble the actual subjects with CVD.[28] As accuracy is not meant for highly unbalanced outcomes, we evaluated the models' performance on the test dataset using the following performance metrics: recall (sensitivity), precision, and F1-score. The latter is the combination of precision and recall and is used for model comparison.[29] Our goal is to maximise the rate of true positives – the proportion of correct CVD predictions out of those subjects with CVD. We present, here, the model with the highest recall, which is the Gaussian Naïve Bayes (GNB) classifier (with a smoothing variance for the Gaussian distribution of 0.9). We arrived at this model after performing a grid search with 10-fold cross-validation to tune hyperparameters (priors and variance smoothing) on the training dataset. Moreover, we carried out a SHAP (SHapley Additive exPlanations) analysis[30] to interpret the predictors of the GNB model. For comparison purposes, a logistic regression (LR) model– with the same predictors and over-oversampling algorithm for the outcome variable (CVD) – was trained and tested on the same training and test datasets respectively as the GNB model.

4. Results

Table 2 shows the participants' baseline demographic characteristics by cardio-vascular status at follow-up. Individuals with CVDs at follow-up were older than those without CVDs, and most were male and smokers, with a slightly higher BMI and waist-to-height ratio than those without CVDs.

Table 2. Participants' baseline demographic characteristics by cardiovascular status at follow-up

Characteristic	Statistic	Participants without CVD (N = 7,767)	Participants with CVD (N = 164)
Age (years)	n	7,767	164
	Mean (SD)	50.8 (15.2)	62.8 (10.8)
	Median	51	64
Sex, n (%)	Male	3,354 (43)	105 (64)
	Female	4,413 (57)	59 (36)
Height (cm)	N	7,726	163
	Mean (SD)	167.8 (9.4)	168.4 (9.4)
	Median	167.2	168.6
Weight (kg)	N	7,609	157
	Mean (SD)	78.5 (15.9)	84.7 (16.8)
	Median	76.9	83.7
BMI (kg/m^2)	n	7,601	157
	Mean (SD)	27.9 (5.1)	29.9 (5.5)
	Median	27.2	28.9
Waist-to-height ratio	n	7,677	157
	Mean (SD)	0.56 (0.08)	0.62 (0.08)
	Median	0.55	0.60
Smoking, n (%)	Yes	1,010 (15)	25 (19)

BMI: Body Mass Index; SD: Standard Deviation

Table 3 (see in Appendix) summarizes our samples' clinical characteristics, including biomarkers, systolic blood pressure, and blood pressure-related medications at baseline by cardiovascular status at follow-up. Participants with CVDs at follow-up had higher levels of several cardiovascular-related biomarkers at baseline (i.e., triglycerides, c-reactive protein, haemoglobin) than those without CVDs, although these were all within the normal range.[31] Likewise, participants with CVD at follow-up had higher levels of liver disease-related biomarkers (i.e., alkaline phosphatase, alanine transaminase, aspartate transaminase, gamma-glutamyl transferase) than those without CVD, although these were also within normal levels [32] at baseline. Similar patterns were observed for kidney disease-related biomarkers (i.e., creatinine and urea), diabetes-related biomarkers (i.e., glycated haemoglobin), and biomarkers related to hormones (i.e., higher for testosterone and lower for insulin-like-growth factor 1 and dehydroepiandrosterone sulphate). Nevertheless, participants who developed CVDs at follow-up had higher systolic blood pressure at baseline and had taken more blood pressure-related medications, such as diuretics, beta-blockers, ace inhibitors, calcium blockers, as well as lipid-lowering medications (to reduce cholesterol) than those who did not develop CVD at follow-up.

Table 4 shows the confusion matrix and performance metrics of both GNB and LR models. As seen, the GNB classifier exhibited 82% recall, implying that the model predicted correctly 82% (27/33) of all CVD cases. The false negative rate was 18% (6/33), while the false positive was 54% (834/1554). Moreover, its precision was 3% (27/861), meaning that out of all predictions, 3% were CVD cases. This is slightly higher than the baseline prevalence of CVD (2%) in our sample. In comparison, the recall of the logistic regression model was much lower (48%) than that of the GNB model. Accordingly, the LR's false positive and false negative rates were 52% (17/33) and 25% (392/1554), respectively. Its precision and F1-score were slightly better than the GNB's model.

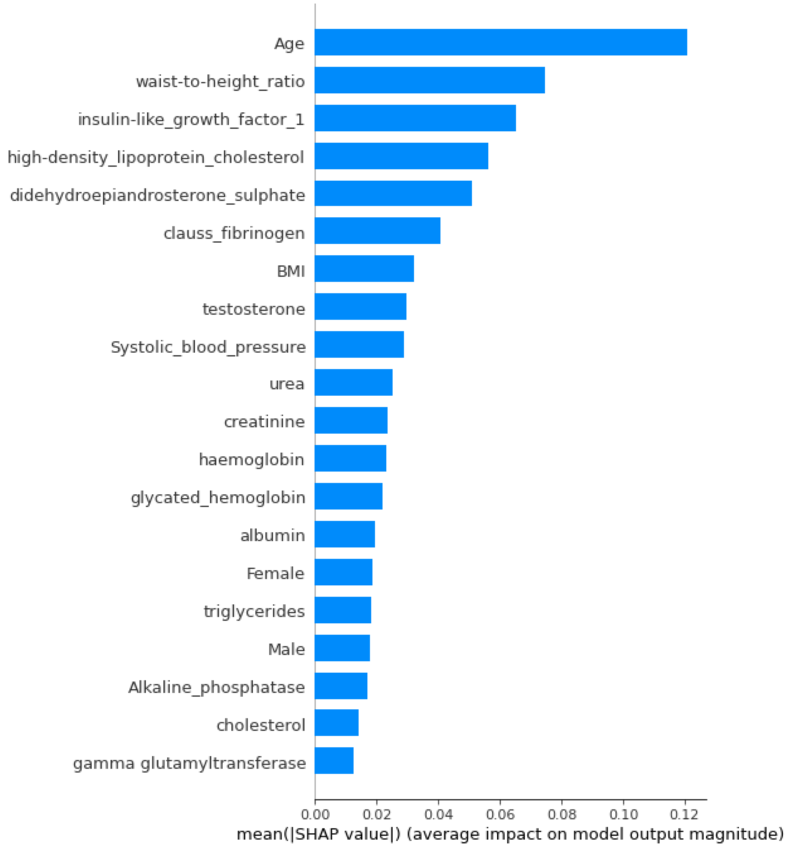
Figure 2 presents the mean absolute SHAP values for each feature (predictor) across all data. Features with higher mean SHAP values are the most influential, i.e., they contribute the most to GNB's predictions. Age was the most influential predictor of CVD followed by waist-to-height ratio, insulin-like growth factor 1, high-density lipoprotein cholesterol, didehydroepiandrosterone sulphate, and clauss fibrinogen. Body mass index (BMI), testosterone, and systolic blood pressure were less important followed by urea, creatinine, haemoglobin and glycated haemoglobin. Towards the bottom of the figure, we find other important predictors including albumin, sex (male, female), triglycerides, alkaline phosphatase, total cholesterol levels, and gamma-glutamyltransferase.

Table 4. Models' performance on the test dataset (n=1587)

Gaussian Naive Bayes		**Predicted**	
		No CVD	**CVD**
Observed	**No CVD**	720	834
	CVD	6	27
Recall (%)	82		
Precision (%)	3		
False positive rate (%)	54		
False negative rate (%)	18		
F1-score	0.06		
Logistic regression		**Predicted**	
		No CVD	**CVD**
Observed	**No CVD**	1162	392
	CVD	17	16
Recall (%)	48		
Precision (%)	4		
False positive rate (%)	52		
False negative rate (%)	25		
F1-score	0.07		

CVD: Cardiovascular disease. Recall: The percentage of correctly predicted CVD cases of those CVD cases observed. Precision: The percentage of correctly predicted CVD cases of the total (CVD and non-CVD) predicted. False positive rate: The percentage of incorrectly predicted CVD cases of all no-CVD cases observed. False negative rate: The percentage of incorrectly predicted non-CVD cases of all CVD cases observed.

Figure 2. Features' influence on GNB's predictions by order of importance



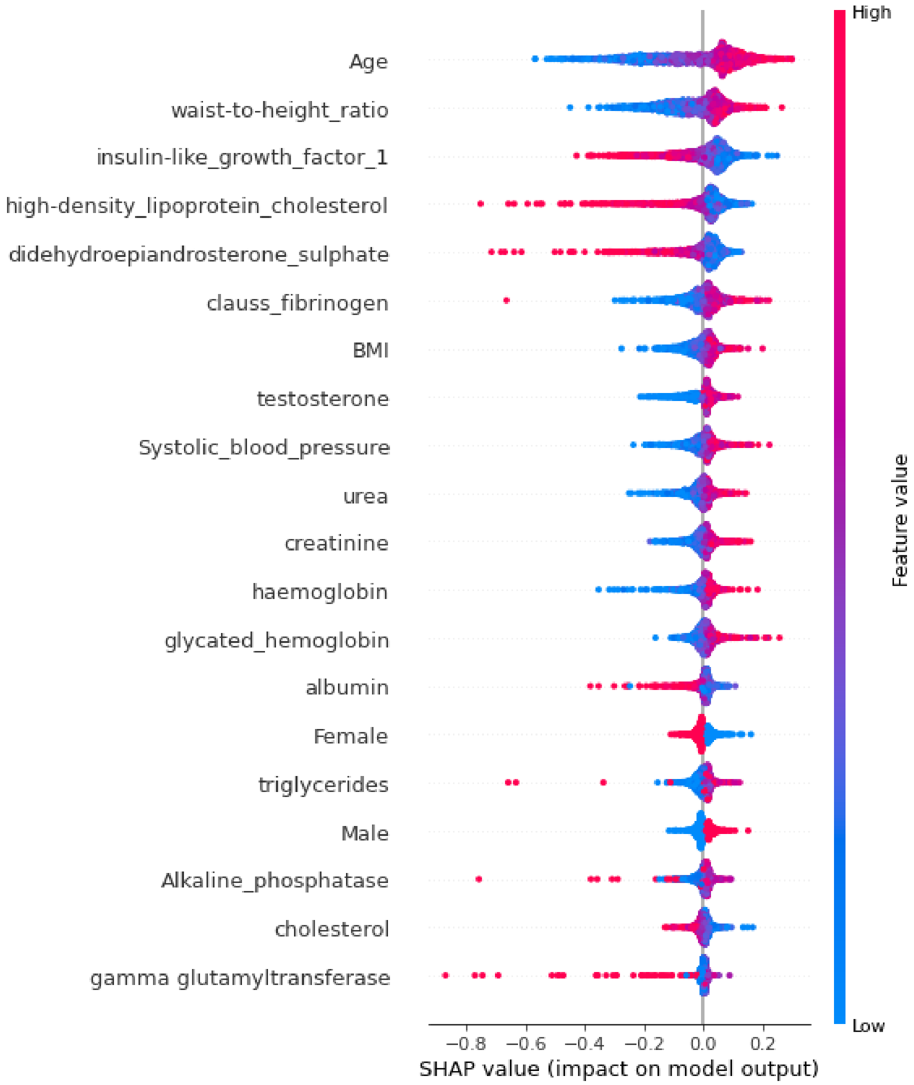
The above biomarkers can be related to underlying comorbidities as shown in Table 5.

Table 5. Biomarkers' underlying comorbidities by order of importance

Biomarkers	Comorbidities
Insulin-like growth factor 1 (nmol/l)	Cardiovascular disease
High-density lipoprotein cholesterol(mmol/l)	Cardiovascular disease
Dihydroepiandrosterone sulphate (µmol/l)	Cardiovascular disease
Clauss fibrinogen (g/l)	Cardiovascular disease
Testosterone (nmol/l)	Diabetes (men), polycystic ovarian syndrome (women)
Urea (mmol/l)	Kidney disease
Creatinine (µmol/l)	Kidney disease
Haemoglobin (g/l)	Cardiovascular disease
Glycated haemoglobin (mmol/mol)	Diabetes
Albumin (g/l)	Liver disease
Triglycerides (mmol/l)	Cardiovascular disease
Alkaline phosphatase(u/l)	Liver disease
Cholesterol (mmol/l)	Cardiovascular disease
Gamma-glutamyl transferase(u/l)	Liver disease

Figure 3, known as beeswarm plot, shows not only the relative importance of each feature but also its relationship with the predicted outcome (CVD).

Figure 3. Relationship of the biomarkers and personal factors with CVDs.



High age values (red colour) had a positive impact on model's output, i.e., a higher likelihood of CVD, while low values (blue colour) had a negative impact, i.e., lower likelihood of CVD. Such an association (i.e., high values of a feature associated with higher chance of CVD) was also observed for waist-to-height ratio, clauss fibrinogen, BMI, testosterone, systolic blood pressure, urea, creatinine, haemoglobin, glycated heamoglobin, and triglycerides. Male subjects were more likely to develop CVD than female ones. In contrast, low (blue) values of insulin-like growth factor 1 were associated with higher chance of CVD. The same was observed for low high-density lipoprotein cholesterol values, dihydroepiandrosterone sulphate, albumin, and total cholesterol levels.

5. Discussion

In this study we trained a Gaussian Naïve Bayes (GNB) model to identify a set of biomarkers from a UK population without CVDs at baseline and predicted—with high sensitivity (82%) —CVD cases ten years later In comparison, Blankenberg's, Shlipak's and Wang's proportional hazard ratio models [15, 11, 14] achieved a sensitivity of 70%, 75% and 80% respectively. GNB is a simple and fast algorithm that performs well with predicting 'zero probability' phenomena such as the rare occurrence of disease.[27] Several of the biomarkers and risk factors identified (i.e., age, sex, systolic blood pressure, body mass index) have also been reported in other studies[10-14] corroborating our findings. Additionally, our findings suggest that liver and kidney disease-related biomarkers (i.e., albumin, alkaline phosphatase, urea and creatinine) and glycated haemoglobin - a diabetes-related biomarker - are also important predictors of CVDs. These, along with haemoglobin, were also associated with increased odds of respiratory treatment,[33] supporting the link between liver disease, diabetes, lung disease and CVD.[33-37] Unlike previous studies,[10-15] we took a wider definition of CVDs to include congestive heart failure, coronary heart disease, angina, heart attack or myocardial infarction, and stroke. While previous studies[10-15] included myocardial infarction, stroke, or CVD death. This broader definition allowed us to obtain a more comprehensive understanding of CVD, including causes not previously captured in studies, thus leading to more accurate predictions; one that extends to causes (e.g., heart failure, coronary heart disease, angina, and heart attack). Despite the lower prevalence of CVD at follow-up (2%) in comparison to that of previous studies[10-15], where the incidence of CVD at follow-up ranged from 4% to 16% (Table 1), our GNB classifier was able to predict correctly 82% of observed CVD cases (aka. recall or sensitivity; Table 4). This was much better than the respective 48% recall of the logistic regression model. When preventing CVDs is the primary objective, a high false positive rate (FPR) is rather preferred to a high false negative rate, as a falsely predicted CVD case may lead to further testing to confirm the initial diagnosis. In contrast, by failing to predict a true CVD case can be irreversible and even fatal. To this end, both models had higher false positive than false negative rates, although these rates were slightly better for the GNB classifier than those of the logistic regression, i.e., higher false positive rate (54% vs 52%) and lower false negative rate (18% vs 25%).

Despite the noteworthy outcomes, the study has some limitations. First, the data lacked information on conventional cardiovascular biomarkers (e.g., N-BNP, MR-proADM, Troponin) already associated with increased odds of CVD events in several studies.[10-12, 15] The addition of the above-mentioned biomarkers is likely to further improve the precision of our model's predictions. The reason for the low precision (3% for GBM and 4% for logistic regression) can also be due to data imbalance. Although this was tackled in the training dataset using a commonly used resampling method (SMOTE), our test dataset - where the models' performance was assessed - was still highly imbalanced due to low CVD prevalence (2%). Further research into resampling methods and inclusion of the above-mentioned important biomarkers (e.g. troponin) would likely help to improve our model's precision. Second, the GNB classifier assumes independence of the predictors,[27] which is not a valid assumption as several biomarkers have some degree of correlation. Nevertheless, this did not challenge our results given the inferences drawn from the SHAP analyses are rational. The only spurious association was the inversed association between low cholesterol levels and increased risk of CVD. This association can be attributed to confounding that almost half of the subjects with CVD had received lipid-lowering medication (Table 3) to reduce their cholesterol levels.

6. Conclusions

Our findings suggest that using Machine Learning can improve healthcare systems in several ways. (1) Early CVD diagnosis and prediction can be achieved by incorporating demographic and clinical characteristics such as age, sex, BMI and routine-collected biomarkers to identify patterns and relationships that may be indicative of CVD or other conditions in general. This can help clinicians make more accurate diagnoses and predict the likelihood of developing CVD in the future. (2) A personalized treatment plan based on patients' medical history, lifestyle and response to previous treatments can be recommended. Medical practitioners can optimize existing treatments for individual patients resulting in better outcomes and fewer side effects. (3) Patients' health including weight and vital signs can be monitored remotely via wearable devices as well by having regular blood tests for biomarkers assessment. This will help healthcare providers to identify signs of health worsening, intervene early and prevent hospitalizations or future heart failure.

Overall, the findings demonstrate that early CVD diagnosis is possible with the identified biomarkers, which allows CVD prevention and relinquishes the burden on society and healthcare systems.

7. Declaration of competing interest

The authors have no competing interest to declare.

8. Funding statement

There was no funding for this study.

9. References

1. NHS - cardiovascular disease (https://www.nhs.uk/conditions/cardiovascular-disease/). Accessed on 12/08/2022
2. WHO - cardiovascular diseases (CVDs) (https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)) Accessed on 12/08/2022.
3. Colombi AM, Wood GC. Obesity in the workplace: Impact on cardiovascular disease, cost, and utilization of care. American Health & Drug Benefits. 2011 Sep;4(5):271.
4. Kahn R, Robertson RM, Smith R, Eddy D. The impact of prevention on reducing the burden of cardiovascular disease. Circulation. 2008 Jul 29;118(5):576-85.
5. Ridker, P. M., Buring, J. E., Rifai, N., & Cook, N. R. (2007). Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. Jama, 297(6), 611-619.
6. Kavanagh A, Bentley RJ, Turrell G, Shaw J, Dunstan D, Subramanian SV. Socioeconomic position, gender, health behaviours and biomarkers of cardiovascular disease and diabetes. Social science & medicine. 2010 Sep 1;71(6):1150-60.
7. Ryan M, Gallagher S, Jetten J, Muldoon OT. State level income inequality affects cardiovascular stress responses: Evidence from the Midlife in the United States (MIDUS) study. Social Science & Medicine. 2022 Oct 1;311:115359.
8. Alessie RJ, Angelini V, van den Berg GJ, Mierau JO, Viluma L. Economic conditions at birth and cardiovascular disease risk in adulthood: Evidence from post-1950 cohorts. Social Science & Medicine. 2019 Mar 1;224:77-84.
9. Browning CR, Cagney KA, Iveniuk J. Neighborhood stressors and cardiovascular health: Crime and C-reactive protein in Dallas, USA. Social science & medicine. 2012 Oct 1;75(7):1271-9.
10. Melander O, Newton-Cheh C, Almgren P et al. Novel and conventional biomarkers for prediction of incident cardiovascular events in the community. Jama. 2009 Jul 1;302(1):49-57.
11. Shlipak M. G., Ix J. H., Bibbins-Domingo K., Lin F., & Whooley M. A. (2008). Biomarkers to predict recurrent cardiovascular disease: the Heart and Soul Study. The American journal of medicine, 121(1), 50-57.
12. Zethelius B, Berglund L, Sundström Jet al. Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. New England Journal of Medicine. 2008 May 15;358(20):2107-16.
13. Folsom AR, Chambless LE, Ballantyne CM et al. An assessment of incremental coronary risk prediction using C-reactive protein and other novel risk markers: the atherosclerosis risk in communities study. Archives of internal medicine. 2006 Jul 10;166(13):1368-73.
14. Wang TJ, Gona P, Larson MG et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. New England Journal of Medicine. 2006 Dec 21;355(25):2631-9.
15. Blankenberg S, McQueen MJ, Smieja M et al. Comparative impact of multiple biomarkers and N-Terminal pro-brain natriuretic peptide in the context of conventional risk factors for the prediction of recurrent cardiovascular events in the Heart Outcomes Prevention Evaluation (HOPE) Study. Circulation. 2006 Jul 18;114(3):201-8.
16. Cao J, Li J, Gu Z et al. Combined metabolomics and machine learning algorithms to explore metabolic biomarkers for diagnosis of acute myocardial ischemia. International Journal of Legal Medicine. 2022 Mar 29:1-2.
17. Alaa AM, Bolton T, Di Angelantonio E, Rudd JH, Van der Schaar M. 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.
18. Nikolaou V, Massaro S, Garn W, Fakhimi M, Stergioulas L, Price D. The cardiovascular phenotype of Chronic Obstructive Pulmonary Disease (COPD): Applying machine learning to the prediction of cardiovascular comorbidities. Respiratory Medicine. 2021 Sep 1;186:106528.
19. The UK Household Longitudinal Study. Available online: https://www.understandingsociety.ac.uk/ (accessed on 31 July 2022).
20. Browning LM, Hsieh SD, Ashwell M. A systematic review of waist-to-height ratio as a screening tool for the prediction of cardiovascular disease and diabetes: 0.5 could be a suitable global boundary value. Nutr Res Rev. 2010;23(02):247-269
21. Ribeiro RC, Coutinho M, Bramorski MA, Giuliano IC, Pavan J. Association of the waist-to-height ratio with cardiovascular risk factors in children and adolescents: the Three Cities Heart Study. Int J Prev Med. 2010;1(1):39-49

22. Khoury M, Manlhiot C, McCrindle BW. Role of the waist/height ratio in the cardiometabolic risk assessment of children classified by body mass index. J Am Coll Cardiol. 2013;62(8):742-751
23. Van Buuren, S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. doi: 10.18637/jss.v045.i03
24. Rokach, Lior; Maimon, O. (2014). Data mining with decision trees: theory and applications, 2nd Edition. World Scientific Pub Co Inc. doi:10.1142/9097
25. Breiman, L. Random Forests. Machine Learning 45, 5-32 (2001).
26. Chen T, He T, Benesty M et al. Xgboost: extreme gradient boosting. R package version 0.4-2. 2015 Aug 1;1(4):1-4
27. Bhuvaneswari R, Kalaiselvi K. Naive Bayesian classification approach in healthcare applications. International Journal of Computer Science and Telecommunications. 2012 Jan;3(1):106-12.
28. Zhu T, Lin Y, Liu Y. Synthetic minority oversampling technique for multiclass imbalance problems. Pattern Recognition. 2017 Dec 1;72:327-40.
29. Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". Journal of Machine Learning Technologies. 2 (1): 37-63
30. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems. 2014 Dec;41(3):647-65.
31. Benzeval M., Davillas A., Kumari M., Lynn P. (2014). “Understanding Society: The UK Household Longitudinal Study. Biomarker User Guide and Glossary”.
32. Nikolaou V, Massaro S, Fakhimi M, Garn W. Using Machine Learning to Detect Theranostic Biomarkers Predicting Respiratory Treatment Response. Life. 2022 May 24;12(6):775.
33. Chang, W.H., Mueller, S.H., Chung, S.-C., Foster, G.R., Lai, A.G. Increased burden of cardiovascular disease in people with liver disease: Unequal geographical variations, risk factors and excess years of life lost. J. Transl. Med. 2022, 20, 2
34. Petrie, J., Guzik, T.J., Touyz, R.M. Diabetes, Hypertension, and Cardiovascular Disease: Clinical Insights and Vascular Mechanisms. Can. J. Cardiol. 2018, 34, 575–584
35. Kong, K.A., Jung, S., Yu, M., Park, J., Kang, I.S. Association Between Cardiovascular Risk Factors and the Severity of Coronavirus Disease 2019: Nationwide Epidemiological Study in Korea. Front. Cardiovasc. Med. 2021, 8
36. Ssentongo, P., Ssentongo, A.E., Heilbrunn, E.S., Ba, D.M., Chinchilli, V.M. Association of cardiovascular disease and 10 other pre-existing comorbidities with COVID-19 mortality: A systematic review and meta-analysis. PLoS ONE 2020, 15, e0238215
37. Viglino, D., Jullian-Desayes, I., Minoves, M. et al. Nonalcoholic fatty liver disease in chronic obstructive pulmonary disease. Eur. Respir. J. 2017, 1, 49

APPENDIX

Table 1. Overview of research investigating biomarkers related to the prediction of CVD events.

Study	Study design/Population	Outcomes	Incidence of CVDs at follow-up	Biomarkers associated with CVDs	Risk factors	CVDs definition
Melander, 2009 [10]	Prospective observational study/ 5,067 participants without CVD	Cardiovascular and coronary events after 12.8 years of follow-up	8% cardio-vascular events and 5% coronary events	CRP and N-BNP for cardiovascular events and MR-proADM and N-BNP for coronary Events	Age, sex, systolic blood pressure, diastolic blood pressure, use of antihypertensive therapy, current smoking, diabetes, HDL, body mass index	Myocardial infarction, Stroke, Coronary death
Shlipak, 2008 [11]	Prospective observational study/ 979 patients with pre-existing coronary artery disease	Cardiovascular events after 3.5 years of follow-up	15%	N-BNP, albuminuria, CRP	Demographic, lifestyle, and behavior variables; Cardio-vascular risk factors; Cardiovascular disease severity; medication use; Left ventricular ejection fraction	Stroke, Myocardial infarction, Coronary heart disease death
Zethelius, 2008 [12]	Prospective observational Uppsala Longitudinal Study of Adult Men (ULSAM)/ 1,135 participants; 661 of them without CVD	CVD death after 10 years of follow-up	12%	Troponin I, N-BNP, cystatin C, C-reactive protein	Age, systolic blood pressure, use or non-use of antihypertensive treatment, total cholesterol, HDL use or non-use of lipid-lowering treatment, presence or absence of diabetes, smoking status, body-mass index	CVD death
Folsom, 2006 [13]	Prospective observational Atherosclerosis Risk in Communities (ARIC) Study/ 15,792 participants without CHD	CHD after 16 years of follow-up TNR	4%	C-reactive protein	Age, race, sex, total and HDL, systolic blood pressure, anti-hypertensive medication use, smoking status, and diabetes	Myocardial infarction, Fatal CHD, or coronary revascularization
Wang, 2006 [14]	Prospective observational Framingham Heart Study/ 3,209 participants without CVD event	Any cause death and CVD event after 7 years of follow-up	6% for CVD death and 5% for major CVD event	B-type natriuretic peptide level, C-reactive protein level, urinary albumin-to-creatinine ratio, homocysteine level, renin level for any cause of death/ B-type natriuretic peptide level, the urinary albumin-to-creatinine ratio for major CVD events	Age, sex, cigarette smoking, blood pressure, use of antihypertensive therapy, total cholesterol, HDL, diabetes, body mass index, serum creatinine level	Fatal and nonfatal myocardial infarction, coronary insufficiency (prolonged angina with documented electrocardiographic changes), heart failure, stroke
Blankenberg, 2006 [15]	Multicenter, randomized, clinical trial Heart Outcomes	CVD event after 4.5 years of follow-up	16%	N-BNP, soluble intercellular adhesion molecule-	Age, sex, the ratio of LDL to HDL cholesterol,	Myocardial infarction, Stroke, Cardiovascular death

Study	Study design/Population	Outcomes	Incidence of CVDs at follow-up	Biomarkers associated with CVDs	Risk factors	CVDs definition
	Prevention Evaluation (HOPE) Study/ 3,199 participants with the previous CAD			1, microalbuminuria, soluble in-terleukin-1 receptor antagonist, fibrinogen	diabetes mellitus, smoking status, systolic blood pressure, waist-hip ratio, triglycerides, glucose, creatinine, microalbuminuria, lipid-lowering drugs, ramipril allocation, and peripheral vascular disease	

CVDs: Cardiovascular diseases; CAD: Coronary artery disease; CHD: Coronary heart disease; CRP: C-reactive protein; N-BNP: N-terminal pro-B-type natriuretic peptide; MR-proADM: mid-regional proadrenomedullin; HDL: high-density lipoprotein cholesterol level.

Table 3. Participants' baseline clinical characteristics by cardiovascular status at follow-up

Characteristic	Statistic	Participants without CVD (N = 7,767)	Participants with CVD (N = 164)
Biomarkers at baseline			
Cholesterol (mmol/l)	n	7,632	162
	Mean (SD)	5.5 (1.1)	5.2 (1.3)
	Median	5.4	5.0
HDL cholesterol (mmol/l)	n	7,621	162
	Mean (SD)	1.6 (0.5)	1.4 (0.4)
	Median	1.5	1.3
Triglycerides (mmol/l)	n	7,636	162
	Mean (SD)	1.8 (1.2)	2.1 (1.1)
	Median	1.5	1.9
Glycated haemoglobin (mmol/mol)	n	7,237	151
	Mean (SD)	36.7 (7.3)	41.1 (10.2)
	Median	36	38
High sensitivity c-reactive protein (mg/l)	n	7,425	161
	Mean (SD)	2.9 (6.7)	4.1 (7.4)
	Median	1.4	2.1
Clauss fibrinogen (g/l)	n	7,608	160
	Mean (SD)	2.7 (0.6)	2.9 (0.6)
	Median	2.7	2.9
Haemoglobin (g/l)	n	7,235	151
	Mean (SD)	137.3 (13.4)	140 (14.3)
	Median	137	137
Ferritin (ug/l)	n	7,633	162
	Mean (SD)	136.5 (177.0)	172.9 (169.5)

Characteristic	Statistic	Participants without CVD (N = 7,767)	Participants with CVD (N = 164)
	Median	100	135
Albumin (g/l)	n	7,647	162
	Mean (SD)	46.9 (2.7)	46.4 (2.6)
	Median	47	46
Alkaline phosphatase(u/l)	n	7,566	159
	Mean (SD)	70.4 (21.8)	74.5 (20.2)
	Median	68	72
Alanine transaminase(u/l)	n	7,561	159
	Mean (SD)	28.4 (26.7)	32.6 (17.4)
	Median	24	28
Aspartate transaminase(u/l)	n	7,314	155
	Mean (SD)	30.6 (28.7)	34.4 (11.7)
	Median	29	32
Gamma glutamyl transferase (u/l)	n	7,584	159
	Mean (SD)	33.9 (52.3)	43.9 (35.4)
	Median	23	32
Creatinine (μmol/l)	n	7,645	162
	Mean (SD)	75.3 (17.6)	81.8 (19.5)
	Median	73	81
Urea (mmol/l)	n	7,649	162
	Mean (SD)	6.2 (1.6)	6.7 (1.7)
	Median	6	6.4
Testosterone (nmol/l)	n	4,496	113
	Mean (SD)	11.7 (7.9)	12.7 (6.2)
	Median	12.5	13.2
Insulin-like growth factor 1 (nmol/l)	n	7,598	160
	Mean (SD)	18.3 (6.9)	15.2 (5.0)
	Median	17	15
Dehydroepiandrosterone sulphate (μmol/l)	n	7,623	161
	Mean (SD)	4.6 (3.1)	3.6 (2.7)
	Median	3.9	2.8
Blood pressure and medications at baseline			
Systolic blood pressure (mmhg)	n	6,480	130
	Mean (SD)	126.0 (16.1)	132.5 (16.6)
	Median	124.5	131.0

Characteristic	**Statistic**	**Participants without CVD (N = 7,767)**	**Participants with CVD (N = 164)**
Diuretics, n (%)	Yes	550 (7)	33 (20)
Beta blockers, n (%)	Yes	404 (5)	36 (22)
Ace inhibitors, n (%)	Yes	652 (8)	44 (27)
Calcium blockers, n (%)	Yes	496 (6)	33 (20)
Lipid lowering medication, n (%)	Yes	1,108 (14)	78 (48)

SD: Standard Deviation; CVDs: Cardiovascular diseases