Investigating factors related to cardiovascular disease in the UK Biobank cohort

Luis Chaves, MEng; Catriona Mackay, BSc; Fatima Shahid, BSc

Code available in GitHub

Abstract

Introduction. One third of all deaths each year are cause by cardiovascular disease (CVD). CVD is highly dependent on lifestyle and environmental exposures. In this study we employ an 'omics' approach to investigate the factors most associated with CVD as well as candidate models for risk score tools. *Methods.* Our research focuses on CVD prevalence in the UK population, sourcing data from UK Biobank. This resource contains biomarker, genetic and basic demographic variables on 500,000 participants from across the UK. A biological health score (BHS) and polygenic risk score (PRS) were computed to aggregate higher-dimensional data. We then built a range of statistical models to determine the biological fingerprint of CVD as well as the most predictive factors and model of CVD. *Results.* Only 3.1% of our population (median age 57; 53.5% female) had CVD at recruitment. Our models identified cystatin C and apolipoprotein B as the biomarkers most positively associated with CVD (OR: 1.77 (95% CI: 1.73-1.80) and OR: 1.78 (95% CI: 1.68-1.89) respectively). Different CVD subtypes manifested differing biomarker signatures, as shown by sPLS analyses. Female sex was the greatest protective factor for CVD, compared to males (OR: 0.54 (95% CI: 0.52-0.55). We obtained an AUC of 0.75 when combining information on demographics, biomarkers and the PRS, using a LASSO logistic regression model. Using a more sophisticated machine learning model (XGBoost) did not improve predictive performance. *Discussion.* We identified biomarker and demographic risk factors, consistent with previous research. Differing biomarker signatures between CVD subtypes could enable more specific risk prediction and better targeting of interventions. We obtained the greatest predictive performance when combining a range of data types: aggregate risk scores may be useful at summarising complex and high-dimensional data for predictive models.

1 Introduction

According to the World Health Organisation (WHO), 31% of all deaths in 2016 were caused by cardiovascular disease (CVD) [1]. The majority of those (85%) were due to heart attack or stroke. In addition to early mortality, CVD causes life-long disability, and has high cost to both the NHS and the economy: the total healthcare cost of CVD in the UK is estimated at £9 billion each year [2]; the non-healthcare cost to the UK economy is estimated to be between £15.8 and £19 billion [2][3]. Risk of CVD is heavily dependent on lifestyle factors such as diet, smoking and exercise: there is considerable evidence that a reduction in harmful lifestyle factors would reduce CVD incidence [4][5]. As highlighted in [6], the greatest clinical danger that CVD presents is the build-up of atherosclerotic plaque, which is the underlying factor that ultimately triggers most CVD events. Atherosclerosis develops slowly and can remain unnoticed if the patient is not closely followed-up.

Due to these factors, many healthcare systems focus on CVD prevention. The NHS Long Term Plan sets CVD prevention as a clinical priority and aims at preventing 150,000 CVD events in the next 10 years [5]. This is set to be achieved through early screening and patient information campaigns that promote healthier lifestyles.

In 2010, Rappaport & Smith expanded Wild's [7] definition of the *exposome* to include "all chemicals that reach the internal chemical environment" [7][8]. With this conceptual break-through, everything that interacts with an organism, and that can be measured, falls under the umbrella of omics. When it comes to CVD, an omics approach can help understand not only the effect of genetic variants (like with genome-wide association studies (GWAS)), but also the effect of exposures such as diet and lifestyle, which will be translated to biomarker fluctuations along with changes in the epigenome and the transcriptome. From a clinical perspective, omics can help improve personalised medicine and enable earlier detection of diseases through screening.

All data in this study was sourced from the UK Biobank (UKBB). We first explored the demographic and biological factors associated to CVD, using a range of statistical methods. We then investigated what predictive model, and what combination of data (biomarkers, SNPs, demographic covariates), could best predict CVD prevalence.

2 Methods

2.1 Study population & data source

All data used in this study was obtained from UKBB, a large-scale, prospective population-based cohort study [9]. Over 500,000 participants between the ages of 40 and 69, of both sexes, were recruited between 2006 and 2010.

We focused on three datasets: a biomarker set, a demographic covariates set and a single-nucleotide polymorphism (SNP) set. Hereafter, we will refer to these as the biomarker, covariate and SNP datasets respectively.

We focused primarily on the biomarker dataset, which contains measurements for 30 biomarkers, such as cholesterol and glucose, at two separate time points, for 502,536 individuals (see Table C.2 for the full list). The covariates dataset contained data on 38 demographic variables, such as age, gender and body mass index (BMI), for 366,749 individuals (see Table C.1 for a list of available variables). The third dataset consisted of the number of copies of 177 CVD-related SNPs, for 487,409 individuals. This dataset was generated using results from a GWAS that identified 202 SNPs significantly associated with coronary artery disease after multiple testing correction (CARDIoGRAMplusC4D) [10]. This GWAS did not include UKBB participants, ensuring no overlap in participants with our study. UKBB participants were genotyped, and genotyping data processed as previously described [11]; genotyping data was available for 177 of the SNPs identified.

The main outcome explored was CVD status: a binary variable indicating whether a patient suffers from CVD. Our cases were individuals with CVD, and controls those without CVD. In this sense, our study is a cross-sectional study of the UKBB cohort at the time of recruitment.

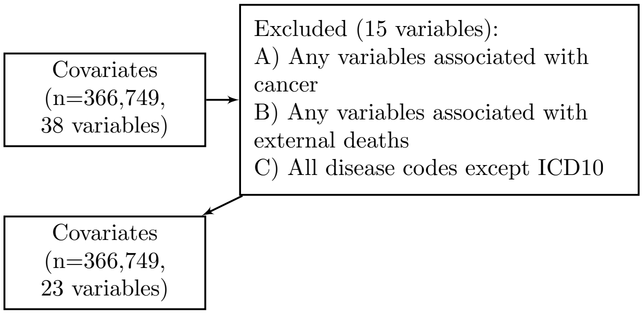

Covariates (n=366,749, 38 variables)
Excluded (15 variables):
A) Any variables associated with cancer
B) Any variables associated with external deaths
C) All disease codes except ICD10
Covariates (n=366,749, 23 variables)

Figure 1: Exclusion Criteria for Covariates Dataset

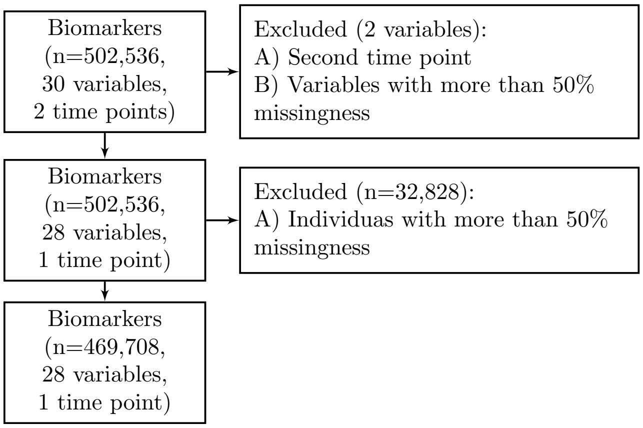

Biomarkers (n=502,536, 30 variables, 2 time points)
Excluded (2 variables):
A) Second time point
B) Variables with more than 50% missingness
Biomarkers (n=502,536, 28 variables, 1 time point)
Excluded (n=32,828):
A) Individuas with more than 50% missingness
Biomarkers (n=469,708, 28 variables, 1 time point)

Figure 2: Exclusion Criteria for Biomarkers Dataset

2.2 Exclusion criteria & preprocessing

Given our focus on CVD, we removed all variables related to diseases other than CVD (e.g. cancer and external deaths) from the covariates dataset. We also removed all disease codes other than ICD-10 (International Classification of Diseases, 10$^{\text{th}}$ revision), the most up-to-date code available. This left us with 23 variables (see Table C.1).

For the biomarkers dataset, we selected only the measurements at recruitment and dropped variables with more than 50% missingness, (Rheumatoid factor and oestradiol, each with around 80% missingness). Participants with more than 50% of measurements missing were removed, leaving 469,708 individuals. For all remaining variables, missingness was below 25%; we imputed the remaining missing values using K-nearest neighbours, assuming a missing-at-random (MAR) mechanism. We decided to impute based on the low prevalence of CVD in our cohort (see Section 3.1). We explored estimating the error of imputation based on a method demonstrated in [12], which can be found in the Appendix (Section C.9).

2.3 Feature engineering

2.3.1 Biological health score

The biological health score (BHS) was first defined in [13]. We used this definition and method to calculate the BHS in our study. Details on how this is performed can be found in the Appendix (Section B.1). In this study, we used three different sets of biomarkers (A, B and C - see Table C.2) to define three different BHS. Additionally, a BHS score was available in the UKBB covariates dataset (referred to as 'BS2'). To make the three references, each biomarker was considered to be harmful or protective in regards to CVD. Reference A results from the overlap of the biomarkers available in UKBB and those used in [14]. The predictive power of each BHS was assessed by determining whether it improved predictive performance when added alongside other covariates (as described in Section 2.4.4). Additionally, the odds ratios (OR) and p-values for the BHS were extracted from these models.

2.3.2 Polygenic risk score

For each participant, a polygenic risk score (PRS) for CVD was computed. The PRS is a weighted average of risk alleles (SNPs) present in each individual: the average number of copies of each SNP (0, 1 or 2), weighted by their effect size. As such, SNPs with a strong association with CVD (a large effect size) will contribute more to the PRS, and a higher PRS corresponds to a greater risk of disease. Risk alleles and their effect sizes were obtained using summary data from the CARDIoGRAMplusC4D GWAS [10]. The PRS was then scaled to lie between -1 and 1, so that a PRS of 0 would indicate neutral risk. The relationship between the PRS and CVD status was assessed in the same way as for the BHS (see previous section).

2.4 Models

2.4.1 Biomarkers univariate analysis

To assess whether the association of each biomarker to CVD status was statistically significant, we fitted a series of logistic regression models with CVD status as the response variable and the respective biomarker as the main explanatory variable. The models were adjusted on age, gender, education quality, ever-smoking status, level of physical activity, level of alcohol consumption, BMI, number of comorbidities and number of medicines currently being taken. From each model, the p-value and effect size for the biomarker was extracted. The p-values were then compared to the Bonferroni-adjusted significance threshold (0.05/number of tests): 28 tests were performed, giving a Bonferroni-adjusted significance level of $1.79 \cdot 10^{-3}$.

2.4.2 Penalised regression

We built penalised regression models that performed feature selection on the data. We performed 100 subsampling iterations of LASSO logistic regression. At each iteration a random sample of 80% of the data was used for cross-validation to find the optimal regularisation parameter, λ. We considered the optimal λ to be the maximum value that ensures the loss to be one standard error away from the minimum loss (known as the 1SE rule), since it has previously been shown that the λ that ensures minimal cross-validation error can result in over-fitting [15].

This method of analysis was implemented for: the biomarker set alone, the biomarker set with the covariates and PRS, and the covariates set with the BHS and PRS.

2.4.3 Partial least squares

Partial least squares discriminant analysis (PLS-DA) was used to assess the biomarker signal associated with CVD status. Variable selection was performed using the sparse (sPLS-DA) extension of this method [16], to determine which biomarkers most strongly relate to CVD status.

The sPLS-DA model was first calibrated, finding the optimal number of variables to be selected by the model (minimising the CVD misclassification rate). The sPLS-DA model was then fitted, and the contribution of the selected biomarkers to the sPLS-DA response component was evaluated using their loading coefficients. This enabled identification of the variables which differ most between CVD cases and controls.

A stability analysis was conducted, to assess consistency of biomarker selection across varying sPLS-DA models. A sub-sampling procedure with 100 iterations was performed, as described in [17]: within each subsample, the number of variables selected by the model was varied from 1 to 28 (total number of biomarkers); the biomarkers selected each time were stored.

Stratified analyses were also performed, to determine whether different subtypes of CVD exhibited differing biomarker signatures. For this analysis, the five most prevalent CVD subtypes in this population were selected using ICD-10 codes: G459 (transient cerebral ischaemic attack), I209 (angina pectoris), I219 (acute myocardial infarction), I251 (atherosclerotic heart disease) and I639 (cerebral infarction). Five datasets were created, each containing the cases of one subtype and all controls. Five sPLS-DA models were then fitted on each of these datasets, using the calibrated parameter value. Loading coefficients of the biomarkers selected by each model were evaluated.

2.4.4 Comparison of predictive model performance

Providing a disease risk score can be useful in clinical settings, even if the underlying disease mechanism is not fully understood. In this section, we assessed the predictive performance of several models, training each of them on different sets of data. We tested the extreme-gradient boosted (XGBoost) ensemble method and a penalised classifier. The different sets of data comprised all possible combinations of biomarkers, covariates, PRS and BHS, excluding combinations that included both the BHS and biomarkers data. Each model was trained by cross-validation on 5 identical folds of 80% of the original data; the AUC of the predictions on the testing set (the remaining 20% of the data) is reported for the best model (hypertuned by cross-validation). This is better illustrated in Table 1.

3 Results

3.1 Exploratory data analysis

General statistics about our population are presented in Table 2.

The median age of our population is 57 years old, with 53.3% of individuals being female. 3.1% of the population have CVD. The majority of the population (65.6%) have a BMI over 25, which is considered clinically as overweight. CVD cases were found to be older than controls, with an average difference of 4.47 years. Cases were also found to be predominantly male (68%), less educated (22.6% of cases have a low-quality education versus 13.9% of controls), and displayed poorer lifestyle factors (53.6% of cases are/have been smokers versus 43.2% of

	Logit	SVM	XGBoost
Bio	🟧	🟧	🟧
Cov	🟨	🟨	🟨
Bio+ Cov	🟠	🟠	🟠
Cov+ BHS	🟨	🟠	🟨
Cov+ PRS	🟨	🟠	🟨
Cov+ PRS+ BHS	🟧	🟠	🟨
Cov+ Bio+ PRS	🟨	🟨	🟥

Table 1: Illustration of the method used to compare models. *Note: 1. The colours are simply to represent that the values within those cells will differ; 2. Bio stands for biomarkers, Cov stands for Covariates.*

controls; 77% of cases have a BMI over 25 versus 65.2% of controls).

Biomarker distributions were assessed (Figure C.5), with most biomarkers displaying similar distributions for cases and controls. However, some differences can be observed: testosterone follows a bi-modal distribution (with the left peak representing females and the right peak males), showing that a greater proportion of males have CVD than females. Other biomarker distributions are marginally shifted, such as for urate and alanine aminotransferase.

3.2 Feature engineering

3.2.1 Biological health score

The three new BHS we calculated were plotted by CVD status, alongside the BS2 score provided in UKBB (Figure C.1). The effect size (OR), 95% confidence interval (CI) and p-values from logistic regression models (f: BHS ↦ CVD) are shown in Table 3. Due to the BHS from reference A having the largest OR (6.59, (95% CI: 5.91-7.33)), we decided to use it as the main BHS and we compared it to BS2 where relevant.

3.2.2 Polygenic risk score

The PRS was found to be marginally higher in cases than controls, with an average score of -0.058 in controls and -0.016 in cases. A simple logistic regression of CVD status on PRS was performed, giving an OR of 2.52 (95% CI: 2.31–2.74).

3.3 Models

3.3.1 Biomarkers univariate analysis

The p-values and effect size (OR) for the biomarkers can be visualised in Figure 3. Another perspective is presented in Figure C.3, with the biomarkers ranked by increasing p-value. The p-values and ORs for all biomarkers are available in the Appendix (Table C.3). As shown in Figure C.3, imputation does not greatly change the effect of any biomarker, suggesting that the observed associations are robust and not impacted by our processing method.

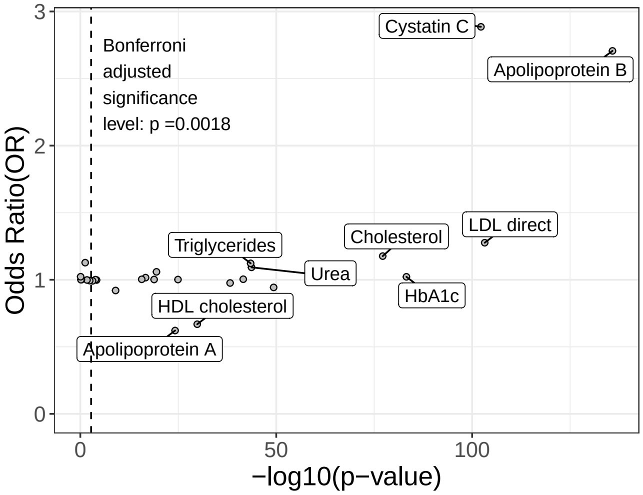


Figure 3: Odds ratio vs. -log10 of the p-value from univariate analyses of the association of each biomarker with CVD status.

3.3.2 Penalised regression

After obtaining the coefficients for 100 subsampling iterations of LASSO logistic regression, we calculated the selection proportion for each variable (proportion of iterations in which the variable was selected), as well as the mean and standard deviation of the effect size over all the iterations. These are shown in Figure 4 for biomarkers selected in at least 50% of iterations. The results for all biomarkers are presented in more detail in Table C.4. Additionally, the results for the models including covariates are shown in Figure 5.

3.3.3 Partial least squares

Calibration of the sPLS-DA model gave an optimum of 9 variables to be selected. 7 of the selected biomarkers (testosterone, urea, cystatin C, creatinine, urate, triglycerides, HbA1c) have positive loading coefficients (Figure 6a), indicating higher levels in CVD cases than controls. The remaining 2 biomarkers (apolipoprotein A and HDL cholesterol) have negative loading coefficients, indicating

CVD status	**0 (N=355096)**	**1 (N=11653)**	**Total (N=366749)**	**p-value**
Gender				<0.001
Female	191828 (54.0%)	3728 (32.0%)	195556 (53.3%)	
Male	163268 (46.0%)	7925 (68.0%)	171193 (46.7%)	
Age at recruitment				<0.001
Mean (SD)	55.522 (8.100)	59.996 (6.750)	55.664 (8.099)	
Range	38.000 - 73.000	40.000 - 71.000	38.000 - 73.000	
Age class				<0.001
<50	96749 (27.2%)	1104 (9.5%)	97853 (26.7%)	
50-64	203662 (57.4%)	7062 (60.6%)	210724 (57.5%)	
>64	54685 (15.4%)	3487 (29.9%)	58172 (15.9%)	
BHS (BS2 reference)				<0.001
Mean (SD)	0.224 (0.163)	0.261 (0.166)	0.225 (0.164)	
Range	0.000 - 1.000	0.000 - 0.900	0.000 - 1.000	
Quality of education				<0.001
low	49349 (13.9%)	2628 (22.6%)	51977 (14.2%)	
intermediate	179820 (50.6%)	5872 (50.4%)	185692 (50.6%)	
high	125927 (35.5%)	3153 (27.1%)	129080 (35.2%)	
Ever-smoke				<0.001
no	201783 (56.8%)	5408 (46.4%)	207191 (56.5%)	
yes	153313 (43.2%)	6245 (53.6%)	159558 (43.5%)	
Physical activity				<0.001
no	127288 (35.8%)	4773 (41.0%)	132061 (36.0%)	
yes	227808 (64.2%)	6880 (59.0%)	234688 (64.0%)	
Alcohol consumption				<0.001
Non-drinker	24838 (7.0%)	1021 (8.8%)	25859 (7.1%)	
Social drinker	77397 (21.8%)	2542 (21.8%)	79939 (21.8%)	
Moderate drinker	92600 (26.1%)	2959 (25.4%)	95559 (26.1%)	
Daily drinker	160261 (45.1%)	5131 (44.0%)	165392 (45.1%)	
BMI				<0.001
[18.5,25[	123665 (34.8%)	2677 (23.0%)	126342 (34.4%)	
[25,30[	151398 (42.6%)	5424 (46.5%)	156822 (42.8%)	
[30,40[	74011 (20.8%)	3308 (28.4%)	77319 (21.1%)	
>=40	6022 (1.7%)	244 (2.1%)	6266 (1.7%)	
Number of co-morbidities				<0.001
0	298600 (84.1%)	9985 (85.7%)	308585 (84.1%)	
1	56496 (15.9%)	1668 (14.3%)	58164 (15.9%)	
Number of medicines				<0.001
0	184417 (51.9%)	4414 (37.9%)	188831 (51.5%)	
1	57230 (16.1%)	1854 (15.9%)	59084 (16.1%)	
2	113449 (31.9%)	5385 (46.2%)	118834 (32.4%)	

Table 2: Basic statistics of our population by CVD case-control status (from the covariates dataset). The p-values in the right-hand column were calculated with a χ^2 test for categorical variables & with ANOVA for continuous variables.

lower levels in cases than controls. Stability analysis of this model, with 100 iterations, shows that this combination of variables is very stably selected (row 9 of Figure C.4 in the Appendix). Further stability analyses, covering all parameter values, shows that cystatin C is selected by every model (Figure C.4).

Across the five sPLS-DA subtypes models fitted, general consistency in biomarker selection and loading coefficients is observed (Figure 6b). Cystatin C was found to have the greatest loading coefficient in all subtypes except I209 (for which it has the second-largest coefficient). However, some differences are observed: apolipoprotein

Reference	OR(95%CI)	-log10(p-value)
A	6.59 (5.91, 7.33)	256
B	4.61 (4.14, 5.13)	170
C	2.48 (2.23, 2.72)	71
BS2	3.66 (3.29, 3.66)	127

Table 3: Effect size and p-values (-log10 scale) obtained from logistic regression models of each BHS (f: BHS ↦ CVD).

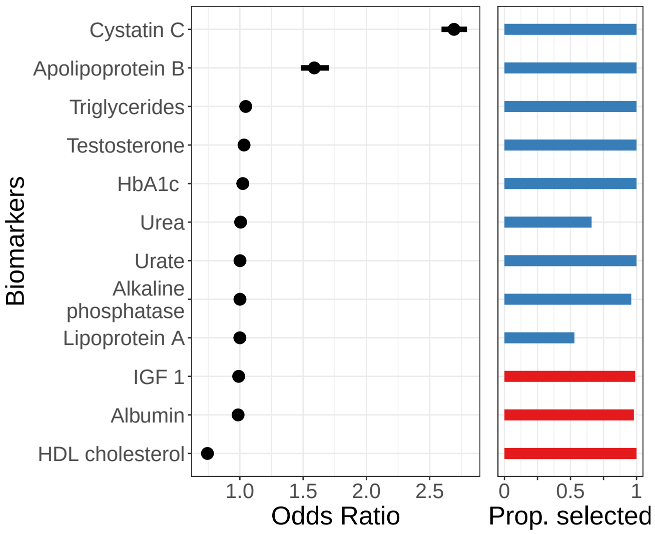


Figure 4: Odd ratios of CVD incidence for biomarkers selected more than 50% of the time over 100 perturbation cycles of LASSO regression. Blue bars represent a positive association and red bars a negative association.

A, HDL cholesterol and triglycerides were only selected by the models for I209, I219 and I251 subtypes.

3.3.4 Comparison of predictive model performance

Models were trained on different sets of data, as described. The AUC of predictions by these models on the test set were retrieved, shown in Figure 7.

4 Discussion

4.1 Biomarkers most associated with CVD case-control status

We employed three different statistical methods to explore the biomarker signature of CVD: univariate analyses, penalised multiple regression and partial least squares analysis. Univariate analyses and penalised multiple regression both identified cystatin C (OR: 2.69 (95% CI: 2.59-2.80) and apolipoprotein B (OR: 1.59 (95% CI: 1.48-1.70) as the topmost positively associated biomarkers with CVD; sPLS-DA also identified cystatin C as the most strongly associated biomarker. This comes as no surprise: the association of cystatin C with CVD has been researched in depth [18][19][20]. It is known for example that cystatin C directly participates in the development of atherosclerotic plaque [21]. Moreover, as early as 2009, researchers were advocating for the use of apoliprotein B as a better marker than LDL cholesterol for CVD patient management [22]. It is now known that LDL and other lipoprotein particles contain apolipoprotein B, and that apolipoprotein B is thereby a more direct way of measuring atherogenesis [23]. Our results support the before-mentioned research and also provide further support for these molecules being substantially better at identifying CVD cases than LDL cholesterol. Additionally, and in agreement with decades of literature, our results identify low levels of HDL cholesterol as a CVD risk factor (OR: 0.74 (95% CI: 0.72-0.77) in the LASSO model), suggesting a protective effect.

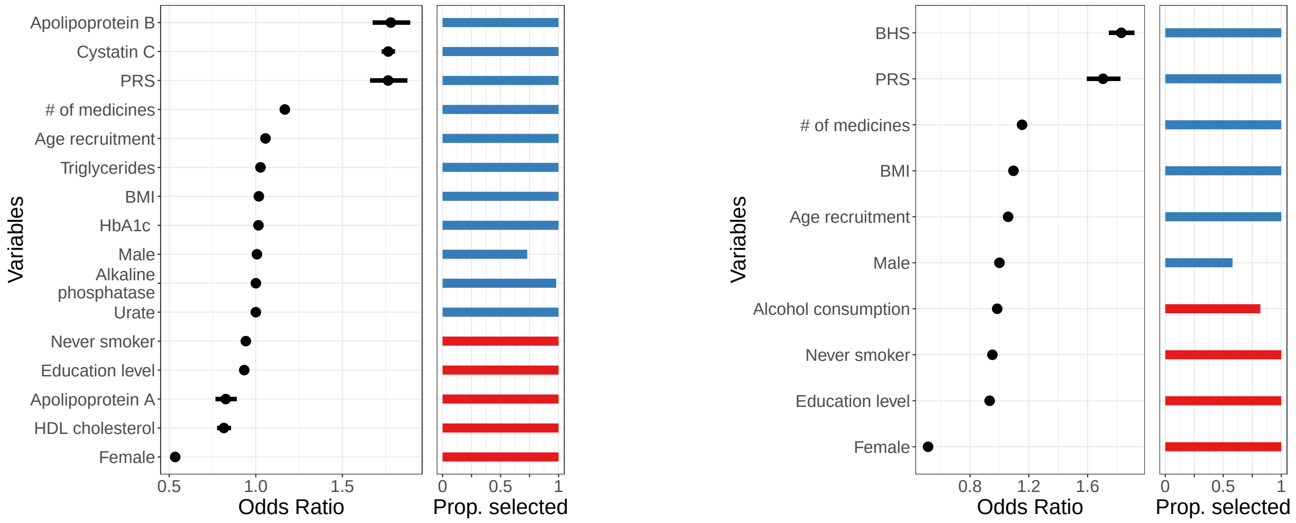
4.2 Identifying molecular signatures of different CVD subtypes

Our results (Figure 6b) show consistent and strong selection of cystatin C by all five subtype-specific models; this suggests that cystatin C is a universal CVD risk factor. Some differences in biomarker selection across subtypes can be observed: HDL cholesterol and apolipoprotein A (selected with negative loading coefficients), and triglycerides (selected with a positive loading coefficient) were only selected by models for ICD-10 subtypes I209, I219 and I251. These three CVD subtypes are all ischaemic heart diseases, while G459 and I639 subtypes are cerebrovascular diseases. These sub-classes of CVD may have different disease mechanisms: high triglyceride levels are recognised as a marker for other atherogenic lipoproteins [24], and HDL (with apolipoprotein A being the major component of HDL) has been suggested to protect against atherosclerosis [25] ("The HDL Hypothesis"). Atherosclerosis may therefore be a more prominent disease mechanism in the ischaemic heart diseases.

Importantly, these differing biomarker signatures could enable subtype-specific risk prediction, allowing more appropriate targeting of interventions. Future research should examine this possibility in broader categories of CVD, grouping by disease class rather than the highly specific ICD-10 code. This could provide more medically-applicable insights.

4.3 Factors most associated with CVD

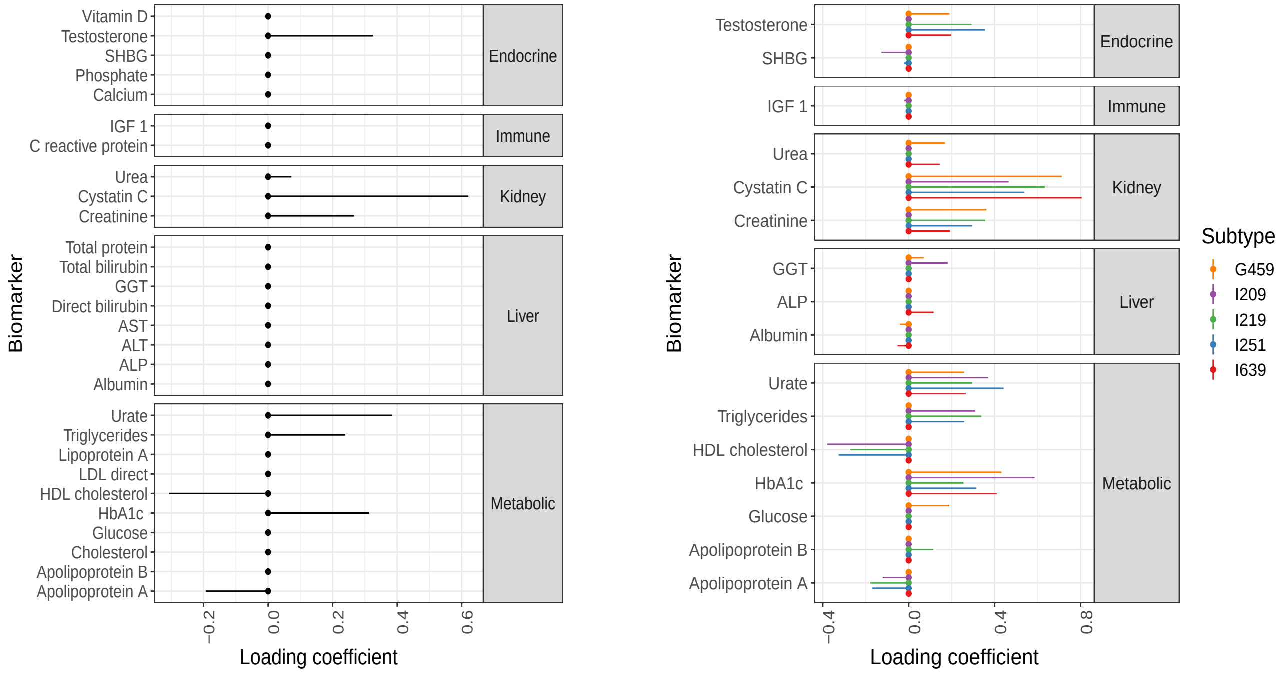
As mentioned previously, lifestyle factors and gender are strong determinants of CVD incidence. Our results support this, showing for example that being a never-smoker or having a greater education level confers lower odds of CVD (OR: 0.94 (95% CI: 0.93-0.96 and OR: 0.93 (95% CI: 0.92-0.95) respectively, from Figure 5a). An interesting finding from our study is that, even when adjusting



(a) Penalised regression with covariates, biomarkers and PRS

(b) Penalised regression with covariates, BHS and PRS

Figure 5: Penalised regression results for different sets of data. Odd ratios for each variable, with 95% CI, are shown. The selection proportion for each variable in the stability analyses is also shown; blue bars indicate a positive association, and red bars a negative association.



(a) sPLS-DA

(b) Stratified sPLS-DA. Only biomarkers selected by at least one model are shown.

Figure 6: Loading coefficients are shown for each biomarker (grouped into their respective body systems), for the sPLS-DA and stratified sPLS-DA models. A positive loading coefficient corresponds to a higher biomarker level in cases than controls.

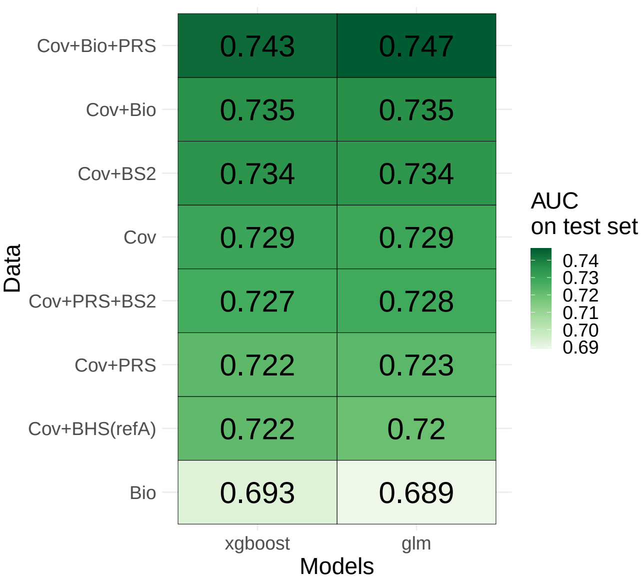


Figure 7: Results from the comparison of glm and XGBoost models with different combinations of data. *Cov stands for covariates, Bio for biomarkers.*

for biomarkers and lifestyle-related confounders, female sex remains the greatest protective factor for CVD, with 46.5% decreased odds of having CVD compared to males (OR: 0.54 (95% CI: 0.519-0.551). This finding is supported by the literature: oestrogen plays an important role in lipid metabolism, and it has been shown that women with oestrogen deficiency have a seven-fold increased risk of developing CVD [26]. Nonetheless, CVD remains the number-one killer among women [27], and much effort is and must be put into eradicating the view of CVD as a “men’s disease”.

4.4 “Black box” approaches in understanding disease

With the rise of machine learning, more researchers are using these methods to study high-dimensional biological data, both for personalised medicine and a deeper understanding of health [28][29]. These methods can be used to classify patients as cases or controls based on patterns drawn from their available data. Our results (Figure 7) show that the combination of covariates, the full biomarker dataset (as opposed to the BHS) and the PRS offers the highest predictive power, with an AUC of 0.75 for the logistic LASSO model. It is worth noting that there is only a 5% range in AUC between the “worst performing” model (using only the biomarkers dataset as explanatory variables) and the best performing model. According to [30], the range of AUC values we have obtained (0.69-0.75) is “reasonable” and “strong” in vascular research. In our case, the use of a more sophisticated model (XGBoost) did not improve the results obtained from a simpler more traditional LASSO logistic regression. Surprisingly, using the BHS with reference A yielded a lower predictive performance than using BS2 in the model comparison section (Figure 7), despite the former having a larger effect size than the latter (Table 3).

In some instances, adding more variables did increase predictive performance: the models trained on just the covariates outperformed other models that were trained on covariates and additional features (PRS and BHS). This implies that there is interaction between the many features considered in the models, and therefore the addition of certain variables does not contribute in a linear and modular way to the predictive performance of that model.

4.5 Are aggregate risk scores valuable?

Risk scores can be used as diagnostic and screening tools, and often play a role in treatment allocation [31]. The exploration of several models and combinations of data does not allow us to confirm whether PRS and BHS improve predictive performance, as previously discussed in Section 4.4. However, when looking at the results from penalised regression models (Figure 5) it becomes evident that the BHS and PRS are important in predicting CVD case-control status. When the BHS is considered along with other confounders, its effect size is estimated at an OR of 1.83 (95% CI: 1.75- 1.92). This is somewhat misleading: this result indicates that a person with a BHS of 1 (maximum value) has an 83% increased odds compared to someone with a BHS of 0 (minimum). Nevertheless, this a meaningful result and implies that the BHS may be a good measure to summarise the complexity of many biomarkers. The PRS effect size is also large, with an OR of 1.71 (95% CI: 1.59-1.82) when adjusting for covariate confounders and BHS, and an OR of 1.76 (95% CI: 1.73-1.81) when adjusting for covariates and biomarkers, but again this result may be misleading due to the narrow range of possible values (-1 to 1). Furthermore, even though the PRS and BHS were found to be relevant in multiple regression settings, their predictive performance when considered as single explanatory variables is poor, with an AUC of 0.5 (results not shown). This shows that the PRS and BHS are useful at aggregating higher dimensional data with minimal damage to predictive performance, but they fail to be powerful predictive scores alone.

4.6 Limitations

First, as this is a cross-sectional study, there is no follow-up assessment or repeated measurements. This introduces a bias, as the phenotype of individuals diagnosed with CVD has been affected by changes in their medication and their behaviour since the time of diagnosis. Time-to-event studies can provide risk scores of a patient developing a disease over a period of time (see [31]); this would be an interesting route to explore. Additionally, we failed to report on the range of the biomarkers which makes the interpretation of effect sizes confusing. In future research, steps will be taken to standardise numerical measurements whenever it is feasible.

Second, the research was restricted to participants who were primarily of European descent. Thus, studies into ethnically different populations are needed to improve the generalisability of our results. The PRS would have been biased by this pitfall: due to demographic activities (such as migrations) which may contribute to prejudice in detection (detection bias), disease-alleles may have substantially different concentrations across populations. Furthermore, selection of SNPs based on results from a single GWAS could have hindered results, especially since the GWAS was examining genetic associations with coronary artery disease which, although the most prevalent type, is just one type of CVD. Possibly, examining numerous GWAS studies and choosing recurrent SNPs would have increased the validity of the PRS.

Third, the creation of the BHS score was reliant on dichotomising each biomarker dependent on quartiles, as opposed to implementing an 'at-risk' cut-off point for each individual biomarker. This could be an issue for two reasons. Firstly, many biomarkers can be harmful both in excess and in shortage; such as iron (when low can cause anaemia, when high can be carcinogenic). Secondly, the arbitrary cut-off point method may lead to misclassification of risk status: for example, our study population were predominantly overweight, so people with high cholesterol levels could have been regarded as not at risk when comparing to the population. This makes the BHS unstably dependent on the population structure instead of stably dependent on biological phenotype and disease outcomes. Therefore, a more robust approach which takes into consideration these constraints is needed.

The usage of the ROC AUC in medical studies has previously been criticised: the ROC curve is independent from prevalence [32], and when class distribution is skewed it provides over-optimistic results [33]. For classification problems with high imbalance and particular interest in the minority class, as in our case, the precision-recall curve (PRC) is preferred [33]. Future work on CVD prevalence classification should use the PRC and the PRC AUC to evaluate model performance.

Finally, the incorporation of some other blood-based biomarkers that have been reported to be associated with CVD risk, such as natriuretic peptides (especially B-type and N-terminal prohormone BNP), fibrinogen and troponin [34], were not available in UKBB. Inclusion of these additional biomarkers could boost the predictive power of all models tested in this study.

5 Conclusion

We found consistent identification of cystatin C as the biomarker most strongly correlated with CVD status, across all statistical models used. A LASSO logistic regression model using variables from all data sources (covariates, biomarkers, and PRS) yielded the best predictive performance. However, the range in AUC between all combinations was minimal, indicating, perhaps, the need to incorporate other model types.

References

[1] Mark Rishniw. *Cardiovascular Diseases.* 2017. DOI: 10.1016/B978-1-4377-0660-4.00020-X. URL: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] British Heart Foundation. "UK Factsheet". In: *British Heart Foundation* January (2020), pp. 1–21. ISSN: 16742001.

[3] *Health matters: preventing cardiovascular disease - GOV.UK.* URL: https://www.gov.uk/government/publications/health-matters-preventing-cardiovascular-disease/health-matters-preventing-cardiovascular-disease.

[4] Edward Yu et al. *Diet, lifestyle, biomarkers, genetic factors, and risk of cardiovascular disease in the nurses' health studies.* Sept. 2016. DOI: 10.2105/AJPH.2016.303316.

[5] *The NHS Long Term Plan.* Tech. rep. 2019. URL: www.longtermplan.nhs.uk.

[6] Maria G Barderas et al. "Metabolomic Profiling for Identification of Novel Potential Biomarkers in Cardiovascular Diseases". In: *Journal of Biomedicine and Biotechnology* 2011 (2011). DOI: 10.1155/2011/790132.

[7] Sonia Dagnino and Anthony Macherone. *Unraveling the exposome.* 2019. URL: https://www.springer.com/gp/book/9783319893204.

[8] Stephen M. Rappaport and Martyn T. Smith. *Environment and disease risks.* Oct. 2010. DOI: 10.1126/science.1192603.

[9] Cathie Sudlow et al. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age". In: *PLoS Medicine* 12.3 (Mar. 2015), e1001779. ISSN: 15491676. DOI: 10.1371/journal.pmed.1001779. URL: https://dx.plos.org/10.1371/journal.pmed.1001779.

[10] Majid Nikpay et al. "A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease". In: *Nature Genetics* 47.10 (Sept. 2015), pp. 1121–1130. ISSN: 15461718. DOI: 10.1038/ng.3396.

[11] Clare Bycroft et al. "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562.7726 (Oct. 2018), pp. 203–209. ISSN: 14764687. DOI: 10.1038/s41586-018-0579-z. URL: http://www.ncbi.nlm.nih.gov/pubmed/30305743%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6786975.

[12] Trevor Hastie et al. *Imputing Missing Data for Gene Expression Arrays*. Tech. rep. 1999.

[13] Maryam Karimi et al. "Early-life inequalities and biological ageing: a multisystem Biological Health Score approach in Understanding Society". In: *J Epidemiol Community Health* 73 (2019), pp. 693–702. DOI: 10.1136/jech-2018-212010. URL: http://dx.doi.org/10.1136/jech-2018-212010.

[14] Maryam Karimi et al. "Early-life inequalities and biological ageing: A multisystem Biological Health Score approach in U nderstanding S ociety". In: *Journal of Epidemiology and Community Health* 73.8 (Aug. 2019), pp. 693–702. ISSN: 14702738. DOI: 10.1136/jech-2018-212010.

[15] Trevor Hastie, Jerome Friedman, and Robert Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. ISSN: 0014-4886. DOI: 10.1016/j.expneurol.2008.01.011.

[16] Florian Rohart et al. "mixOmics: An R package for 'omics feature selection and multiple data integration". In: *PLOS Computational Biology* 13.11 (Nov. 2017). Ed. by Dina Schneidman, e1005752. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005752. URL: https://dx.plos.org/10.1371/journal.pcbi.1005752.

[17] Roel Vermeulen et al. "Pre-diagnostic blood immune markers, incidence and progression of B-cell lymphoma and multiple myeloma: Univariate and functionally informed multivariate analyses". In: *International Journal of Cancer* 143.6 (Sept. 2018), pp. 1335–1347. ISSN: 10970215. DOI: 10.1002/ijc.31536.

[18] Magdalena Madero and Mark J Sarnak. "Association of cystatin C with adverse outcomes." In: *Current opinion in nephrology and hypertension* 18.3 (May 2009), pp. 258–63. ISSN: 1473-6543. DOI: 10.1097/mnh.0b013e328326f3dd. URL: http://www.ncbi.nlm.nih.gov/pubmed/19374014%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2890263.

[19] Christos Angelidis et al. "Cystatin C: An Emerging Biomarker in Cardiovascular Disease". In: ().

[20] Paul Muntner et al. "Serum Cystatin C and Increased Coronary Heart Disease Prevalence in US Adults Without Chronic Kidney Disease". In: *American Journal of Cardiology* 102.1 (July 2008), pp. 54–57. ISSN: 00029149. DOI: 10.1016/j.amjcard.2008.02.098.

[21] Adeera Levin and James H. Lan. *Cystatin C and Cardiovascular Disease: Causality, Association, and Clinical Implications of Knowing the Difference*. Aug. 2016. DOI: 10.1016/j.jacc.2016.06.037.

[22] John H Contois et al. "Apolipoprotein B and Cardiovascular Disease Risk: Position Statement from the AACC Lipoproteins and Vascular Diseases Division Working Group on Best Practices". In: *Clinical Chemistry* 55.3 (Mar. 2009), pp. 407–419. ISSN: 0009-9147. DOI: 10.1373/clinchem.2008.118356. URL: https://academic.oup.com/clinchem/article/55/3/407/5629355.

[23] Allan D. Sniderman et al. *Apolipoprotein B Particles and Cardiovascular Disease: A Narrative Review*. Dec. 2019. DOI: 10.1001/jamacardio.2019.3780. URL: http://www.ncbi.nlm.nih.gov/pubmed/31642874.

[24] Beatriz G. Talayero and Frank M. Sacks. "The role of triglycerides in atherosclerosis". In: *Current Cardiology Reports* 13.6 (Dec. 2011), pp. 544–552. ISSN: 15233782. DOI: 10.1007/s11886-011-0220-3.

[25] Menno Vergeer et al. *The HDL hypothesis: Does high-density lipoprotein protect from atherosclerosis?* Aug. 2010. DOI: 10.1194/jlr.R001610.

[26] A. H.E.M. Maas and Y. E.A. Appelman. *Gender differences in coronary heart disease*. 2010. DOI: 10.1007/s12471-010-0841-y.

[27] Mark Woodward. *Cardiovascular disease and the female disadvantage*. Apr. 2019. DOI: 10.3390/ijerph16071165.

[28] Jina Ko et al. *Machine learning to detect signatures of disease in liquid biopsies-a user's guide*. Feb. 2018. DOI: 10.1039/c7lc00955k.

[29] Jordi Martorell-Marugán et al. "Deep Learning in Omics Data Analysis and Precision Medicine". In: *Computational Biology*. Codon Publications, Nov. 2019, pp. 37–53. DOI: 10.15586/computationalbiology.2019.ch3.

[30] Celine Foote, Mark Woodward, and Meg J. Jardine. *Scoring Risk Scores: Considerations Before Incorporating Clinical Risk Prediction Tools Into Your Practice*. May 2017. DOI: 10.1053/j.ajkd.2017.02.005.

[31] Ralph B. D'Agostino et al. *Cardiovascular disease risk assessment: Insights from Framingham*. 2013. DOI: 10.1016/j.gheart.2013.01.001.

[32] Steve Halligan, Douglas G. Altman, and Susan Mallett. "Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach". In: *European Radiology* 25.4 (Mar. 2015), pp. 932–939. ISSN: 14321084. DOI: 10.1007/s00330-014-3487-0.

[33] Jesse Davis and Mark Goadrich. *The Relationship Between Precision-Recall and ROC Curves*. Tech. rep.

[34] Ying Huang et al. *Biomarkers of Cardiovascular Disease*. 2017. DOI: 10.1155/2017/8208609.

[35] Runmin Wei et al. "Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data". In: *Scientific Reports* 8.1 (Dec. 2018), pp. 1–10. ISSN: 20452322. DOI: 10.1038/s41598-017-19120-0.

[36] *statistical significance - What sense does it make to compare p-values to each other? - Cross Validated*. URL: https://stats.stackexchange.com/questions/21419/what-sense-does-it-make-to-compare-p-values-to-each-other.

A Response to reviewers

Reviewers expressed concern regarding several features of an early state of our project:

1. A not clear enough definition of what we considered as patient with cardiovascular disease and our definition of cases.
2. Concern about the processing of the SNPs where we removed SNPs with a degree of similarity above 90%.
3. Rationale for removing individuals with 50% or more of measurements missing
4. Robustness of imputation and benefit of using multiple imputation
5. Rationale and added-value behind employing LASSO and PLS. Better definition of intent of using both
6. Validity of p-value comparison

Our responses to these issues are listed below:

1. We define our cases as those individuals with prevalent CVD at time of recruitment in the UK Biobank. These are the individuals for which the variable `CVD_status` in the UKBB covariates dataset is 1. (controls have `CVD_status` = 0).
2. This measure was not intentionally taken and was an artefact in our processing. Rare SNPs can be specially insightful.
3. Imputation for missing biomarkers for each individual was based on values of measured biomarkers. A great amount of missing measurements could result in very erroneous imputation estimates, thus we set the threshold for missingness at 50%. Regarding the bias this introduces we believe that due to large sample size of UK Biobank bias due to removing 33,000 out of 502,000 individuals does not generate too much bias, although associations between missingness and other covariates was not statistically evaluated.
4. We based our method of imputation on the paper by Hastie & Tibshirani [12]. In their paper they show how kNN imputation yields the best result when compare to other imputation methods. We tested our imputation error in two ways. First, we evaluated the results for one of our analysis and evaluated whether or not they greatly differed. We then went onto getting estimation errors by artificially injecting missingness in random subsets of sets of complete of data (see C.7). Based on the results of the first experiment (see C.3) we decided to progress with that method of imputation as the results did not differ greatly between the imputed data and the non-imputed data (no statistical test was performed - though though the effect size did not change greatly). The results from C.7, were not promising as some biomarkers had normalised root mean square error (NRMSE) around 1 (i.e. 1 standard deviation). Initially we thought this was due to coding errors perhaps but further review of the literature showed that kNN imputation in metabolic data can indeed yield this kind of errors ([35]). We also implemented Multivariate Imputation by Chained Equations (MICE) but did not run any experiment with that data for reasons not stated. In future iterations of our research we will not employ kNN imputation and lean towards exploring MICE or random forest imputation as suggested in [35].
5. In a previous iteration of our research project, the boundary between the usefulness of sPLS and that of LASSO was less clear. Both models perform feature selection and differences arising between them could be indeed an artefact of the model instead of the data. Despite this issue, coherent results across modelling methods ensures robustness. Nevertheless, we hope to have made clear now that we used LASSO as our main tool to assess effect sizes for each variable studied and PLS mainly for the CVD sub-type analysis. One could argue using PLS for such endeavour is not necessary, though we thought worth exploring to get a different perspective on the data.
6. Previously, it appeared from presenting Figure C.3 that we were focusing on the magnitude of the -$\log_{10}$(p-value) to rank and compare the effect of biomarkers. Of course p-values do not inform of effect size and we therefore decided to change our main plot for the univariate analysis, showing both the p-value and effect size (OR) (see Figure 3). Nonetheless, it is arguable that given a fixed sample size N, p-values are comparable as they are monotonically related to the t-value which in turn is related to the effect size [36].

B Detailed methods

B.1 Biological health score(BHS)

In short, the BHS calculation method considers biomarkers to be either harmful in excess or harmful in shortage. For each biomarker an 'at-risk' quartile is defined: if the biomarker is considered harmful in excess, the 'at-risk' quartile is the 3rd quartile, otherwise it is the 1st quartile. More over, the quartiles are calculated per strata (by age group and gender). For each biomarker, a patient scores 1 if the patient's measurement belongs to the at-risk quartile for the patient's strata. Finally, the BHS is calculated as the average score across all biomarkers. A higher BHS is considered harmful. The BHS can be calculated:

$$BHS^p = \frac{1}{B}\sum_{b=1}^{B} I_b^p$$

Where p represents a patient, b is a biomarker, B the total number of biomarkers assessed and I_b^p the binary score for a biomarker b, for the patient p.

System-specific sub-scores can also be calculated, where each biomarker is considered to belong to non-overlapping subsystems (i.e. one biomarker cannot belong to multiple systems).

These can be calculated as follows:

$$BHS_s^p = \frac{1}{B_s}\sum_{b=1}^{B_s} I_b^p$$

Where s represents a given system (e.g. metabolic, immune...), B_s is the number of biomarkers in system s and BHS_s^p, represents the BHS score for patient p and the system s.

C Extra figures and tables

C.1 BHS plot

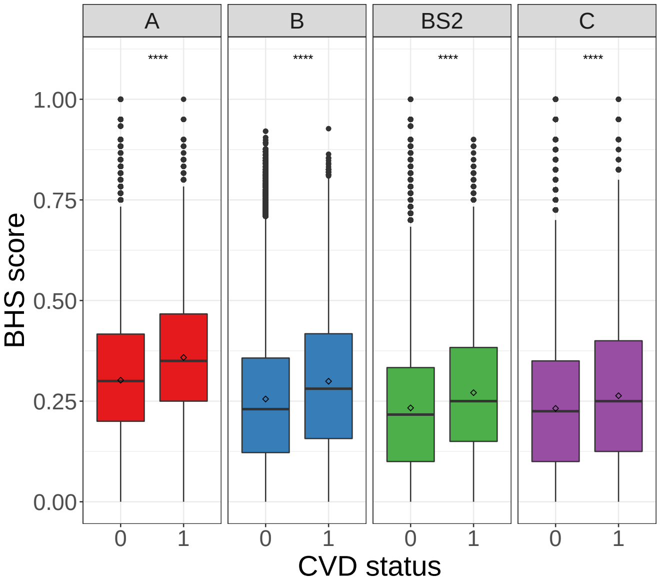


Figure C.1: BHS scores by CVD status for the four different references that were considered in this study. *The diamonds represent the mean BHS value for each BHS reference, by CVD case-control status.*

C.2 Available variables in the covariates dataset

Variable	Description
vit_status	1 if the participant is dead, 0 otherwise
dc_cancer_st	1 if death from cancer, 0 otherwise
dc_cvd_st	1 if death from cardiovascular disease (CVD), 0 otherwise
dc_external_st	1 if death from external causes, 0 otherwise
age_recruitment.0.0	Age at recruitment
age_cl	Age class with 3 levels: <50, 50-65, >64
stop	Age at death or end of follow up
stop_cancer	Age at death from cancer or end of follow up
stop_cvd	Age at death from CVD or end of follow up
stop_external	Age at death from external cause or end of follow up
cancer_status	Participant diagnosed with cancer
CVD_status	Participant diagnosed with CVD
age_cancer	Age at cancer diagnosis
age_CVD	Age at CVD diagnosis
BS2_all	Biological Health Score at baseline
qual2	Education level (possible values: low, intermediate, high)
smok_ever_2	Ever smoker status (yes: has ever smoked; no: has never smoked)
physical_activity_2	Physical activity (yes: if vigorous physical activity for more than 10 minutes at least once a week; no: otherwise)
alcohol_2	Alcohol consumption, (possible values: non-drinker, social drinker, moderate drinker, daily drinker)
BMI_5cl_2	BMI category ([18.5,25[, [25,30[, [30,40[, >40)
no_cmrbt_cl2	Number of co-morbidities (range: 1 to 3)
no_medicines	Number of pharmacological treatments (range: 1 to 3)
cvd_final_icd10	ICD code 10 of CVD diagnosis (linkage to Hospital Episode Statistics)
cvd_final_icd9	ICD code 9 for CVD diagnosis (linkage to Hospital Episode Statistics)
cvd_final_nc_illness_code	Non cancer illness code of CVD diagnosis (self-reported in UKBB nurse-administered questionnaire)
cvd_final_opcs4	Operation code linked to CVD (linkage to Hospital Episode Statistics)
cvd_final_ukb_oper_code	Operation code linked to CVD (UKBB questionnaire)
primary_cause_death_ICD10	ICD 10 of the primary cause of death
cancer_death	1 if death from cancer
cvd_death	1 if death from cardiovascular disease (CVD)
external_cause_death	1 if death from external causes
other_cause_death	1 if death from other cause (not CVD, cancer or external)
cancer_final_icd10	ICD 10 of cancer diagnosis (linkage to cancer registry)
cancer_incident	1 if incident cancer case
cancer_prevalent	1 if prevalent cancer case (i.e. diagnosed before recruitment)
cvd_incident	1 if incident CVD case
cvd_prevalent	1 if prevalent CVD case
Gender	Male or Female

Table C.1: Available variables and their descriptions in the covariates dataset. Variables not in bold were dropped at all stages of our analysis. Variables in light blue were used as confounders when mentioned in the main text. CVD_status (in orange) was our main response variable.

C.3 BHS references tables

Biomarker name	**System A**	*****	**System B**	*****	**System C**	*****
Alanine aminotransferase	Liver	1	Liver	1	Metabolic	1
Albumin	NA	NA	Liver	0	NA	NA
Alkaline phosphatase	NA	NA	Liver	1	NA	NA
Apolipoprotein A	NA	NA	Metabolic	1	NA	NA
Apolipoprotein B	NA	NA	Metabolic	1	NA	NA
Aspartate aminotransferase	Liver	1	Liver	1	Liver	1
C-reactive protein	Inflammatory /Immune	1	Inflammatory/Immune	1	Inflammatory/Immune	1
Calcium	NA	NA	Endocrine	-1	NA	NA
Cholesterol	Metabolic	1	Metabolic	1	Metabolic	NA
Creatinine	Kidney	1	Kidney	-1	Kidney	1
Cystatin C	NA	NA	Kidney	1	NA	NA
Direct bilirubin	NA	NA	Liver	1	NA	NA
Gamma glutamyltransferase	Liver	1	Liver	1	Liver	1
Glucose	NA	NA	Metabolic	-1	NA	NA
Glycated haemoglobin (HbA1c)	Metabolic	1	Metabolic	1	Metabolic	1
HDL cholesterol	Metabolic	0	Metabolic	0	Metabolic	0
IGF-1	Inflammatory/Immune	0	Inflammatory/Immune	-1	Inflammatory/Immune	0
LDL direct	NA	NA	Metabolic	1	Metabolic	1
Lipoprotein A	NA	NA	Metabolic	1	NA	NA
Oestradiol	NA	NA	Endocrine	-1	NA	NA
Phosphate	NA	NA	Endocrine	-1	NA	NA
Rheumatoid factor	NA	NA	Inflammatory/Immune	1	NA	NA
SHBG	NA	NA	Endocrine	-1	NA	NA
Testosterone	Endocrine	0	Endocrine	-1	NA	NA
Total bilirubin	NA	NA	Liver	1	NA	NA
Total protein	NA	NA	Liver	-1	NA	NA
Triglycerides	Metabolic	1	Metabolic	1	Metabolic	1
Urate	NA	NA	Metabolic	1	NA	NA
Urea	NA	NA	Kidney	1	NA	NA
Vitamin D	NA	NA	Endocrine	0	NA	NA

Table C.2: [* stands for "Harmful in excess?"] Reference values for the calculation of BHS. In the column "Harmful in excess" a 1 means True (i.e. that biomarker is considered is harmful in excess) whereas a 0 means False. -1 means does not apply and NA means not considered.

C.4 Exploratory data analysis

C.4.1 PCA on biomarkers

PCA was also performed on the biomarkers dataset, to assess the quantity of correlation in the data (figure C.2). The first two principal components explained 26.61% of the variance in the data. However, the PCA failed to separate cases and controls, with no clusters forming. Lipid metabolism-related biomarkers were found to contribute heavily to the second principal component.

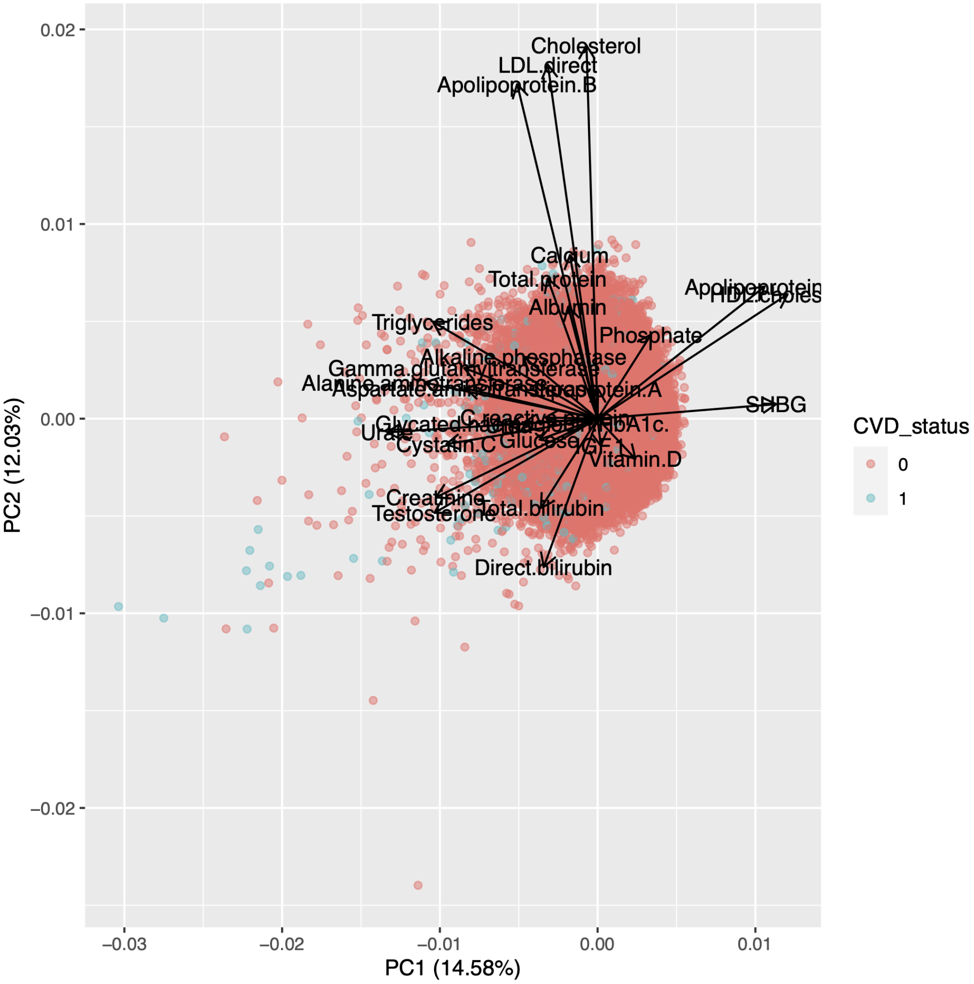


Figure C.2: Projection of the samples and original features on the top 2 principal components (as per explained variance).

C.5 Univariate analysis extra-results

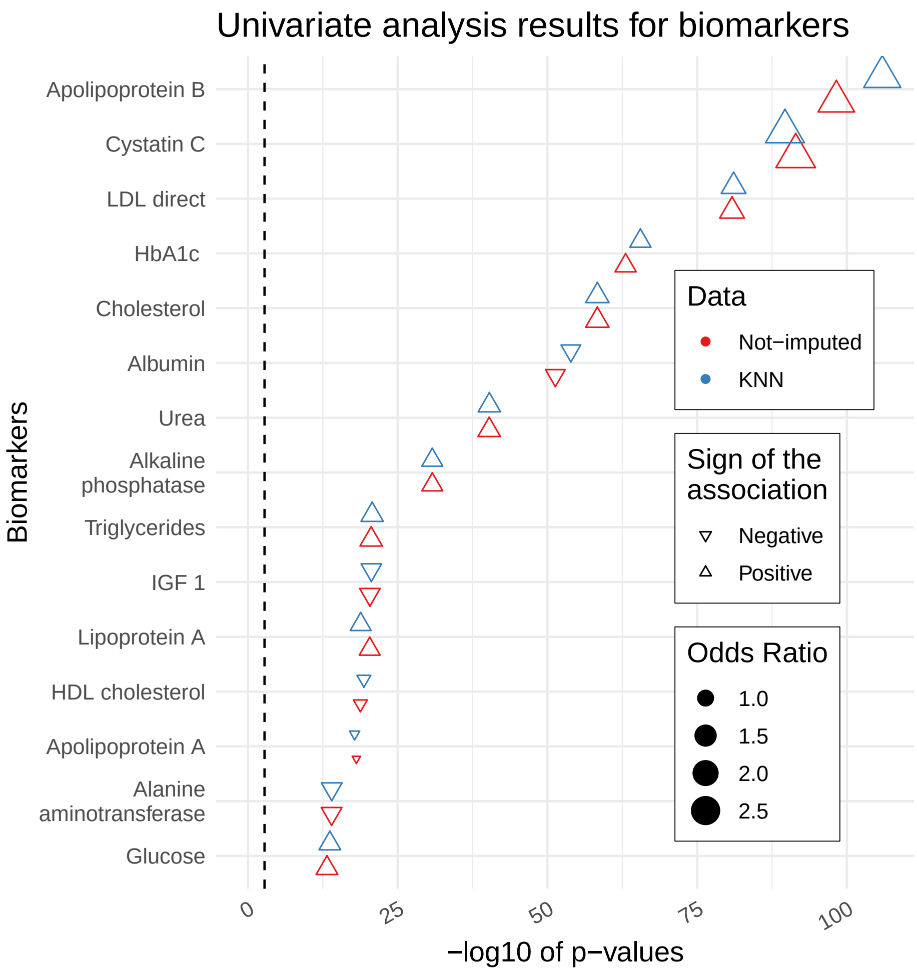


Figure C.3: -log10 of the p-values for the top 10 most associated biomarkers to CVD status. The dashed line corresponds to the Bonferroni adjusted p-value.

Biomarkers	Data	p-value	Lower 95%CI	OR	Upper 95%CI
Apolipoprotein B	KNN	1.50E-136	2.5018	2.7062	2.9269
Apolipoprotein B	Not-imputed	1.05E-127	2.4367	2.6373	2.8541
Cystatin C	Not-imputed	5.09E-106	2.7241	3.0049	3.3177
LDL direct	KNN	6.13E-104	1.2481	1.276	1.3045
LDL direct	Not-imputed	1.01E-103	1.248	1.2759	1.3044
Cystatin C	KNN	5.69E-103	2.6223	2.8856	3.1796
HbA1c	KNN	5.92E-84	1.0196	1.0219	1.0241
HbA1c	Not-imputed	1.90E-80	1.0193	1.0216	1.0238
Cholesterol	KNN	6.67E-78	1.1565	1.1763	1.1965
Cholesterol	Not-imputed	6.87E-78	1.1565	1.1764	1.1966
Albumin	KNN	4.62E-50	0.9357	0.9429	0.9503
Albumin	Not-imputed	2.60E-47	0.9367	0.9441	0.9515
Urea	KNN	2.03E-44	1.079	1.0925	1.1061
Urea	Not-imputed	2.20E-44	1.079	1.0925	1.1061
Triglycerides	KNN	3.44E-44	1.1043	1.1224	1.1407
Triglycerides	Not-imputed	5.33E-44	1.104	1.1222	1.1405
Alkaline phosphatase	KNN	2.64E-42	1.0034	1.004	1.0046
Alkaline phosphatase	Not-imputed	2.72E-42	1.0034	1.004	1.0046
IGF 1	KNN	5.91E-39	0.9722	0.9758	0.9794
IGF 1	Not-imputed	1.80E-38	0.9724	0.976	0.9795
HDL cholesterol		1.46E-30	0.624	0.6685	0.7159
HDL cholesterol	Not-imputed	2.50E-29	0.6247	0.67	0.7183
Urate	KNN	1.21E-25	1.0012	1.0015	1.0018
Urate	Not-imputed	1.32E-25	1.0012	1.0015	1.0018
Apolipoprotein A	Not-imputed	4.67E-25	0.5598	0.6141	0.6735
Apolipoprotein A	KNN	6.52E-25	0.5675	0.6214	0.6801
Lipoprotein A	Not-imputed	6.99E-21	1.0016	1.002	1.0024
Glucose	KNN	3.68E-20	1.046	1.059	1.0719
Lipoprotein A	KNN	1.48E-19	1.0015	1.0019	1.0023
Glucose	Not-imputed	1.73E-19	1.0453	1.0584	1.0713
C reactive protein	KNN	2.23E-17	1.0121	1.0158	1.0195
C reactive protein	Not-imputed	2.26E-17	1.0121	1.0158	1.0195
Creatinine	Not-imputed	1.78E-16	1.0025	1.0032	1.004
Creatinine	KNN	2.13E-16	1.0024	1.0032	1.004
Direct bilirubin	KNN	9.99E-10	0.8954	0.9199	0.9446
Direct bilirubin	Not-imputed	1.86E-07	0.9054	0.9305	0.9558
Vitamin D	KNN	7.02E-05	0.9971	0.998	0.999
Gamma glutamyltransferase	KNN	1.48E-04	1.0004	1.0007	1.0011
Vitamin D	Not-imputed	1.51E-04	0.9971	0.9981	0.9991
Gamma glutamyltransferase	Not-imputed	2.06E-04	1.0003	1.0007	1.0011
Total protein	Not-imputed	3.91E-04	0.9862	0.9911	0.996
Total protein	KNN	7.53E-04	0.9867	0.9916	0.9965
Total bilirubin	KNN	0.004046429	0.9889	0.9934	0.9979
Total bilirubin	Not-imputed	0.005043518	0.9891	0.9936	0.998
Alanine aminotransferase	KNN	0.019911685	0.9969	0.9983	0.9997
Alanine aminotransferase	Not-imputed	0.021839892	0.9969	0.9984	0.9997
Phosphate	KNN	0.057902112	0.996	1.1275	1.2763
Phosphate	Not-imputed	0.07883033	0.9871	1.1186	1.2676
Testosterone	KNN	0.518967775	0.9959	1.002	1.0081
SHBG	Not-imputed	0.554196353	0.9993	1.0003	1.0012
SHBG	KNN	0.612903188	0.9993	1.0002	1.0012
Aspartate aminotransferase	Not-imputed	0.67396292	0.9986	1.0004	1.0021
Aspartate aminotransferase	KNN	0.685509075	0.9986	1.0004	1.002
Testosterone	Not-imputed	0.823721778	0.9928	0.9993	1.0058
Calcium	KNN	0.838670067	0.8249	1.0225	1.267
Calcium	Not-imputed	0.989143314	0.8044	0.9985	1.2388

Table C.3: p-values, odd-ratios and 95% confidence intervals for the imputed and non-imputed data for the 28 biomarkers assesed.

C.6 Penalised regression values

Biomarker	Prop. Selec	OR	OR-SD	OR+SD
Cystatin C	1.00	2.692	2.593	2.795
Apolipoprotein B	1.00	1.588	1.481	1.703
Triglycerides	1.00	1.046	1.040	1.053
Testosterone	1.00	1.033	1.030	1.036
HbA1c	1.00	1.023	1.022	1.024
Urea	0.66	1.006	0.999	1.013
Urate	1.00	1.001	1.001	1.001
Alkaline phosphatase	0.96	1.001	1.001	1.001
Lipoprotein A	0.53	1.000	1.000	1.000
C reactive protein	0.08	1.000	1.000	1.000
Creatinine	0.04	1.000	1.000	1.000
Gamma glutamyltransferase	0.01	1.000	1.000	1.000
Alanine aminotransferase	0.00	1.000	1.000	1.000
Apolipoprotein A	0.00	1.000	1.000	1.000
Aspartate aminotransferase	0.00	1.000	1.000	1.000
Direct bilirubin	0.00	1.000	1.000	1.000
Calcium	0.00	1.000	1.000	1.000
Cholesterol	0.00	1.000	1.000	1.000
Glucose	0.00	1.000	1.000	1.000
LDL direct	0.00	1.000	1.000	1.000
Phosphate	0.00	1.000	1.000	1.000
SHBG	0.00	1.000	1.000	1.000
Total bilirubin	0.00	1.000	1.000	1.000
Total protein	0.00	1.000	1.000	1.000
Vitamin D	0.00	1.000	1.000	1.000
IGF 1	0.99	0.991	0.988	0.994
Albumin	0.98	0.986	0.980	0.992
HDL cholesterol	1.00	0.744	0.720	0.768

Table C.4: Aggregate values of the proportion selection(Prop. Selec), odds ratio (OR) as well as the OR one standard deviation (SD) away (either side) for the penalised regression methods on biomarkers only.

Variable	Proportion selected	OR-SD	OR	OR+SD
BHS	1.00	1.745	1.830	1.919
PRS	1.00	1.594	1.705	1.824
# of medicines	1.00	1.143	1.154	1.165
BMI	1.00	1.083	1.095	1.108
Age at recruitment	1.00	1.057	1.059	1.061
Male	0.58	1.000	1.000	1.000
Has ever smoked	0.41	1.000	1.000	1.000
No physical activity	0.05	0.999	1.000	1.002
Physical activity	0.03	1.000	1.000	1.000
# of co-morbidites	0.00	1.000	1.000	1.000
Alcohol consumption	0.82	0.972	0.985	0.998
Never smoked	1.00	0.931	0.952	0.974
Education level	1.00	0.923	0.933	0.944
Female	1.00	0.499	0.514	0.530

Table C.5: Aggregate values of the proportion selection(Prop. Selec), odds ratio (OR) as well as the OR one standard deviation (SD) away (either side) for the penalised regression methods with covariates, BHS and PRS (no Biomarkers).

Variable	Proportion selection	OR-SD	OR	OR+SD
Apolipoprotein B	1.00	1.675	1.780	1.892
Cystatin C	1.00	1.726	1.765	1.804
PRS	1.00	1.660	1.764	1.876
# of medicines	1.00	1.153	1.168	1.182
Age at recruitment	1.00	1.054	1.056	1.058
Triglycerides	1.00	1.021	1.027	1.033
BMI	1.00	1.008	1.018	1.027
HbA1c	1.00	1.015	1.015	1.016
Male	0.73	0.989	1.007	1.024
Alkaline phosphatase	0.98	1.000	1.001	1.001
Urate	1.00	1.000	1.000	1.000
Has ever smoked	0.45	1.000	1.000	1.000
Lipoprotein A	0.24	1.000	1.000	1.000
Testosterone	0.19	1.000	1.000	1.001
C reactive protein	0.18	1.000	1.000	1.001
Vitamin D	0.12	1.000	1.000	1.000
IGF 1	0.04	1.000	1.000	1.000
Gamma glutamyltransferase	0.01	1.000	1.000	1.000
Alanine aminotransferase	0.00	1.000	1.000	1.000
Aspartate aminotransferase	0.00	1.000	1.000	1.000
Direct bilirubin	0.00	1.000	1.000	1.000
Urea	0.00	1.000	1.000	1.000
Calcium	0.00	1.000	1.000	1.000
Cholesterol	0.00	1.000	1.000	1.000
Creatinine	0.00	1.000	1.000	1.000
Glucose	0.00	1.000	1.000	1.000
LDL direct	0.00	1.000	1.000	1.000
Phosphate	0.00	1.000	1.000	1.000
SHBG	0.00	1.000	1.000	1.000
Total bilirubin	0.00	1.000	1.000	1.000
Total protein	0.00	1.000	1.000	1.000
# of co-morbidities	0.00	1.000	1.000	1.000
No physical activity	0.00	1.000	1.000	1.000
Physical activity	0.00	1.000	1.000	1.000
Alcohol	0.13	0.996	0.999	1.002
Albumin	0.45	0.995	0.998	1.001
Never smoked	1.00	0.925	0.942	0.960
Education level	1.00	0.922	0.933	0.945
Apolipoprotein A	1.00	0.767	0.827	0.890
HDL cholesterol	1.00	0.777	0.816	0.857
Female	1.00	0.519	0.535	0.551

Table C.6: Aggregate values of the proportion selection(Prop. Selec), odds ratio (OR) as well as the OR one standard deviation (SD) away (either side) for the penalised regression methods with covariates, biomarkers and PRS (no BHS).

C.7 sPLS Stability Analyses

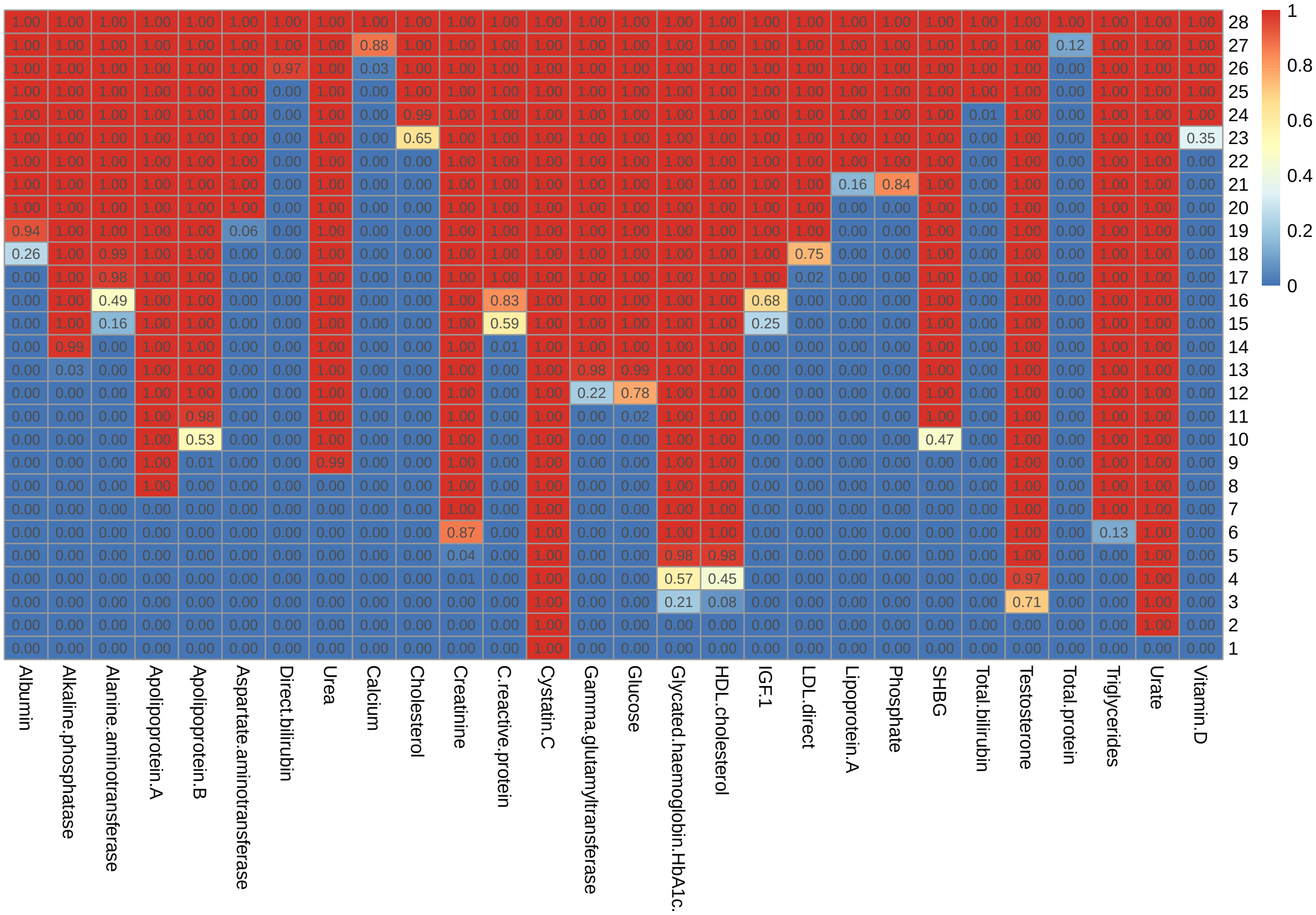


Figure C.4: Heatmap of results from the sPLS-DA stability analysis. The values in each box represent the selection proportion (across the 100 iterations) for each biomarker and each parameter value (number of variables selected by the model).

C.8 Exploratory Analysis Figures

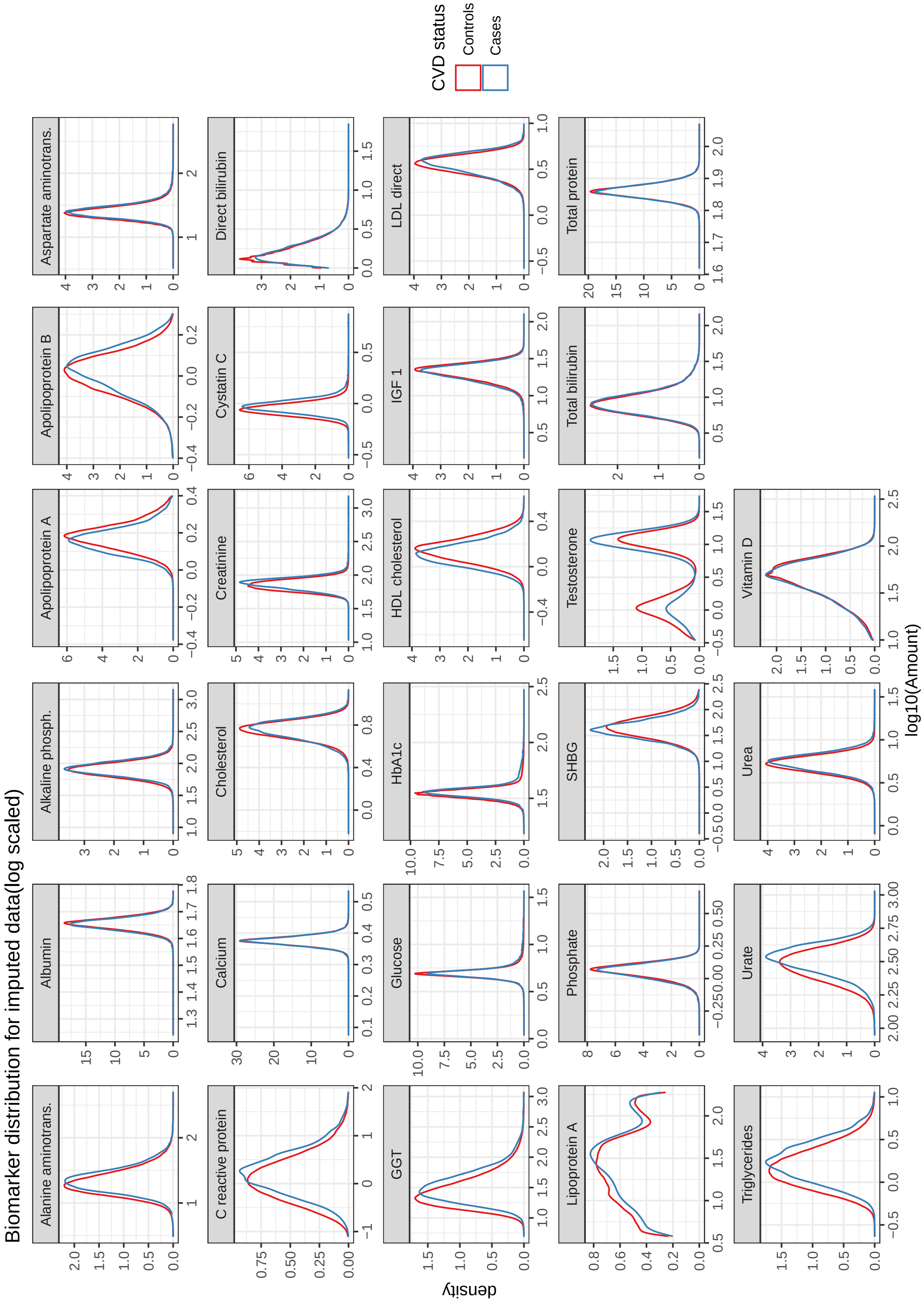


Figure C.5: Biomarkers distribution

C.9 Estimation of KNN imputation error

Estimating error from imputation is a difficult task as because of the nature of data missingness the true value of the missing data is unknown. Not knowing how wrong the performed imputation is, is frustrating. Therefore, and inspired by [12], we decided to write a program that could allows us to estimate the imputation RMSE for all biomarkers and for a range of k nearest neighbours. We did this by first considering only the complete cases in our data and then injecting missingness based on a missing data pattern from the original data set. Then we perform knn imputation with a range of k values and because we know the original values of the missing values we obtain the RMSE for imputation for each choice of k. For robustness, we perform this operation several times (50), each time considering a different set or the original data to obtain the missing data pattern. For computational efficiency we also implemented this program on 40% of the complete data at any instance.

In addition we standardise all biomarkers with respect to their mean and standard deviation, so that we can compare imputation error for biomarkers that have different ranges. Below can be seen the mean RMSE over 50 runs for several values of k (1 to 19 in steps of 2) for all biomarkers considered in this study.

The results are not the most promising given that a lot of biomarkers converge (as k grows) to a mean RMSE of around 1 standard deviation, which if true is a terrible result and kNN imputation should not be considered. Due to this work being experimental and the congruence between imputed and non imputed data in our univariate analysis, we believe kNN imputation does not cause any major disturbances in our data and allows us to have a greater statistical power. Because this work was experimental and we did not dive deeper into this method, we believe the results shown in figure C.7 should not be considered to hold true. There may be intricacies we have not come across or some error in our code, though we believed this method was worth looking into instead of blindly imputing our data with the default parameters of a third party imputation function.

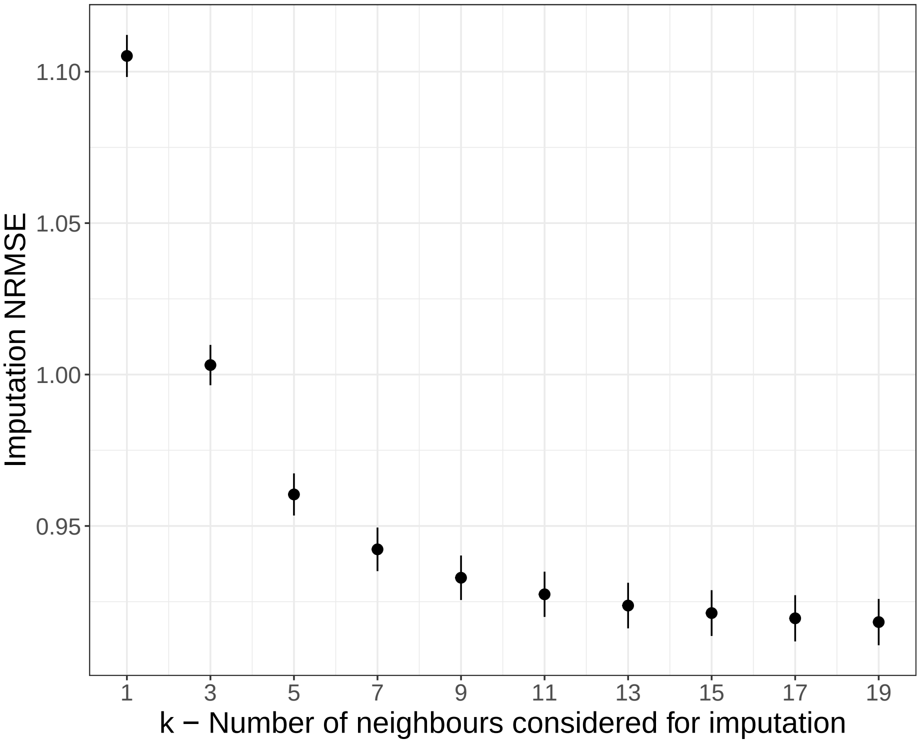
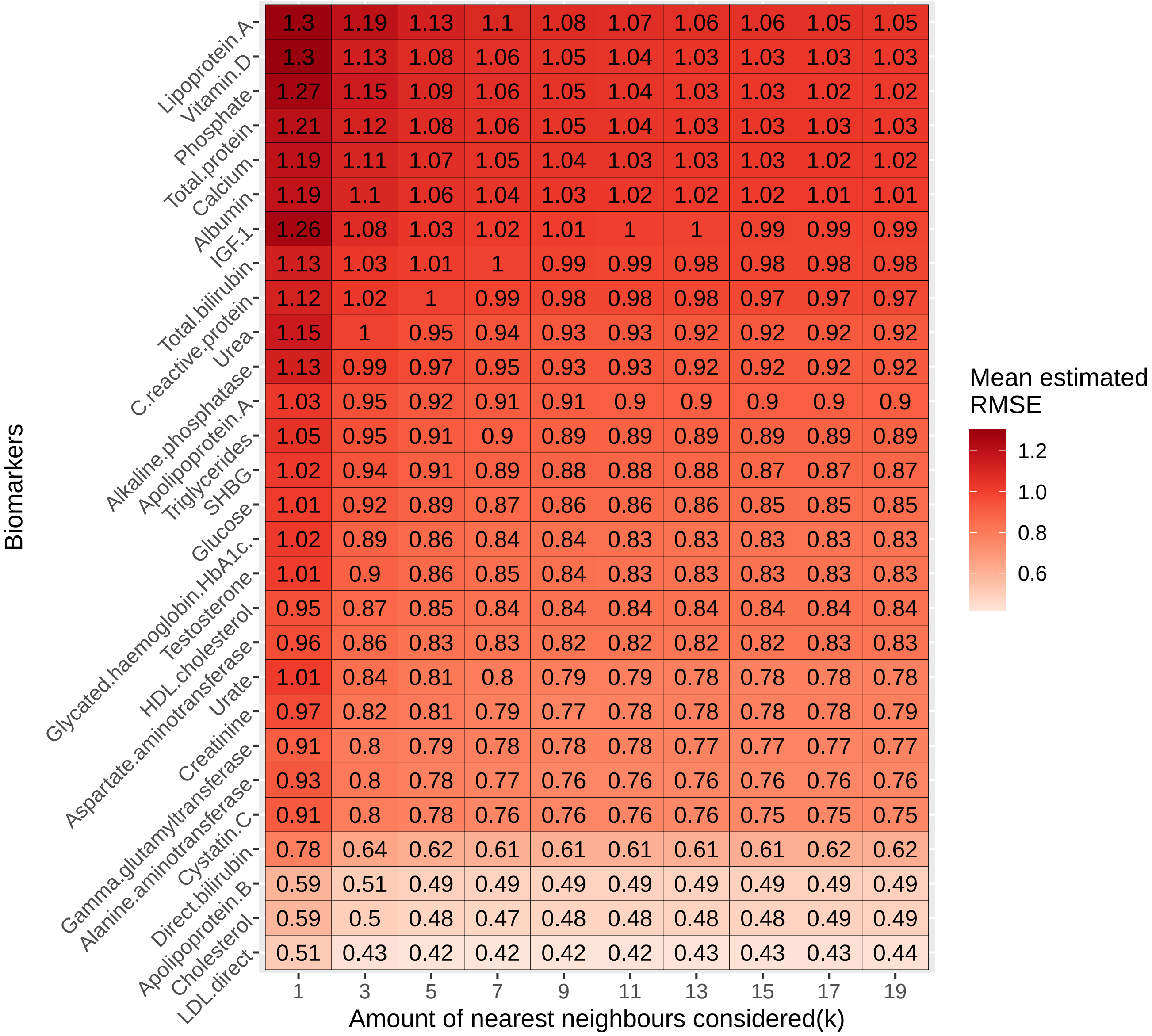


Figure C.6: Mean of the standardised (mean 0, sd 1) estimated imputation RMSE for all possible values of k)



Biomarkers	1	3	5	7	9	11	13	15	17	19
Lipoprotein.A	1.3	1.19	1.13	1.1	1.08	1.07	1.06	1.06	1.05	1.05
Vitamin.D	1.3	1.13	1.08	1.06	1.05	1.04	1.03	1.03	1.03	1.03
Phosphate	1.27	1.15	1.09	1.06	1.05	1.04	1.03	1.03	1.02	1.02
Total.protein	1.21	1.12	1.08	1.06	1.05	1.04	1.03	1.03	1.03	1.03
Calcium	1.19	1.11	1.07	1.05	1.04	1.03	1.03	1.03	1.02	1.02
Albumin	1.19	1.1	1.06	1.04	1.03	1.02	1.02	1.02	1.01	1.01
IGF.1	1.26	1.08	1.03	1.02	1.01	1	1	0.99	0.99	0.99
Total.bilirubin	1.13	1.03	1.01	1	0.99	0.99	0.98	0.98	0.98	0.98
C.reactive.protein	1.12	1.02	1	0.99	0.98	0.98	0.98	0.97	0.97	0.97
Urea	1.15	1	0.95	0.94	0.93	0.93	0.92	0.92	0.92	0.92
Alkaline.phosphatase	1.13	0.99	0.97	0.95	0.93	0.93	0.92	0.92	0.92	0.92
Apolipoprotein.A	1.03	0.95	0.92	0.91	0.91	0.9	0.9	0.9	0.9	0.9
Triglycerides	1.05	0.95	0.91	0.9	0.89	0.89	0.89	0.89	0.89	0.89
SHBG	1.02	0.94	0.91	0.89	0.88	0.88	0.88	0.87	0.87	0.87
Glucose	1.01	0.92	0.89	0.87	0.86	0.86	0.86	0.85	0.85	0.85
Glycated.haemoglobin.HbA1c.	1.02	0.89	0.86	0.84	0.84	0.83	0.83	0.83	0.83	0.83
Testosterone	1.01	0.9	0.86	0.85	0.84	0.83	0.83	0.83	0.83	0.83
HDL.cholesterol	0.95	0.87	0.85	0.84	0.84	0.84	0.84	0.84	0.84	0.84
Aspartate.aminotransferase	0.96	0.86	0.83	0.83	0.82	0.82	0.82	0.82	0.83	0.83
Urate	1.01	0.84	0.81	0.8	0.79	0.79	0.78	0.78	0.78	0.78
Creatinine	0.97	0.82	0.81	0.79	0.77	0.78	0.78	0.78	0.78	0.79
Gamma.glutamyltransferase	0.91	0.8	0.79	0.78	0.78	0.78	0.77	0.77	0.77	0.77
Alanine.aminotransferase	0.93	0.8	0.78	0.77	0.76	0.76	0.76	0.76	0.76	0.76
Cystatin.C	0.91	0.8	0.78	0.76	0.76	0.76	0.76	0.75	0.75	0.75
Direct.bilirubin	0.78	0.64	0.62	0.61	0.61	0.61	0.61	0.61	0.62	0.62
Apolipoprotein.B	0.59	0.51	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49
Cholesterol	0.59	0.5	0.48	0.47	0.48	0.48	0.48	0.48	0.49	0.49
LDL.direct	0.51	0.43	0.42	0.42	0.42	0.42	0.43	0.43	0.43	0.44

Figure C.7: Mean of the standardised (mean 0, sd 1) estimated imputation RMSE for all biomarkers and different values of k (choice of nearest neighbours considered in imputation)