Augmenting clinical risk prediction of cardiovascular disease through protein and epigenetic biomarkers

Aleksandra D Chybowska[1], Danni A Gadd[1], Yipeng Cheng[1], Elena Bernabeu[1], Archie Campbell[1], Rosie M Walker[2], Andrew M McIntosh[3], Nicola Wrobel[4], Lee Murphy[4], Paul Welsh[5], Naveed Sattar[5], Jackie F Price[6], Daniel L McCartney[1], Kathryn L Evans[1], Riccardo E Marioni[1,*]

[1] Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, EH4 2XU, UK.

[2] School of Psychology, University of Exeter, Perry Road, Exeter, EX4 4QG, UK.

[3] Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, EH10 5HF, UK.

[4] Edinburgh Clinical Research Facility, University of Edinburgh, Western General Hospital, Edinburgh, EH4 2XU, UK.

[5] Institute of Cardiovascular and Medical Sciences, BHF Glasgow Cardiovascular Research Centre, University of Glasgow, 126 University Place, Glasgow G12 8TA, UK.

[6] Usher Institute, University of Edinburgh, Old Medical School, Teviot Place, Edinburgh, EH8 9AG, UK.

*Corresponding author:

Name: Riccardo Marioni

Contact Details: riccardo.marioni@ed.ac.u

Abstract

Background and Aims

Cardiovascular disease (CVD) is among the leading causes of death worldwide. Therefore, it is important to identify CVD risk as early as possible and to translate this knowledge into effective preventative strategies. Here, ASSIGN – a cardiovascular risk calculator recommended for use in Scotland – was examined in tandem with epigenetic and proteomic features in risk prediction models in 12,790 participants in the Generation Scotland cohort.

Methods

Previously generated DNA methylation-derived epigenetic scores (EpiScores) for 109 protein levels were considered, in addition to both measured levels and an EpiScore for cardiac troponins. The associations between potentially informative multi-omic features and the disease status were examined using univariate Cox PH models (n_{cases}=1,288). Splitting the cohort into independent training (n≥3,681) and test (n≥1,996) subsets composed of unrelated individuals, two composite scores were developed: CVD TroponinScore and CVD EpiScore.

Results

There were 65 protein EpiScores that associated with incident CVD (n_{cases}=1,288) independently of ASSIGN ($P_{Bonferroni}$<0.05), over a follow up of up to 15 years of electronic health record linkage. The CVD TroponinScore (based on the concentration of two cardiac troponins) and CVD EpiScore (based on 52 protein EpiScores) were both associated with CVD risk in Cox models in test sets (Hazard Ratio HR=1.28 and 1.23, P=5.9×10^{-4} and 3.8×10^{-3}, n_{test}=1,996 and 3,711, respectively).

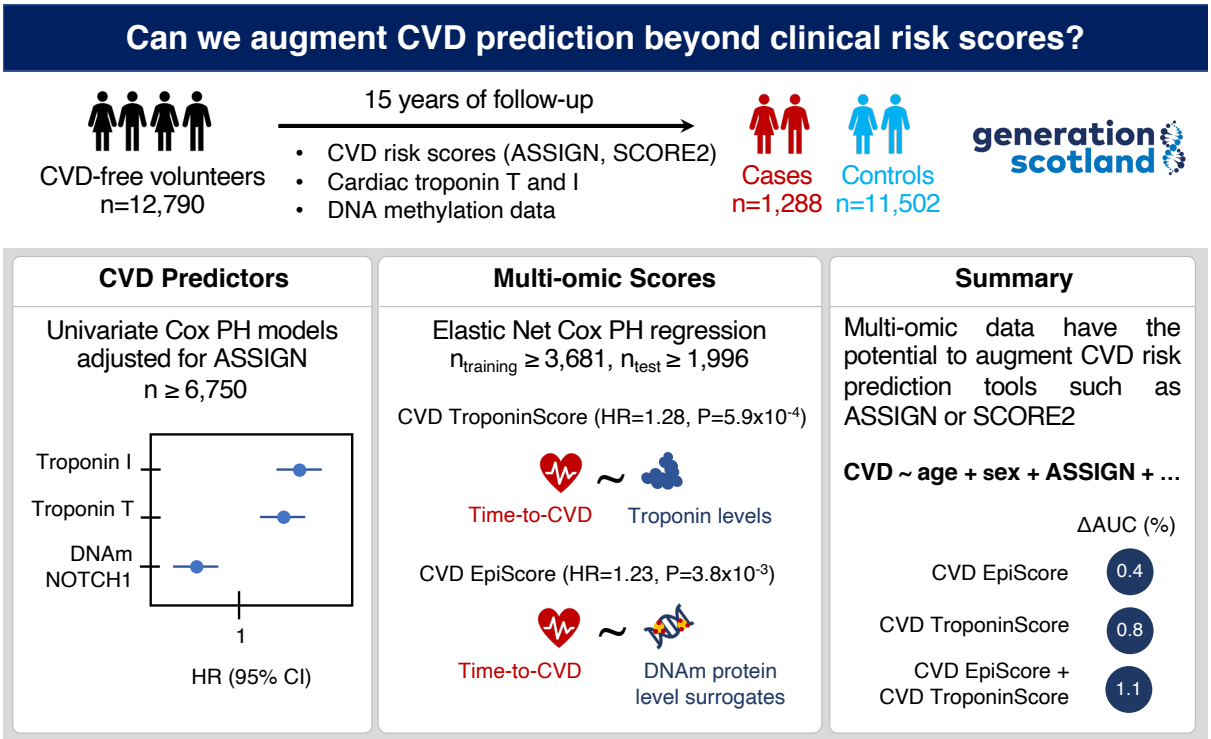
Conclusions

EpiScores for circulating protein levels have the potential to improve the prediction of CVD and be useful tools for precision medicine.

Keywords

ASSIGN, SCORE2, epigenetics, EpiScores, multi-omics, CVD prediction

Structured graphical abstract



Structural graphical abstract. CVD – Cardiovascular Disease, EpiScore – Epigenetic Score, Cox PH – Cox Proportional Hazards Regression, DNAm – DNA methylation

Structured abstract text

Key Question: Can we augment CVD prediction beyond clinical risk scores?

Key Finding: Newly derived multi-omic scores (CVD TroponinScore and CVD EpiScore) were both associated with CVD risk in holdout test sets.

Take-home Message: Multi-omic data have the potential to augment CVD risk prediction tools such as ASSIGN or SCORE2.

Introduction

For the past 20 years, cardiovascular disease (CVD) has been a leading cause of mortality, accounting for nearly one-third of global deaths in 2019 [1]. Although our knowledge of risk factors and treatment methods has improved, the prevalence of CVD is still increasing [1]. This trend may be partially explained by the rising prevalence of CVD risk factors, such as hypertension or obesity [2,3].

To help clinicians assess the chance of future CVD events in asymptomatic people, risk prediction tools have been developed. Most are based on conventional risk factors observed in a cohort of participants recruited either from a general population or from those in primary care. ASSIGN (Assessing cardiovascular risk using SIGN guidelines to ASSIGN preventive treatment) is a risk prediction tool recommended for use in Scotland [4]. It incorporates age, sex, smoking status, systolic blood pressure, total cholesterol, high density lipoprotein cholesterol, family history of premature CVD, diagnosis of rheumatoid arthritis, diagnosis of diabetes, and a deprivation score (the Scottish Index of Multiple Deprivation – SIMD) [5] to estimate the 10 year risk of a CVD diagnosis. ASSIGN was developed using data collected from the general population, with participants recruited during the 1980s and 1990s.

Given the wealth of biomarker and omics data now available across many cohort studies, it may be possible to augment existing risk prediction models. Studies have shown that cardiac-specific troponin I (cTnI) and troponin T (cTnT) are highly sensitive markers of myocardial damage [6]. The concentration of cardiac troponin rapidly rises following cardiac tissue degradation and for this reason, it is used in clinical settings to detect myocardial infarction. In the general population, a high-sensitivity cardiac troponin test may play a role in CVD diagnosis and risk stratification [6]. However, elevation of cardiac troponin in the bloodstream

can be also observed in non-cardiac conditions, which can lead to a confusion with myocardial damage [7]. Elevated cardiac troponin levels have been also reported in healthy individuals after strenuous exercise [8] and in kidney disease patients, as a result of impaired renal clearance [9].

It may be possible to derive stable biomarkers of CVD risk factors using blood-based DNA methylation (DNAm). DNAm is an epigenetic modification characterised by the addition of methyl groups to DNA. Typically, the methyl group is added at cytosine-phosphate-guanine (CpG) dinucleotides, which are found mostly (but not exclusively) in gene promoters [10]. Blocking promoters, to which activating transcription factors should bind in order to initiate transcription, is one of the mechanisms by which DNA methylation can precisely regulate gene expression [11]. Previous studies have shown that DNAm-derived proxies for smoking [12] and BMI [13] can show stronger associations with health outcomes than their self-reported or measured counterparts. We have also shown that DNAm-based epigenetic scores for 109 circulating protein levels – known as protein EpiScores – are associated with the time-to-diagnosis for a host of leading causes of morbidity and mortality, including cardiovascular outcomes [14]. Protein EpiScores are therefore useful biomarker tools for disease risk stratification. Protein EpiScores capture long-term changes in the biological state of an individual and may be more precise than singular protein measurements, which are known to fluctuate between measurements [15–18].

Here, we examine whether the ASSIGN score prediction of CVD hospitalisation can be augmented using multi-omics data from 12,790 participants of the Generation Scotland (GS) study. We first run individual Cox proportional hazards (PH) models to test relationships between individual omic features (namely protein EpiScores and measured levels of cardiac troponin) and incident CVD in the full sample. We then create a CVD EpiScore (based on the protein EpiScores) and a CVD TroponinScore (based on measured cardiac troponins) and test the additional predictive performance offered by these scores for CVD risk stratification.

Methods

Generation Scotland

Generation Scotland is a population-based and family-structured cohort study of individuals from Scotland [19]. Between 2006 and 2010, patients at collaborating general medical practices in Scotland aged 35-65 years were invited to join the study. Subsequently, participants were asked to identify first-degree relatives aged 18 and over who were then invited to participate. 24,088 participants, aged 18-99 years, completed a health survey. Clinical and physical characteristics of 21,521 individuals who attended a clinic were measured using a standardized protocol. Fasting blood samples were obtained in clinic using a standard operating procedure.

Measurement of high-sensitivity troponin

The concentrations of high sensitivity cTnT (Roche Diagnostics) and high sensitivity cTnI (Abbott Diagnostics) were measured in 10201 and 19130 GS individuals, respectively. To facilitate measuring the concentration of both proteins together during the first thaw of stored serum samples, a standard operating procedure was created [20]. Before the assay, samples were spun for 5 minutes at 2000g. The measurements were taken using i1000SR and Cobas e411

devices using the manufacturers' quality controls and calibrators. The limits of detection set by manufacturers of these devices are 3 ng/L for cTnT and 1.2 ng/L for cTnI. Anything below these limits is reported as a blank value. The results below the limit of blank were analysed as both raw concentrations and corrected concentrations. Correcting for blank values consisted of reporting results below the threshold of blank as 1.5 ng/L for cTnT and 0.6 ng/L for cTnI [6].

DNA Methylation

DNAm was profiled in blood samples using the Illumina EPIC array. Quality control details have been described previously [21]. Filtering for poorly detected probes and samples, non-blood (e.g., saliva) samples and outliers was performed. Subsequently, non-CpG probes as well as probes on the X and Y chromosomes were removed. Missing CpG values were mean-imputed. To ensure any signatures generalise to as many cohorts as possible, the sites were subset to those also present on the older Illumina 450k array (n=453,093 CpGs). The full quality-controlled dataset contained 18,413 individuals.

Methylation profiling was carried out in three sets. Set 1 contained 5,087 individuals, Set 2 contained 4,450 individuals and Set 3 contained 8,876 individuals. Participants in Set 2 were genetically unrelated to each other and to those in Set 1 (genetic relationship matrix (GRM) threshold <0.05) while more complex relationship structures were present within and between Sets 1 and 3.

The 109 protein EpiScores described by Gadd *et al.* [14] were projected into the cohort via the publicly available MethylDetectR Shiny App [22].

ASSIGN and SCORE2 scores

ASSIGN scores were calculated for 16,366 GS individuals with complete information about the component parts. Variables were adjusted as per official guidance (*FAQs - ASSIGN Score*, 2022). This included modifying the number of cigarettes smoked per day in recent quitters, adding 20mmHg to systolic blood pressure in individuals on blood pressure medications, and adding 10 to the number of cigarettes smoked per day in rheumatoid arthritis patients. The ASSIGN score was calculated in R using publicly available coefficients and formulae [24]. A subset of the obtained scores was validated against the scores produced by the online tool [25].

SCORE2 values were calculated for 16,934 GS individuals using external coefficients [26]. The estimates were calibrated according to region-specific scaling factors and validated against the online tool [27].

The R scripts used to derive ASSIGN and SCORE2 are available at https://github.com/aleksandra-chybowska/troponin_episcores.

Cardiovascular disease events

CVD cases were ascertained through data linkage to NHS Scotland hospital records. Participants were followed to the end of September 2021. A composite CVD outcome was defined as per Welsh *et al.* [6] and included the following *International Classification of Diseases, 10th Revision codes*: I20–25, G45, I60–69, I00–I99, L29.5, L31.1, K40–46, K49, K75. A total of 1,288 incident CVD cases were observed over a follow up period of up to 15 years **(Supplementary Figure 1).**

Statistical analysis for risk associations

Cox PH models adjusted for the ASSIGN score were prepared to investigate the relationship between a composite CVD outcome and the following potential biomarkers: measured cardiac troponin, troponin EpiScores, and 109 protein EpiScores. These predictors were rank-based inverse normalised prior to the analyses. CVD cases comprised individuals diagnosed after baseline who subsequently died as well as of those who remained alive after receiving a diagnosis. Controls were censored at the end of the follow up period (September 2021) or at time of death (CVD-free survival). Models based on data from merged Sets 1, 2, and 3 were generated using coxme library (v 2.2.16) with a kinship matrix fitted as a random effect to adjust for relatedness. Cox model assumptions were checked by running coxph models. These models were not adjusted for relatedness. The cox.zph function from the survival library (v. 3.2.7) [28] was used to examine Schoenfeld residuals (local and global tests with significance threshold $p < 0.05$).

Composite scores

Two composite scores for incident CVD were trained using Cox PH elastic net models (**Supplementary Figure 2**).

1. The first was an epigenetics-based score that considered the 109 DNAm scores reflecting the concentrations of plasma proteins (protein EpiScores) from Gadd *et al* [14] as potentially informative features ($n_{training}$=6,936, n_{test}=3,711). This is referred to as the *CVD EpiScore*.
2. The second was a proteomics-based score that considered raw concentration of cTnT and a corrected concentration of cTnI ("blank" correction) as potentially informative features ($n_{training}$=3,681, n_{test}=1,996). This is referred to as the *CVD TroponinScore*.

Protein EpiScores and troponin concentrations were rank-based inverse normalised prior to training. Models were trained using the glmnet package (v 4.0.2) [29]. The L1-L2 mixing parameter was set to alpha=0.5, and ten-fold cross-validation was applied. The features with non-zero coefficients were used to generate composite scores for CVD – a weighted linear combination of individual features. Finally, to test whether the CVD EpiScore and CVD TroponinScore measures improved prediction of CVD over and above ASSIGN, the following Cox PH models were fit to the test set: a null model adjusted for age, sex, and ASSIGN, then full models, which included age, sex, ASSIGN score and 1) the CVD EpiScore, 2) the CVD TroponinScore and 3) the CVD EpiScore + the CVD TroponinScore.

Evaluating Performance of the Composite Scores

Cox PH analysis can be used model the risk of incident CVD for different prediction periods. As the ASSIGN score estimates the risk of developing CVD over 10 years, we used 10-year CVD incidence to evaluate the classification performance of our models. To calculate the binary CVD outcome, time-to-event was truncated at 10-years. Prediction probabilities were obtained in the test set for each model by calculating the cumulative baseline hazard using the Breslow estimator [30]. Prediction metrics (Area Under the Curve (AUC), Precision-Recall AUC and C-index) were then generated using the pROC package [31].

Results

Study overview

A series of Cox PH models were run with time-to-CVD as the outcome. The first considered ASSIGN as the only predictor. Subsequent models were then adjusted for the ASSIGN score and a different potential predictor of CVD (Figure 1). This was followed by a prediction analysis where composite scores were trained, based on either 109 protein EpiScores (referred to as the CVD EpiScore), or cardiac troponin protein measurements (referred to as the CVD TroponinScore). The CVD EpiScore and TroponinScores were then assessed in the test sample, to quantify the additional predictive performance offered by these scores over and above ASSIGN. This was tested in Cox PH models that adjusted for age, sex, and ASSIGN with CVD as the outcome.

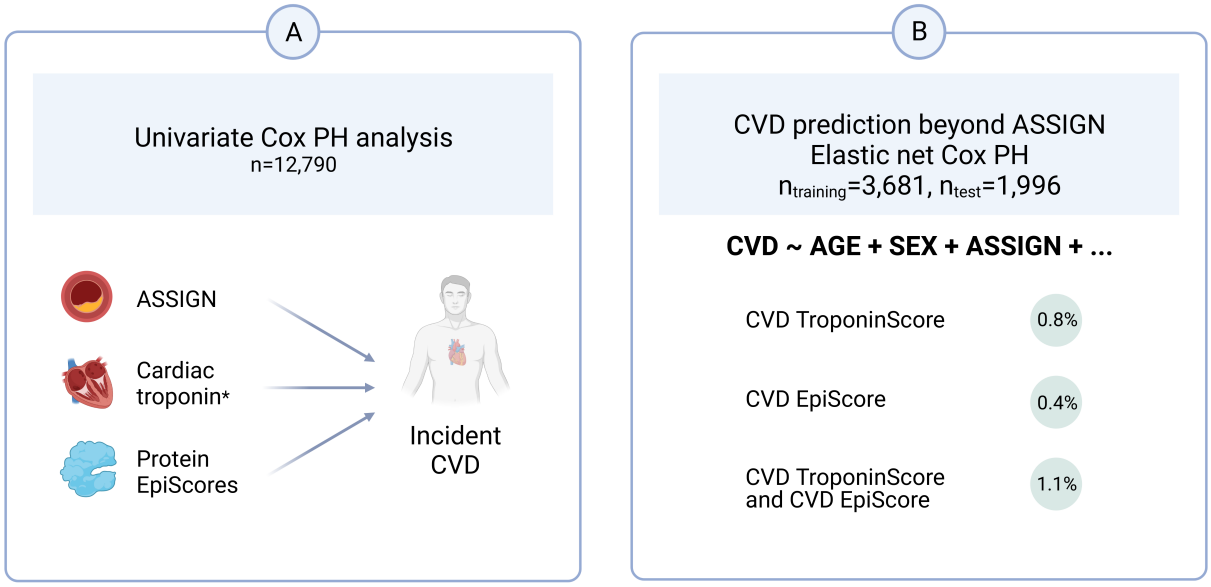


Figure 1. The main steps of the analysis pipeline. The numbers in the green circles correspond to increases in the Area Under the Curve observed after adding the newly created predictors to the model. *$n_{troponin} \geq 6{,}750$. Created with BioRender.com

Calculating ASSIGN and comparing it with SCORE2

ASSIGN scores were calculated for 16,366 individuals who had complete information available with 3,918 (23.9%) receiving a score higher than the clinically used threshold of 20. The rank-based inverse normalised ASSIGN score was strongly positively correlated with age (Spearman r=0.90). More than 80% of hospital admissions associated with CVD concerned people aged 50 or over. ASSIGN risk scores were also strongly positively correlated with estimates produced by SCORE2 – an algorithm validated to predict 10-year CVD risk in European populations. In a sample of GS individuals within the recommended SCORE2 age range (40-69 years old, n = 11,348), the Spearman's correlation coefficient of ASSIGN and SCORE2 was 0.89 (**Supplementary Figure 3**). Given that the ASSIGN score was tailored to Scottish population, we decided to use it as the primary CVD risk measure, with SCORE2 used in sensitivity analyses.

Cox model assumptions

Splitting the dataset by decade of age revealed strong associations between the transformed ASSIGN score and CVD risk in those under 30 (n=2,580, n_{events}= 20, HR per SD of ASSIGN = 2.42, 95% CI 1.43, 4.10; P=9.4×10^{-4}) and weaker associations in those aged 70 or more (n = 803, n_{events}= 252, HR per SD of ASSIGN=1.27, 95% CI 1.12, 1.44; P=2.3×10^{-4}). The results split by decade are presented in **Supplementary Table 1**. To meet the proportional hazards assumption of the Cox model, the dataset was filtered to individuals between 30 and 70 years old. To minimise the impact of outliers on the analyses, ASSIGN scores were trimmed of outliers (points beyond 3 SDs of the mean, n=181). This left a cohort of 12,790 individuals for downstream analysis. **Table 1** summarises the training, test, and full datasets after applying the above-mentioned corrections.

Table 1. Summary of training, test, and full datasets prior to adding cardiac troponin to the models. For continuous variables with normal distributions, summary values are reported as mean (SD). Median [Q1, Q3] are given for continuous variables which do not follow a normal distribution. A number and a percentage of samples are reported for categorical variables. SIMD – Scottish Index of Multiple Deprivation, CHD – Coronary Heart Disease.

	Training		Test		Full	
	Cases	**Controls**	**Cases**	**Controls**	**Cases**	**Controls**
n	661	6275	345	3366	1288	11502
Time-to-event (Years to Onset or Censoring)	7.0 [4.1, 9.8]	11.8 [11.1, 13.0]	4.9 [2.7, 7.6]	11.8 [11.0, 13.6]	6.8 [3.7, 9.8]	11.8 [11.1, 13.2]
Age (Years)	58.3 [50.8, 62.6]	50.0 [40.7, 58.8]	56.5 [51.5, 60.0]	51.5 [43.9, 57.5]	57.3 [50.9, 62.2]	50.4 [41.5, 58.2]
Sex (Male)	346 (52.3%)	2475 (39.4%)	168 (48.7%)	1229 (36.5%)	661 (51.3%)	4439 (38.6%)
SIMD (score/10)	41.5 [22.2, 53.2]	45.3 [26.2, 55.0]	44.6 [19.8, 54.8]	44.3 [22.6, 54.6]	41.6 [21.5, 54.0]	44.7 [25.5, 54.8]
Family history of CHD/Stroke (Yes)	443 (67.0%)	3193 (50.9%)	227 (65.8%)	1808 (53.7%)	868 (67.4%)	5941 (51.7%)
Diabetes (Yes)	20 (3.0%)	67 (1.1%)	17 (4.9%)	75 (2.2%)	46 (3.6%)	173 (1.5%)
Rheumatoid arthritis (Yes)	29 (4.4%)	140 (2.2%)	23 (6.7%)	111 (3.3%)	71 (5.5%)	282 (2.5%)
Non-smoker (Yes)	537 (81.2%)	5344 (85.2%)	286 (82.9%)	2786 (82.8%)	1056 (82.0%)	9714 (84.5%)
Systolic blood pressure (mmHg)	142.1 (16.7)	130.8 (16.9)	140.5 (17.7)	130.3 (16.5)	141.6 (17.2)	130.6 (16.8)
Total cholesterol (mmol/l)	5.4 (1.1)	5.2 (1.0)	5.3 (1.1)	5.3 (1.1)	5.4 (1.1)	5.2 (1.0)
HDL cholesterol (mmol/l)	1.3 [1.1, 1.6]	1.4 [1.2, 1.7]	1.4 [1.1, 1.6]	1.5 [1.2, 1.8]	1.3 [1.1, 1.6]	1.4 [1.2, 1.7]
ASSIGN score	19 [12, 29]	9 [4, 18]	18 [11, 28]	10 [5, 17]	18 [12, 28]	9 [4, 17]

Incremental model using cardiac troponin and cardiac troponin EpiScores

We tested whether measured cardiac troponins or EpiScores for cardiac troponins (see **Supplementary Methods** for details) were associated with CVD risk over and above ASSIGN. cTnT was associated with a HR of 1.19 per SD increase (95% CI 1.10, 1.28; $P=3.7x10^{-6}$), while cTnI corrected was associated with an HR of 1.20 per SD increase (95% CI 1.13, 1.29; $P=1.9x10^{-8}$). Both troponins remained significant when modelled together – the cTnT HR was 1.14 (95% CI 1.06, 1.23, $P=4.6x10^{-4}$), while corrected cTnI had a HR of 1.15 (95% CI 1.06, 1.25; $P=1.3x10^{-3}$).

In contrast, EpiScores generated for measured troponin concentrations did not predict CVD risk in Cox models adjusted for ASSIGN. For cTnT, the blood DNAm signal underlying differences in the protein level was too weak to generate a meaningful EpiScore (incremental R^2= 0.004%, P=0.74). For cTnI, after replacing values below the limit of blank with half of this limit, a significant but weak EpiScore was derived (incremental R^2=0.186%, $P=2.3x10^{-3}$). However, this score did not associate with incident CVD (P=0.59). As a result, the EpiScores for troponins were not considered as features in the generation of the composite CVD score.

Incremental model using EpiScores for plasma protein levels

We then tested whether 109 protein EpiScores generated by Gadd *et al.* (protein description available in **Supplementary Table 2**) were associated with the incidence of CVD over the follow up of 15 years [14]. 109 Cox models were created with each adjusting for the ASSIGN score and a different protein EpiScore. Five EpiScores failed to satisfy the proportional hazards assumption (Schoenfeld residual test $P > 0.05$). Of the remaining 104 protein EpiScores, 65 were significantly associated with CVD risk ($P<0.05$, **Figure 2**). After applying a conservative Bonferroni threshold for multiple testing ($P<0.05/104 = 4.8\times10^{-4}$) 37 associations remained statistically significant.

Higher levels of 38 protein EpiScores were associated with an increased hazard of CVD (HR>1 and $P<0.05$). For example, elevated levels of *CRP* and *MMP12* were associated with HR of 1.22 (95% CI 1.15, 1.29; $P=1.7\times10^{-11}$) and 1.13 (95% CI 1.05, 1.22, $P=6.4\times10^{-4}$), respectively. In contrast, higher levels of 27 protein EpiScores were associated with a decreased hazard of CVD (HR<1 and $P<0.05$). Examples of protein EpiScores belonging to this group include *NOTCH1* (HR per SD of 0.83, 95% CI 0.78, 0.88; $P=1.5\times10^{-10}$) and *OMD* (HR per SD of 0.86, 95% CI 0.81, 0.91; $P=9.0\times10^{-8}$).

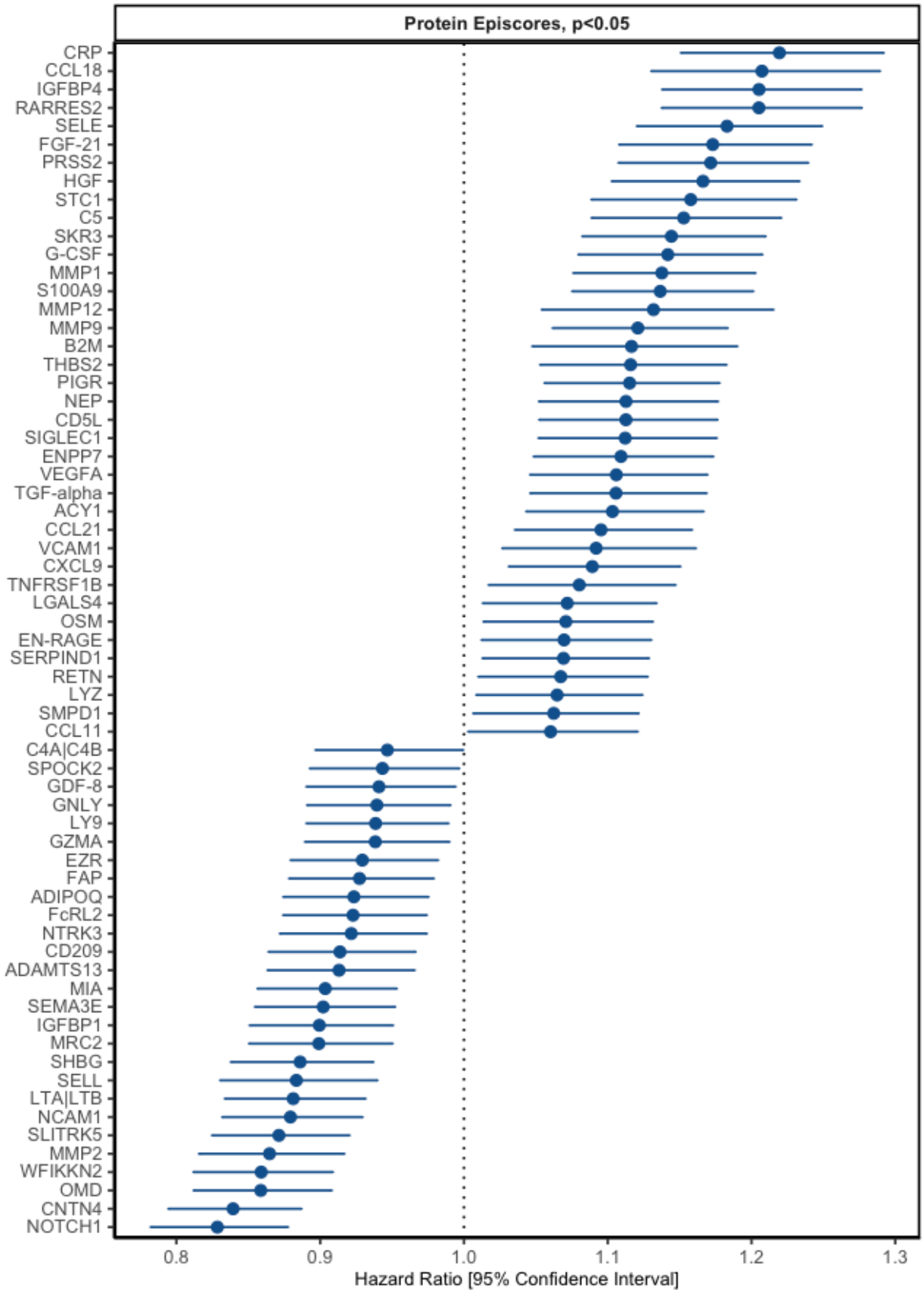


Figure 2. Associations between protein EpiScores and incident CVD. Hazard ratios are plotted for the 65 significant associations with 95% confidence interval limits.

Composite EpiScore and TroponinScore for CVD prediction

Two composite scores for incident CVD prediction were created using Cox PH elastic net models. The first score – known as the CVD EpiScore – was trained using protein EpiScores from Gadd *et al* as potentially informative features ($n_{training}$=6,936, n_{test}=3,711). Of 109 possible protein EpiScores, 52 were selected as features and assigned non-zero coefficients. The weights and annotations for the selected protein EpiScores can be found in **Supplementary Tables 2** and **3.** The second – known as the CVD TroponinScore – considered the two measured troponins as potentially informative features ($n_{training}$=3,681, n_{test}=1,996). Both troponin concentrations were selected as contributing features (weights are shown in **Supplementary Table 4**).

To understand whether the CVD EpiScore and CVD TroponinScore can be used as biomarkers that add additional predictive value over-and-above typically used clinical risk scores, such as ASSIGN and SCORE2, a series of 10-year CVD-risk prediction models were assessed in a hold-out test set with 212 cases and 1,784 controls (n=1,996).

The null model (containing age, sex, and ASSIGN) had an AUC of 0.707. The model with the CVD EpiScore increased the AUC to 0.711, but there was no significant differences in the ROC curves ($P_{EpiScore\ vs\ null}$ = 0.38). The CVD EpiScore was not a significant predictor in the Cox PH model (HR) but was significant with a similar effect size in a larger test set that did not pre-filter for missingness in the troponin data (HR=1.23, P=3.8x10^{-3}, n=3,711, n_{cases}=345). The addition of the CVD TroponinScore (based on measured troponins) to the null model resulted in an AUC of 0.715 (ROC test $P_{TroponinScore\ vs\ null}$= 0.23). In this model, the CVD TroponinScore was a significant predictor of incident CVD (HR=1.28, P=5.9x10^{-4}). The full model (null model + CVD TroponinScore + CVD EpiScore) AUC was 0.718 (ROC test $P_{full\ vs}$

$_{null}$ = 0.16, **Figure 3**). Full output for the CVD models including ASSIGN (and SCORE2) can be found in **Supplementary Tables 5** and **6**, respectively. ASSIGN showed better discrimination than SCORE2, but both the CVD EpiScore and CVD TroponinScore were significant predictors in the full SCORE2 model (P=0.025 and 7.2×10^{-4}, respectively).

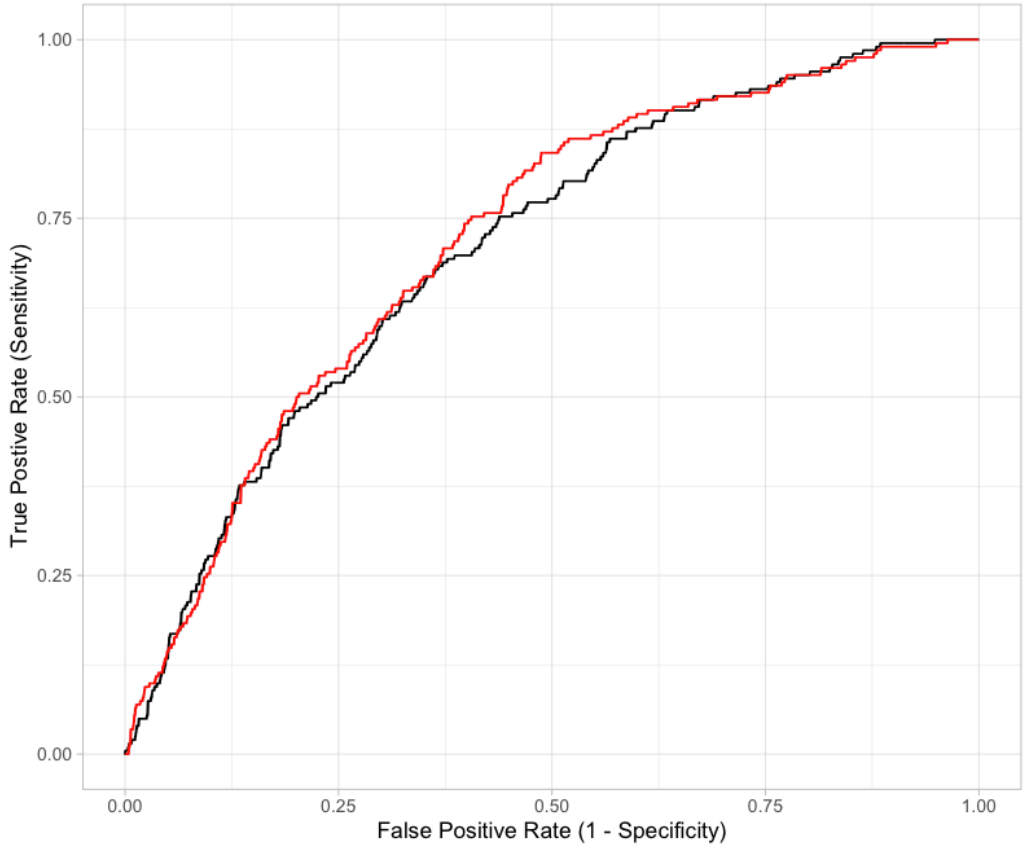


Figure 3. Comparison of model based on age, sex, and ASSIGN (black) with the full multi-omic model (red). The full model (red) considered both the composite CVD TroponinScore (based on measured troponins) and the CVD EpiScore (based on protein EpiScores) and was adjusted for age, sex, and the ASSIGN score. The full model had slightly better discrimination when compared to the null model (AUC improved by 1.1%, ROC test P = 0.16).

Discussion

This study directly compared the added contribution of multi-omics biomarkers to a clinical prediction tool (ASSIGN) for 10-year risk of CVD. Cox regression analyses highlighted 65 protein EpiScores that were significant over and above ASSIGN. When splitting the cohort into training and test sets, newly derived CVD TroponinScore and CVD EpiScore were both associated with CVD risk in inferential models. This was complemented by suggestive evidence from a formal prediction analysis (AUC increase of 1.1%, **Supplementary Table 5**).

One previous study focused on how DNAm biomarkers improve CVD risk prediction [32]. Using time-to-event data and a panel of 60 blood DNAm biomarkers measured in an Italian cohort of 1803 individuals (295 cases), Cappozzo *et al.* trained a composite score for predicting short-term risk of CVD. In comparison, we focused on a more extensive panel of DNAm protein markers in addition to measured troponin levels. We also ran univariate analyses to identify individual proteins and protein classes that associate with CVD. Furthermore, we developed10-year prediction models (the prediction window for which both ASSIGN and SCORE2 are recommended) trained on more than double the number of cases.

Our findings suggest that individual protein EpiScores capture disease-specific biomarker signals relevant to CVD risk prediction. The relationships found between 65 protein EpiScores and incident CVD mirrored previously reported associations between CVD and measured protein concentrations. For example, elevated levels of *CRP*, a marker for systemic low-grade inflammation, have been associated with multiple age-related morbidities, including CVD [33]. *MMP12* and *OMD*, in turn, are involved in maintaining the stability of atherosclerotic plaques. While *MMP12* contributes to the growth and destabilisation of plaques [34], increased levels of

OMD have been observed in macro-calcified plaques from asymptomatic patients [35]. Finally, multiple studies have demonstrated that *NOTCH1* signalling protects the heart from CVD-induced myocardial damage. The Notch1 pathway is involved in neo-angiogenesis and revascularisation of failing heart [36]. It limits the extent of ischemic injury [36], reduces fibrosis [37] and improves cardiac function [38]. Several of the protein EpiScores associated with CVD in our study, such as *SELE* and *C5*, have also been shown to associate with stroke and ischemic heart disease in our previous work [14].

The inferential association between the CVD EpiScore (based on 52 protein EpiScores) and incident CVD suggests that DNAm differences associated with plasma protein levels have the potential to improve current clinical risk prediction tools. As the sample size for training increases, it is expected that both the CVD TroponinScore and CVD EpiScore will augment prediction analyses, boosting the suggestive evidence we observed in the current data. Taking into account the commercial availability of high sensitivity cardiac troponin tests, incorporating these measures into CVD risk models may be a relatively straightforward and affordable way to augment clinical risk prediction tools. Future studies might also consider polygenic risk scores as well as other measured proteins, plus their EpiScores, where available, as features for the prediction analyses.

The protein EpiScores we trained for cardiac troponins T and I were too weak to serve as surrogates for the measured protein concentrations and were therefore excluded from composite CVD score generation. This highlights an important consideration in the development of multi-omics biomarkers, as there are unlikely to be DNAm differences that associate with every blood protein. For example, the 109 protein EpiScores generated by Gadd *et al* that we make use of in our study were extracted as the best-performing EpiScores from a

total set of 953 proteins tested as outcomes [14]. It is therefore not always possible to generate a meaningful protein EpiScore that reflects the protein biology. In the case of troponins, the lack of DNAm signal may be due to the sample source (blood) as cardiac troponin is released upon myocardial damage. As DNAm is tissue and cell specific, blood may not reflect epigenetic changes occurring in the failing heart [39,40]. Moreover, following myocardial damage, circulating levels of cTnT and cTnI increase mainly due to the release of these proteins from the heart and not due to increased transcription [41]. Therefore, the methylation signal associated with the transcription of cardiac troponin may be weak.

Strengths of this study include the precise timing of the CVD event through the electronic health records, the ability to generate a clinical risk predictor in a population cohort, and the very large sample size for DNAm, which also permitted the splitting of the data into train/test sets to formally examine the improvement in risk prediction from our omics biomarkers.

Limitations to this work include the generalisability beyond a Scottish population. In this study, we trained and tested predictors in a Scottish cohort to augment the ASSIGN score. However, the protein EpiScores were trained in a German cohort (KORA) and projected to Generation Scotland [14]. Although the ASSIGN score is tailored to the Scottish population, we observed similar findings across all models when replacing it with SCORE2, which is widely used across Europe.

In conclusion, the widely applied CVD clinical risk prediction tool ASSIGN may be augmented via the integration of epigenetic and proteomic biomarkers. As DNAm and proteins are assessed in increasingly large cohort samples, it will be possible to evaluate more precisely the

potential gains in risk prediction, disease prevention and any associated health economic benefits.

Acknowledgements

We are grateful to all the families who took part, the general practitioners and the Scottish School of Primary Care for their help in recruiting them, and the whole Generation Scotland team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, healthcare assistants and nurses. We would also like to thank Dr Robert Hillary for helpful feedback on the manuscript and analyses.

Funding

This research was funded in whole, or in part, by the Wellcome Trust (104036/Z/14/Z, 108890/Z/15/Z, 220857/Z/20/Z, and 216767/Z/19/Z). For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. A.D.C is supported by a Medical Research Council PhD Studentship in Precision Medicine with funding by the Medical Research Council Doctoral Training Programme and the University of Edinburgh College of Medicine and Veterinary Medicine.

D.A.G. is funded by the Wellcome Trust 4-year PhD in Translational Neuroscience–training the next generation of basic neuroscientists to embrace clinical research (108890/Z/15/Z).

E.B. and R.E.M. are supported by the Alzheimer's Society major project grant AS-PG-19b-010.

Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates (CZD/16/6) and the Scottish Funding Council (HR03006). Genotyping and DNA methylation profiling of the GS samples was carried out by the

Genetics Core Laboratory at the Clinical Research Facility, University of Edinburgh, Scotland and was funded by the Medical Research Council UK and the Wellcome Trust (Reference 104036/Z/14/Z, 220857/Z/20/Z, and 216767/Z/19/Z). The DNA methylation profiling and analysis was supported by Wellcome Investigator Award 220857/Z/20/Z and Grant 104036/Z/14/Z (PI: AM McIntosh) and through funding from NARSAD (Ref: 27404; awardee: Dr DM Howard) and the Royal College of Physicians of Edinburgh (Sim Fellowship; Awardee: Dr HC Whalley). All components of GS:SFHS, including the protocol and written study materials, have received formal, national ethical approval from NHS Tayside Research Ethics Committee (REC reference number 05/S1401/89). In addition local approval has been obtained from NHS Glasgow Research Ethics Committee, and from NHS Glasgow and NHS Tayside Research and Development Offices, as is required.

Author contributions

R.E.M., and A.D.C. were responsible for the conception and design of the study. A.D.C carried out the data analyses. D.A.G, Y.C and E.B. contributed to the analyses and methodology. R.E.M. and A.D.C. drafted the article. A.C. facilitated data linkage. K.L.E. and J.F.P. were involved in conceptualisation and provided consultation on the methodology. S.W.M., R.M.W., N.W., L.M., C.S., A.W.M., K.L.E contributed to data collection and preparation. All authors read and approved the final manuscript.

Conflict of interest

R.E.M. and L.M. have received a speaker fee from Illumina. R.E.M. is an advisor to the Epigenetic Clock Development Foundation. D.A.G. and R.E.M. have received consultancy fees from Optima partners. All other authors declare no competing interests.

Data availability

All code used in the analyses is available with open access at the following Gitlab repository: https://github.com/aleksandra-chybowska/troponin_episcores. The source datasets analysed during the current study are not publicly available due to them containing information that could compromise participant consent and confidentiality. Data can be obtained from the data owners. Instructions for accessing Generation Scotland data can be found here: https://www.ed.ac.uk/generation-scotland/for-researchers/access; the 'GS Access Request Form' can be downloaded from this site. Completed request forms must be sent to access@generationscotland.org to be approved by the Generation Scotland access committee. Videos introducing the 109 protein EpiScores and MethylDetectR application are available at https://www.youtube.com/channel/UCxQrFFTIItF25YKfJTXuumQ.

References

1. Roth GA, Mensah GA, Johnson CO, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019. *Journal of the American College of Cardiology* 2020; 76: 2982–3021.

2. Agha M, Agha R. The rising prevalence of obesity: part A: impact on public health. *Int J Surg Oncol (N Y)* 2017; 2: e17.

3. Tapela N, Collister J, Clifton L, et al. Prevalence and determinants of hypertension control among almost 100 000 treated adults in the UK. *Open Heart* 2021; 8: e001461.

4. About - ASSIGN Score, https://www.assign-score.com/about/beginners/ (2022, accessed 1 June 2022).

5. GovScot, http://www.gov.scot/publications/scottish-index-multiple-deprivation-2020/ (2020, accessed 4 August 2022).

6. Welsh P, Preiss D, Hayward C, et al. Cardiac Troponin T and Troponin I in the General Population. *Circulation* 2019; 139: 2754–2764.

7. Wu AHB. Release of cardiac troponin from healthy and damaged myocardium. *Frontiers in Laboratory Medicine* 2017; 1: 144–150.

8. Michielsen ECHJ, Wodzig WKWH, Van Dieijen-Visser MP. Cardiac Troponin T Release after Prolonged Strenuous Exercise. *Sports Med* 2008; 38: 425–435.

9. Fridén V, Starnberg K, Muslimovic A, et al. Clearance of cardiac troponin T with and without kidney function. *Clinical Biochemistry* 2017; 50: 468–474.

10. Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes & Development* 2011; 25: 1010.

11. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* 2008; 9: 465–476.

12. Corley J, Cox SR, Harris SE, et al. Epigenetic signatures of smoking associate with cognitive function, brain structure, and mental and physical health outcomes in the Lothian Birth Cohort 1936. *Transl Psychiatry* 2019; 9: 248.

13. Hamilton OKL, Zhang Q, McRae AF, et al. An epigenetic score for BMI based on DNA methylation correlates with poor physical health and major disease in the Lothian Birth Cohort. *Int J Obes* 2019; 43: 1795–1802.

14. Gadd DA, Hillary RF, McCartney DL, et al. Epigenetic scores for the circulating proteome as tools for disease prediction. *eLife* 2022; 11: e71802.

15. Moldoveanu AI, Shephard RJ, Shek PN. Exercise elevates plasma levels but not gene expression of IL-1β, IL-6, and TNF-α in blood mononuclear cells. *Journal of Applied Physiology* 2000; 89: 1499–1504.

16. Koenig W, Sund M, Fröhlich M, et al. Refinement of the Association of Serum C-reactive Protein Concentration and Coronary Heart Disease Risk by Correction for Within-Subject Variation over Time: The MONICA Augsburg Studies, 1984 and 1987. *American Journal of Epidemiology* 2003; 158: 357–364.

17. Liu Y, Buil A, Collins BC, et al. Quantitative variability of 342 plasma proteins in a human twin population. *Mol Syst Biol* 2015; 11: 786.

18. Stevenson AJ, McCartney DL, Hillary RF, et al. Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. *Clinical Epigenetics* 2020; 12: 113.

19. Smith BH, Campbell A, Linksted P, et al. Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *International Journal of Epidemiology* 2013; 42: 689–700.

20. Welsh P, Preiss D, Shah ASV, et al. Comparison between High-Sensitivity Cardiac Troponin T and Cardiac Troponin I in a Large General Population Cohort. *Clinical Chemistry* 2018; 64: 1607–1616.

21. Cheng Y, Gadd DA, Gieger C, et al. Development and validation of DNA Methylation scores in two European cohorts augment 10-year risk prediction of type 2 diabetes. 2022; 2021.11.19.21266469.

22. Hillary RF, Marioni RE. MethylDetectR: a software for methylation-based health profiling. *Wellcome Open Res* 2020; 5: 283.

23. Faqs - ASSIGN Score, https://www.assign-score.com/faqs/ (2022, accessed 14 April 2022).

24. Woodward M, Brindle P, Tunstall-Pedoe H, et al. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart* 2005; 93: 172–176.

25. Estimate the risk - ASSIGN Score, https://www.assign-score.com/estimate-the-risk/visitors/ (2022, accessed 14 April 2022).

26. SCORE2 working group and ESC Cardiovascular risk collaboration. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *European Heart Journal* 2021; 42: 2439–2454.

27. SCORE2 risk calculator for low-risk regions, https://heartscore.escardio.org/Calculate/quickcalculator.aspx?model=low (2022, accessed 6 October 2022).

28. Therneau TM, Grambsch PM. *Modeling survival data: extending the Cox model*. New York: Springer, 2000.

29. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 2010; 33: 1–22.

30. Lin DY. On the Breslow estimator. *Lifetime Data Anal* 2007; 13: 471–480.

31. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011; 12: 77.

32. Cappozzo A, McCrory C, Robinson O, et al. A blood DNA methylation biomarker for predicting short-term risk of cardiovascular events. *Clinical Epigenetics* 2022; 14: 121.

33. Sharif S, Van der Graaf Y, Cramer MJ, et al. Low-grade inflammation as a risk factor for cardiovascular events and all-cause mortality in patients with type 2 diabetes. *Cardiovascular Diabetology* 2021; 20: 220.

34. Goncalves I, Bengtsson E, Colhoun HM, et al. Elevated Plasma Levels of MMP-12 Are Associated With Atherosclerotic Burden and Symptomatic Cardiovascular Disease in Subjects With Type 2 Diabetes. *Arteriosclerosis, Thrombosis, and Vascular Biology* 2015; 35: 1723–1731.

35. Gonçalves I, Oduor L, Matthes F, et al. Osteomodulin Gene Expression Is Associated With Plaque Calcification, Stability, and Fewer Cardiovascular Events in the CPIP Cohort. *Stroke* 2022; 53: e79–e84.

36. Li Y, Hiroi Y, Ngoy S, et al. Notch1 in Bone Marrow–Derived Cells Mediates Cardiac Repair After Myocardial Infarction. *Circulation* 2011; 123: 866–876.

37. Ferrari R, Rizzo P. The Notch pathway: a novel target for myocardial remodelling therapy? *European Heart Journal* 2014; 35: 2140–2145.

38. Gude NA, Emmanuel G, Wu W, et al. Activation of Notch-Mediated Protective Signaling in the Myocardium. *Circulation Research* 2008; 102: 1025–1035.

39. Husby A. On the Use of Blood Samples for Measuring DNA Methylation in Ecological Epigenetic Studies. *Integrative and Comparative Biology* 2020; 60: 1558–1566.

40. Herzog EM, Eggink AJ, Willemsen SP, et al. The tissue-specific aspect of genome-wide DNA methylation in newborn and placental tissues: implications for epigenetic epidemiologic studies. *J Dev Orig Health Dis* 2021; 12: 113–123.

41. Daubert MA, Jeremias A. The utility of troponin measurement to detect myocardial infarction: review of the current findings. *Vasc Health Risk Manag* 2010; 6: 691–699.

Figure legends

Structural graphical abstract. CVD – Cardiovascular Disease, EpiScore – Epigenetic Score, Cox PH – Cox Proportional Hazards Regression, DNAm – DNA methylation

Figure 1. The main steps of the analysis pipeline. The numbers in the green circles correspond to increases in the Area Under the Curve observed after adding the newly created predictors to the model. *$n_{troponin} \geq 6{,}750$. Created with BioRender.com

Figure 2. Associations between protein EpiScores and incident CVD. Hazard ratios are plotted for the 65 significant associations with 95% confidence interval limits.

Figure 3. Comparison of model based on age, sex, and ASSIGN (black) with the full multi-omic model (red). The full model (red) considered both the composite CVD TroponinScore (based on measured troponins) and the CVD EpiScore (based on protein EpiScores) and was adjusted for age, sex, and the ASSIGN score. The full model had slightly better discrimination when compared to the null model (AUC improved by 1.1%, ROC test P = 0.16).

Supplement

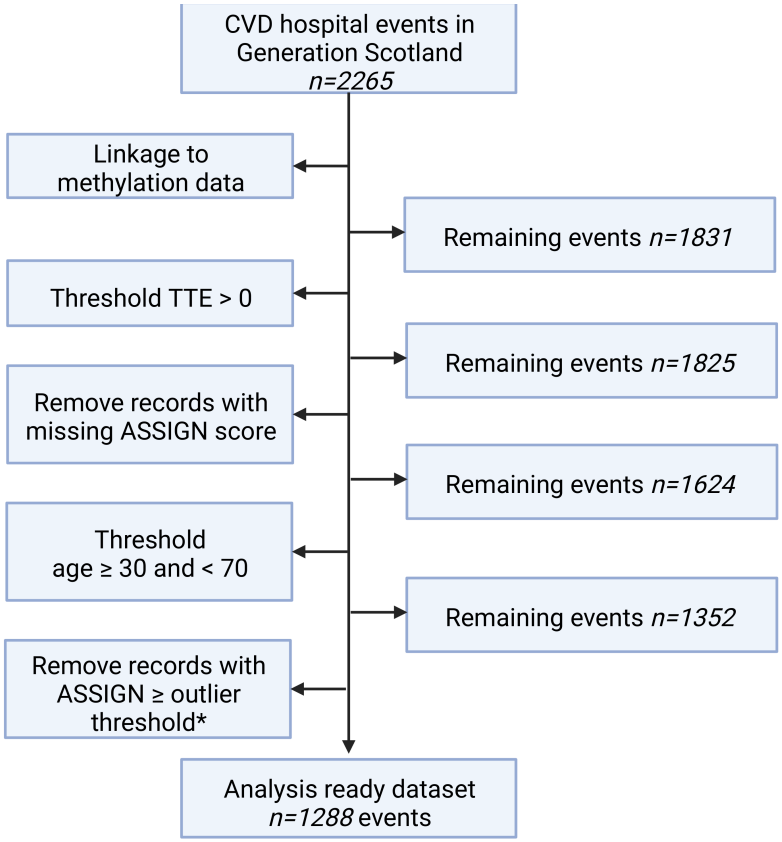


Figure 1. Pre-processing of the full dataset. This pipeline was used to prepare data for Cox PH models aimed at the identification of potential predictors of CVD. *Outlier threshold – an ASSIGN value that is more than 3 SD from the mean. Created with BioRender.com.

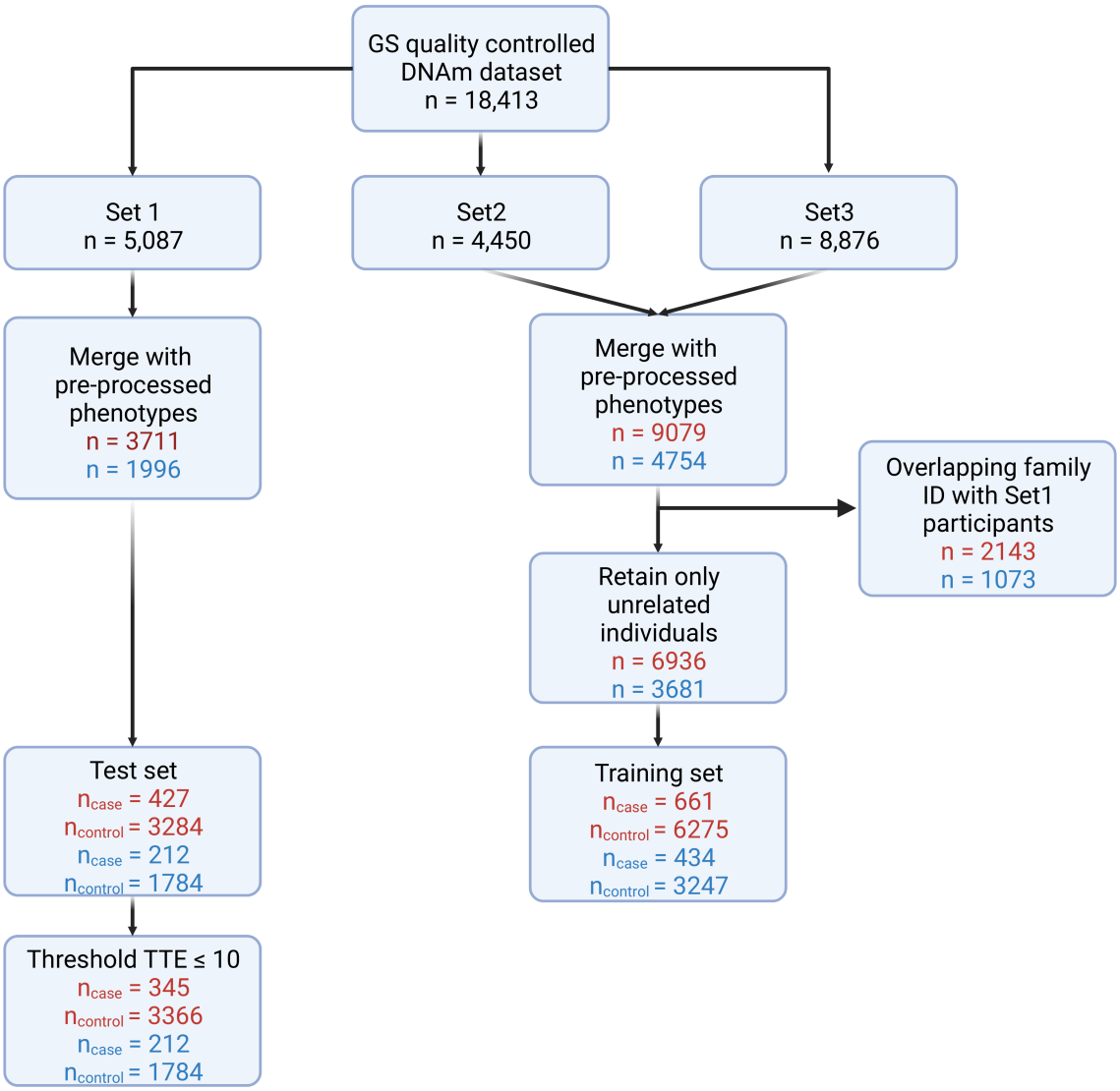


Figure 2. Pre-processing of the training and test sets. While Set 1 was the test set, the training set consisted of Sets 2 and 3 combined. Pre-processing phenotypes (ASSIGN components, composite CVD event status and time-to-event) consisted of filtering out records with time-to-event≤0, removing records with missing ASSIGN score, and thresholding age of participants to 30-70 years. To ensure that there were no individuals with overlapping family ids across the training and the test set, any individuals in the training set associated with the same family id as individuals from the test set were excluded from further analyses. The number of samples has been reported in red for datasets containing cardiac troponin and in blue for datasets that did not include it. Created with BioRender.com.

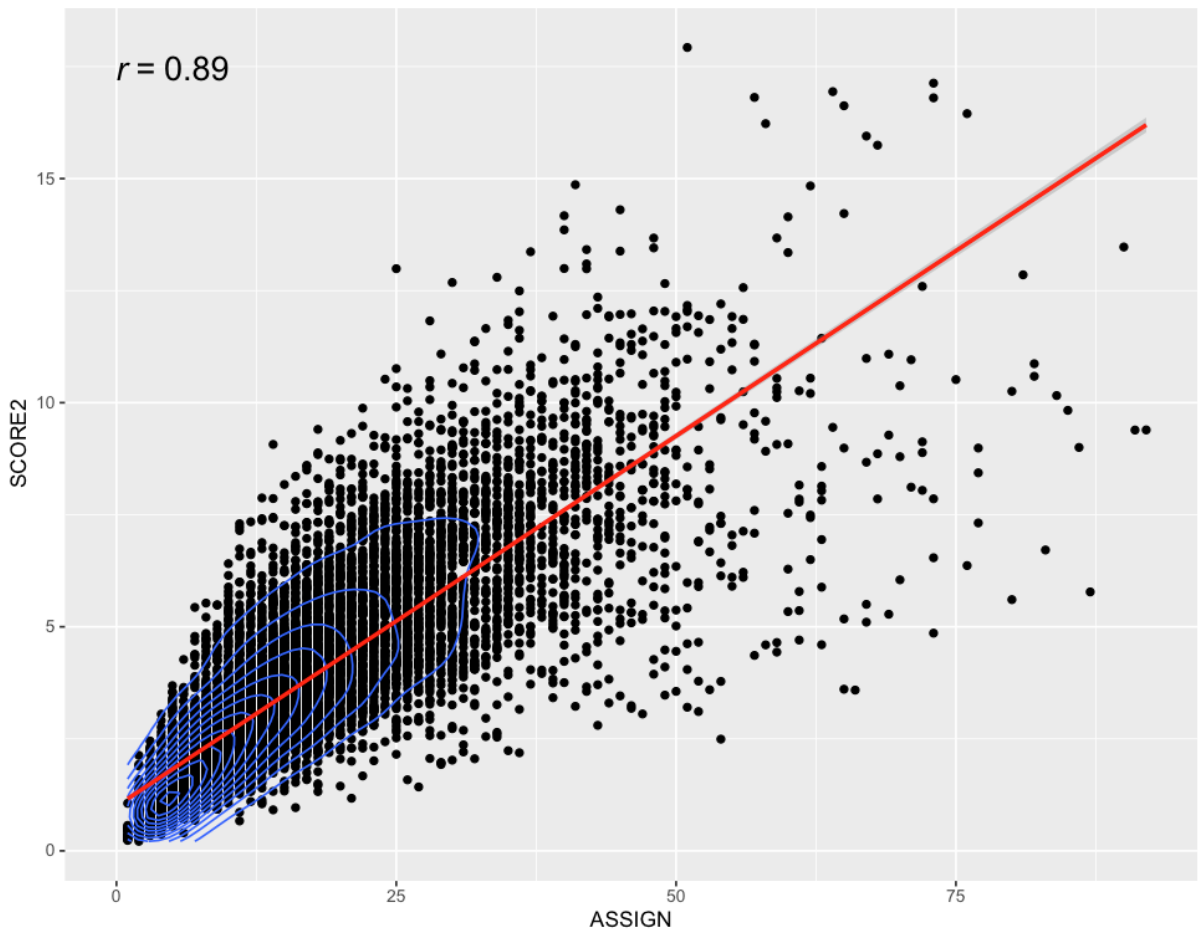


Figure 3. Strongly correlated 10-year CVD risk estimates produced by ASSIGN and SCORE2.

1 **Table 1. Hazard ratio (HR) of ASSIGN estimated by Cox proportional hazards CVD**

2 **model adjusted for relatedness using kinship matrix where the ASSIGN score is the**

3 **only predictor.**

Age category	HR_{ASSIGN}	95% CI	P_{ASSIGN}	n_{cases}	$n_{controls}$
<30	2.42	[1.43, 4.10]	9.4×10^{-4}	20	2560
30 - 40	1.87	[1.48, 2.37]	2.0×10^{-7}	64	2436
40 - 50	1.97	[1.74, 2.22]	$<2.2 \times 10^{-16}$	232	3153
50 - 60	1.67	[1.55, 1.81]	$<2.2 \times 10^{-16}$	526	3763
60 - 70	1.54	[1.42, 1.67]	$<2.2 \times 10^{-16}$	530	2267
≥70	1.27	[1.12, 1.44]	2.3×10^{-4}	252	551

4

5

6 **Table 2. Description of the 109 proteins for which Gadd et al. generated robust**

7 **EpiScores.** Table adapted from Gadd *et. al* [14].

8

Identifier (SOMAScan SeqId or Olink name)	*Protein Panel*	*Gene Name*	*Target*
4407-10	SomaScan	MST1	MSP
4435-66	SomaScan	ENPP7	ENPP7
4148-49	SomaScan	PAPPA	PAPP-A
3617-80	SomaScan	HGFAC	HGFA
3440-7	SomaScan	GZMA	granzyme A
3403-1	SomaScan	TPSB2	TPSB2
3291-30	SomaScan	FCER2	CD23
3195-50	SomaScan	GNLY	Granulysin
5124-69	SomaScan	ICAM5	sICAM-5
3216-2	SomaScan	PIGR	PIGR
5339-49	SomaScan	S100A9	calgranulin B
3175-51	SomaScan	ADAMTS13	ATS13
3311-27	SomaScan	FCGR3B	FCG3B
4500-50	SomaScan	CLEC11A	SCGF-alpha
2948-58	SomaScan	GHR	Growth hormone receptor
2950-57	SomaScan	IGFBP4	IGFBP-4
2966-65	SomaScan	CLEC11A	SCGF-beta
4337-49	SomaScan	CRP	CRP
2579-17	SomaScan	MMP9	MMP-9

4498-62	SomaScan	NCAM1	NCAM-120
3184-25	SomaScan	F7	Coagulation Factor VII
3324-51	SomaScan	LY9	LY9
5034-79	SomaScan	PRSS2	Trypsin 2
3343-1	SomaScan	ACY1	Aminoacylase-1
5358-3	SomaScan	OMD	OMD
3046-31	SomaScan	RETN	resistin
4568-17	SomaScan	SLITRK5	SLIK5
3448-13	SomaScan	INSR	IR
5028-59	SomaScan	CD163	sCD163
2658-27	SomaScan	NTRK3	TrkC
3292-75	SomaScan	CD48	CD48
4929-55	SomaScan	SHBG	SHBG
4924-32	SomaScan	MMP1	MMP-1
2580-83	SomaScan	MPO	Myeloperoxidase
4831-4	SomaScan	SELL	sL-Selectin
3152-57	SomaScan	TNFRSF1B	TNF sR-II
2967-8	SomaScan	VCAM1	VCAM-1
5000-52	SomaScan	LGALS3BP	LG3BP
3038-9	SomaScan	CXCL11	I-TAC
3169-70	SomaScan	IDUA	IDUA
3060-43	SomaScan	C9	C9
2665-26	SomaScan	TNFRSF17	BCMA
2516-57	SomaScan	CCL21	6Ckine
2780-35	SomaScan	LTF	Lactoferrin
4763-31	SomaScan	AFM	Afamin
3485-28	SomaScan	B2M	b2-Microglobulin
2771-35	SomaScan	IGFBP1	IGFBP-1
4481-34	SomaScan	C4A|C4B	C4
3505-6	SomaScan	LTA|LTB	Lymphotoxin a1/b2
3041-55	SomaScan	MRC2	MRC2
3470-1	SomaScan	SELE	sE-Selectin
3554-24	SomaScan	ADIPOQ	Adiponectin
4990-87	SomaScan	GP1BA	GP1BA
4920-10	SomaScan	LYZ	Lysozyme
2851-63	SomaScan	C5	C5a
3293-2	SomaScan	CD5L	CD5L
4496-60	SomaScan	MMP12	MMP-12
2687-2	SomaScan	MIA	MIA
3316-58	SomaScan	SERPIND1	Heparin cofactor II
3473-78	SomaScan	MPL	Thrombopoietin Receptor

2982-82	SomaScan	LGALS4	Galectin-4
3298-52	SomaScan	CNTN4	Contactin-4
3044-3	SomaScan	CCL18	PARC
5107-7	SomaScan	NOTCH1	Notch 1
4153-11	SomaScan	SERPINA3	alpha-1-antichymotrypsin complex
3805-16	SomaScan	ESM1	Endocan
3339-33	SomaScan	THBS2	TSP2
4160-49	SomaScan	MMP2	MMP-2
3079-62	SomaScan	RARRES2	TIG2
3519-3	SomaScan	CCL17	TARC
3035-80	SomaScan	IL19	IL-19
5363-51	SomaScan	SEMA3E	Semaphorin 3E
3235-50	SomaScan	WFIKKN2	WFKN2
2705-5	SomaScan	CCL25	TECK
3600-2	SomaScan	CHIT1	Chitotriosidase-1
4141-79	SomaScan	CXCL10	IP-10
3348-49	SomaScan	BMP1	BMP-1
3029-52	SomaScan	CD209	DC-SIGN
3508-78	SomaScan	CCL22	MDC
2816-50	SomaScan	BCAM	BCAM
2826-53	SomaScan	EDA	EDA
5491-12	SomaScan	SPOCK2	Testican-2
5029-3	SomaScan	FAP	SEPR
4930-21	SomaScan	STC1	Stanniocalcin-1
CRTAM	Olink Neurology	CRTAM	N/A
EZR	Olink Neurology	EZR	N/A
FcRL2	Olink Neurology	FCRL2	N/A
G.CSF	Olink Neurology	CSF3	N/A
GDF.8	Olink Neurology	MSTN	N/A
GZMA	Olink Neurology	GZMA	N/A
NEP	Olink Neurology	MME	N/A
N.CDase	Olink Neurology	ASAH2	N/A
NMNAT1	Olink Neurology	NMNAT1	N/A
NTRK3	Olink Neurology	NTRK3	N/A
SKR3	Olink Neurology	ACVRL1	N/A
SIGLEC1	Olink Neurology	SIGLEC1	N/A
SMPD1	Olink Neurology	SMPD1	N/A
CCL11	Olink Inflammatory	CCL11	N/A
CD6	Olink Inflammatory	CD6	N/A
CXCL10	Olink Inflammatory	CXCL10	N/A

CXCL11	Olink Inflammatory	CXCL11	N/A
CXCL9	Olink Inflammatory	CXCL9	N/A
EN.RAGE	Olink Inflammatory	S100A12	N/A
TGF.alpha	Olink Inflammatory	TGFA	N/A
FGF.21	Olink Inflammatory	FGF21	N/A
HGF	Olink Inflammatory	HGF	N/A
OSM	Olink Inflammatory	OSM	N/A
VEGFA	Olink Inflammatory	VEGFA	N/A
MMP.1	Olink Inflammatory	MMP1	N/A

Table 3. Predictor weights for the 52 protein EpiScores included in the composite protein EpiScore.

Variable	*Coefficient*
GZMA	-0.1271547
N.CDase	-0.0087279
NTRK3	-0.0196634
SKR3	0.03051959
CCL11	0.01233694
CXCL10	0.01221046
CXCL11	-0.0495001
EN.RAGE	-0.012591
FGF.21	0.05384112
HGF	0.06730665
MMP.1	-0.0640031
2705-5	0.04281577
2780-35	0.00792122
2816-50	0.00172923
2851-63	0.08494397
2948-58	0.04343471
2967-8	0.10683444
3029-52	-0.0013619
3038-9	-0.0165371
3044-3	0.25625949
3060-43	-0.0869837
3079-62	0.10752135
3152-57	0.01913877
3169-70	-0.0004891
3175-51	0.04843302
3216-2	-0.0036286
3291-30	0.04410383

3292-75	-0.0158129
3293-2	0.05553758
3316-58	0.00786336
3339-33	0.04545785
3348-49	0.02731157
3403-1	-0.0112186
3448-13	0.03492377
3470-1	0.0955622
3485-28	0.05696234
3505-6	-0.0303161
3519-3	-0.0651195
3600-2	0.04599604
3617-80	0.05262743
4160-49	-0.0566947
4407-10	0.04561812
4435-66	0.01197674
4481-34	-0.0561997
4496-60	0.093213
4763-31	0.04087379
4831-4	-0.0238234
4920-10	-0.0145315
4924-32	0.01811132
5000-52	0.01128698
5124-69	0.01086571
5363-51	-0.0206442

Table 4. Predictor weights for cardiac troponins included in the composite troponin score.

Variable	Coefficient
Troponin_T	0.1825032
cTnI_corrected	0.2157524

Table 5. Summary of CVD risk models based on ASSIGN. N – null model (adjusted for age + sex + ASSIGN)**,** E – CVD EpiScore, T – CVD TroponinScore,

Model	HR_{age}	P_{age}	HR_{sex}	P_{sex}	HR_{ASSIGN}	P_{ASSIGN}	HR_E	P_E	HR_T	P_T	AUC	PRAUC	C-index
N	0.95	0.64	1.06	0.68	2.07	$1.3x10^{-12}$	-	-	-	-	0.707	0.209	0.697
N+E	0.89	0.35	0.99	0.93	1.97	$1.0x10^{-10}$	1.20	0.055	-	-	0.711	0.211	0.701
N+T	0.90	0.35	0.92	0.56	2.06	$2.6x10^{-12}$	-	-	1.28	$5.9x10^{-4}$	0.715	0.226	0.705
N+E+T	0.85	0.17	0.85	0.31	1.97	$1.7x10^{-10}$	1.19	0.065	1.28	$6.8x10^{-4}$	0.718	0.228	0.708

Table 6. Summary of CVD risk models based on SCORE2. N – null model (adjusted for age + sex + ASSIGN)**,** E – CVD EpiScore, T – CVD TroponinScore

Model	HR_{age}	P_{age}	HR_{sex}	P_{sex}	HR_{SCORE2}	P_{SCORE2}	HR_E	P_E	HR_T	P_T	AUC	PRAUC	C-index
N	1.09	0.47	0.89	0.52	1.74	$4.4x10^{-6}$	-	-	-	-	0.683	0.184	0.673
N+E	1.04	0.78	0.85	0.36	1.61	$1.4x10^{-4}$	1.24	0.023	-	-	0.688	0.191	0.679
N+T	1.05	0.71	0.78	0.18	1.71	$9.6x10^{-6}$	-	-	1.28	$6.8x10^{-4}$	0.691	0.195	0.681
N+E+T	0.99	0.92	0.74	0.1	1.59	$2.2x10^{-4}$	1.23	0.025	1.28	$7.2x10^{-4}$	0.696	0.200	0.687

Supplementary Methods

EpiScores for cardiac troponin

EpiScores were trained and tested for cTnT, cTnI, and cTnI adjusted for "blank" levels, further referred to as cTnI corrected. The training set consisted of Sets 2 and 3 combined, whereas Set 1 was the test set. To minimise information from shared environments and relatedness leaking between the training and test sets, any individuals in the training set associated with the same family ID as individuals from the test set were excluded from further analyses ($n_{excluded}$= 1,073). After this step, there were 3,681 individuals in the training set and 1,996 individuals in the test set. Further details on pre-processing steps are provided in **Supplementary Figure 2.**

As part of data pre-processing, all variables were trimmed of outliers (points beyond 4 SDs of the mean). cTnT, cTnI and cTnI corrected underwent rank-based inverse normal transformations. Using linear mixed effects models, age, age^2 and sex were regressed out of all variables, with a pedigree-based kinship matrix fitted as a random effect. The subsequent model residuals were saved and entered as the outcome of elastic net regression models. Methylation levels were treated as the independent variables. Elastic net models were run using biglasso (version 1.5.0) in R (version 4.0.3). The L1-L2 mixing parameter was set to alpha=0.5, and ten-fold cross-validation was applied. Non-zero coefficients/weights from the models were extracted. To calculate EpiScores, CpG methylation values were multiplied by these coefficients and summed for each individual in the test set. The proportion of variance explained by each EpiScore was estimated by comparing R^2 estimates from the null and full linear regression models of a studied trait. Null models were adjusted for age, sex, and the studied trait whereas full models accounted for age, sex, the studied trait and the EpiScore.

43 **EpiScores for 109 circulating proteins**

44 We have previously generated 109 EpiScores for plasma protein levels [14]. Briefly, elastic net
45 penalised regression models were run with 953 possible protein levels as outcomes and up to
46 428,489 DNAm measurements from the Illumina 450k array as input features. Protein levels
47 were adjusted for genetic effects (protein quantitative trait loci – pQTLs) prior to training
48 EpiScores. Protein EpiScores were trained in two cohorts: the Lothian Birth Cohort of 1936
49 (training set contained between 725 and 875 individuals for 160 Olink inflammatory and
50 neurology proteins) and the German cohort KORA (944 individuals, 793 SOMAscan proteins).
51 109 scores (84 trained in KORA and 25 trained in the Lothian Birth Cohort 1936) explained
52 between 1% and 58% of the variance in protein levels (with $r>0.1$ and $P<0.05$) in independent
53 test cohorts. In this study, the 109 protein EpiScores were projected into Generation Scotland
54 (n = 18,413) through the publicly available MethylDetectR Shiny App, which houses EpiScore
55 weights for the 109 scores [22].

56
57

58

59