THE UNIVERSITY *of* EDINBURGH

THE UNIVERSITY *of* EDINBURGH

Augmenting clinical risk prediction of cardiovascular disease through multi-omics

Aleksandra Daria Chybowska

Doctorate of Philosophy in Precision Medicine

The University of Edinburgh - 2025

Abstract

Cardiovascular disease (CVD) remains the leading cause of death worldwide, accounting for nearly one-third of all fatalities. Despite advances in understanding its risk factors and expanding treatment options, CVD prevalence continues to rise in many regions, imposing a substantial public health and economic burden. The disease often develops silently over decades, remaining asymptomatic until a sudden, life-threatening event occurs. Therefore, it is important to identify individuals at risk as early as possible to enable effective preventive interventions. This thesis investigates the use of multi-omic approaches to identify biomarkers of CVD and improve CVD risk prediction. A particular focus is placed on blood-based markers that could enable large-scale screening and personalised risk stratification.

Accurate prediction of CVD risk requires reliable quantification of its risk factors. Among these, smoking remains particularly difficult to measure. Self-reported data are prone to bias and fail to capture passive exposure, while clinical biomarkers such as cotinine have a short half-life and cannot reflect the long-term patterns needed to identify former smokers. In this thesis, I use DNA methylation (DNAm), a type of epigenetic modification that alters DNA without changing its sequence, to develop a long-term biomarker of smoking. I further investigate the genetic architecture of smoking and integrate DNAm data from multiple tissues and cohorts to gain deeper insights into the biological effects of tobacco use.

CVD risk is commonly assessed using clinical calculators such as ASsessing cardiovascular risk using SIGN guidelines (ASSIGN; recommended in Scotland) and Systematic Coronary Risk Evaluation 2 (SCORE2; recommended across Europe). Recent studies suggest that proteomic biomarkers can improve risk prediction beyond traditional factors included in these tools. For some proteins, DNAm-based proxies - known as protein EpiScores - can capture stable, long-term biological signals. However, the associations between protein EpiScores and CVD have not yet been systematically examined. In this thesis, I investigate associations between 109 protein EpiScores in a large cohort of more than 12,000 individuals, and evaluate whether these proxies can improve the predictive performance of ASSIGN and SCORE2.

Finally, while the number of existing proteomic biomarkers of CVD is still limited, large-scale discovery in human cohorts has been constrained by the high cost and low throughput of untargeted methods. Most cohort studies to date have relied on affinity-based platforms that measure only pre-defined subsets of proteins, restricting opportunities to identify novel markers. In this thesis, I use a novel, cost-effective mass spectrometry approach and data from 8,343 Generation Scotland (GS) individuals to study the relationships between the abundances of 439 proteins and CVD. In addition to describing potential biomarkers of early disease, I develop a proteomic risk score that integrates information from multiple protein concentrations into a single measure, with the aim of improving prediction of CVD over established risk factors.

Chapters 1 and **2** form the introduction to this thesis. In **Chapter 1**, I provide an overview of CVD, discussing its epidemiology and prevention strategies, including the use of clinical risk scores. In **Chapter 2**, I introduce the blood-based multi-omic markers investigated in this work - DNA, DNAm, and proteins. The introduction concludes in **Chapter 3**, where I set out the overarching aims of the thesis.

In **Chapter 4**, I describe the population-based cohort studies underpinning the analyses and outline the key methodological approaches used to examine the relationships between omic biomarkers and CVD.

In **Chapter 5**, I present a multi-cohort, multi-array, and multi-tissue Epigenome-Wide Association Study (EWAS) and prediction analysis of smoking. I first conduct a Bayesian EWAS of smoking pack years in GS (n=17,865; ~850k sites, Illumina EPIC array) and extend this by analysing whole-genome methylation data in smokers and non-smokers from the same cohort (n=46; ~4-21 million sites, TWIST Bioscience targeted sequencing and Oxford Nanopore long read sequencing). Next, I develop mCigarette, an epigenetic biomarker of smoking, and evaluate its performance in two independent British cohorts: Lothian Birth Cohort 1936 (LBC1936) and Avon Longitudinal Study of Parents and Children (ALSPAC). Complementary EWAS analyses of brain and blood (n_{brain}=14, n_{blood}=882; >450k sites, Illumina arrays) reveal several loci with near-perfect discrimination of smoking status, though these do not overlap between tissues. Finally, I perform a Genome-Wide Association Study (GWAS) of the epigenetic smoking phenotype, identifying multiple smoking-related loci. Together, these findings improve the accuracy of smoking-related biomarkers and provide new insights into the biological effects of smoking.

In **Chapter 6**, I use Cox proportional hazards regression to examine associations between incident CVD and 110 DNAm proxies for protein concentrations called protein EpiScores (109 externally generated by Gadd *et al*. [1], plus a newly derived score for cardiac troponin I, cTnI) in ≥12,657 GS participants (n_{cases}≥1,274; $n_{controls}$≥11,383). Sixty-five individual protein EpiScores are significantly associated with incident CVD, independent of ASSIGN and measured cTnI, with the strongest signals for proteins involved in metabolism, immune response, and tissue regeneration. I also train a composite CVD EpiScore (n=6,880 individuals, 45 proteins) and evaluate it in an independent test set (n=3,659). The protein EpiScore modestly improves 10-year CVD risk prediction beyond traditional factors and cTnI (Hazard Ratio (HR)=1.32, $P=3.7x10^{-3}$; 0.3% increase in C-statistic).

In **Chapter 7**, I conduct an untargeted mass spectrometry-based analysis of the serum proteome in relation to incident CVD and all-cause mortality. Serum abundances of 439 highly expressed proteins and protein groups are quantified in 8,343 GS participants free of CVD at baseline, with follow-up of up to 17 years ($n_{composite\_CVD}$=666; $n_{all\text{-}cause\_death}$=618). I use Cox proportional hazards models to examine associations between individual proteins and incident outcomes before and after adjustment for known CVD risk factors, with sex-specific effects explored. Forty-eight proteins are significantly associated with incident CVD and death, including 24 previously unreported associations ($P_{Bonferroni}<1.14x10^{-4}$). Notably, proteins involved in immune and oxidative stress responses, lipid metabolism, and the complement cascade show outcome-specific associations. Finally, I develop a protein-based risk score for composite CVD and test in an independent subset, improving 17-year CVD prediction beyond age, sex, and nine lifestyle and clinical risk factors (increase in area under the Receiver Operating Characteristic (ROC) curve of 0.010, ROC P=0.013).

In **Chapter 8**, I synthesise the findings from **Chapters 5-7**, discuss the limitations of these studies, and outline directions for future research.

This thesis demonstrates that DNAm-based proxies for smoking and protein levels can serve as biomarkers of CVD. By expanding the repertoire of potential proteomic markers, this work enhances the ability to detect early changes associated with CVD development. Moreover, multi-omic markers described here provide novel insights into the molecular pathways underlying CVD.

Lay Summary

Cardiovascular disease (CVD), which affects the heart and arteries, is the leading cause of death worldwide. It often develops silently over many years and can remain unnoticed until a sudden, life-threatening event occurs. Detecting people at risk early is crucial to prevent serious complications. This thesis explores how blood-based molecular markers can help identify individuals at risk of CVD.

One important risk factor for CVD is smoking. Traditional ways of measuring smoking, such as self-reported questionnaires or substances produced when the body breaks down nicotine, can be unreliable or only reflect short-term exposure. In this thesis, I develop a new biomarker called mCigarette, based on chemical changes to DNA known as DNA methylation (DNAm), which can capture both long-term and second-hand exposure to tobacco. I validate this biomarker across multiple large British datasets and also examine DNAm changes in different tissues to better understand how smoking affects the body.

In addition to smoking, protein levels in the blood are linked to CVD risk. However, levels of certain proteins can change rapidly in response to diet, infection, or other temporary factors, making them less reliable for long-term risk prediction. I investigate DNAm–based proxies for 110 blood proteins in more than 12,000 Europeans, showing that many of these “protein EpiScores” are associated with future CVD risk beyond traditional clinical tools. I also combine 45 of these protein markers into a single score that improves prediction of CVD over standard risk calculators.

Finally, I use data of 439 highly abundant blood proteins measured directly using a cost-effective mass spectrometry approach in over 8,000 Europeans. I identify 48 proteins linked to future CVD and death, and develop a protein-based risk score that improves prediction of heart disease beyond age, sex, and traditional risk factors.

Together, this work shows that DNAm and protein-based markers found in the blood can help identify people at risk of CVD before the disease becomes serious. Results presented here expand the list of potential biomarkers for early detection, improve our ability to predict heart disease, and provide new insights into the biological processes that drive CVD.

Declaration of Originality

I declare that the work presented in this thesis is my own, except where work which has formed part of jointly authored publications has been included. My contributions and those of other authors to this work are indicated below. Where reference was made within this thesis to the work of others, appropriate credit was given. This thesis has not been submitted, in whole or in part, in any previous application for a degree.

The majority of data used in the analyses presented in this document were pre-processed by the Generation Scotland and Lothian Birth Cohort teams. This included genomic, epigenomic, and proteomic data, alongside all demographic and clinical variables. I was responsible for processing and analysing all next generation sequencing data available in the Generation Scotland cohort (TWIST Bioscience and Oxford Nanopore Technologies samples). Additionally, I calculated CVD risk scores (ASsessing cardiovascular risk using SIGN guidelines to assign preventive treatment and Systematic Coronary Risk Evaluation 2) in Generation Scotland and derived the pack years variable in the Lothian Birth Cohort. I was also involved in pre-processing the mass spectrometry data available in Generation Scotland. The code used to accomplish these tasks can be found at:

https://github.com/aleksandra-chybowska?tab=repositories.

The work presented in **Chapter 5** was previously published in *Nature Communications* as “Blood- and brain-based genome-wide association studies of smoking” by Aleksandra D. Chybowska (candidate), Elena Bernabeu, Paul Yousefi, Matthew Suderman, Robert F. Hillary, Louise MacGillivray, Lee Murphy, Sarah E. Harris, Janie Corley, Archie Campbell, Tara L. Spires-Jones, Daniel L. McCartney, Simon R. Cox, Jackie F. Price (PhD supervisor), Kathryn L. Evans (PhD supervisor), Riccardo E. Marioni (primary PhD supervisor). Author contributions: A.D.C. analysed the data. E.B. and R.F.H developed the Bayesian EWAS pipeline. P.Y. and M.S. replicated results in the ALSPAC cohort. D.L.M., R.F.H., L.McG., L.M., S.E.H., J.C., A.C., T.L.S, S.R.C., and K.L.E. were involved in the data generation. E.B., R.E.M., and C.A.V. drafted the initial manuscript. A.D.C., J.F.P., K.L.E., and R.E.M. designed the study. All authors read and approved the final manuscript. The work was also presented at the European Society of Human Genetics Conference 2023 (poster and oral presentation) as well as the Clinical Epigenetics International Conference 2024 (poster presentation).

The work presented in **Chapter 6** was previously published in Circulation: Genomic and Precision Medicine as “Epigenetic Contributions to Clinical Risk Prediction of Cardiovascular Disease” by Aleksandra D. Chybowska (candidate), Danni A. Gadd, Yipeng Cheng, Elena Bernabeu, Archie Campbell, Rosie M. Walker, Andrew M. McIntosh, Nicola Wrobel, Lee Murphy, Paul Welsh, Naveed Sattar, Jackie F. Price (PhD supervisor), Daniel L. McCartney, Kathryn L. Evans (PhD supervisor) and Riccardo E. Marioni (primary PhD supervisor). Author contributions: A.D.C. and R.E.M. 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. A.D.C and R.E.M. 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. The work was also presented at the Clinical Epigenetics International Conference 2022 (poster presentation) as well as at the UK Molecular Epidemiology Group Conference 2022 (oral presentation).

The work presented in **Chapter 7** has been checked by all co-authors and submitted to medRxiv, the corresponding manuscript is “Untargeted Proteomic Profiling Identifies Candidate Biomarkers for Early Detection of Cardiovascular Disease and Mortality” by Aleksandra D. Chybowska (candidate), Spyros Vernardis Daniel L. McCartney, Jure Mur, Josephine Robertson, Hannah M. Smith, Archie Campbell, Camilla Drake, Hannah Grant, Poppy Adkin, Matthew White, Christoph B. Messner, Arturas Grauslys, Sergej Andrejev, Charles Brigden, David J. Porteous, Caroline Hayward, Jackie F. Price (PhD supervisor), Kathryn L. Evans (PhD supervisor), Aleksej Zelezniak, Markus Ralser, Riccardo E. Marioni (PhD supervisor). Author contributions: R.E.M., and A.D.C. were responsible for the conception and design of the study. R.E.M. and A.D.C. drafted the article. A.D.C and D.L.M. carried out the data analyses. J.M., J.R, H.M.S contributed to the analyses and methodology. C.D. and H.G. were involved in investigation; sample preparation, collating approximately 10000 samples and transferring into required format for mass spectrometry analysis, QC checks and helping coordinate shipping. P.A., M.W., C.B.M, A.G., S.A., C.B., A.Z. and M.R. were responsible for mass spectrometry analysis. A.C. facilitated data linkage. C.H., M.R., A.Z, K.L.E. and J.F.P. were involved in conceptualisation and provided consultation on the methodology. D.J.P contributed to data collection and preparation. All authors read and approved the final manuscript. The work was presented as part of the CGEM work in progress talks (oral presentation).

Aleksandra Chybowska, 1st November 2025

Acknowledgements

I am deeply grateful to my supervisor, Prof. Riccardo Marioni, whose steady support, thoughtful guidance, and genuine encouragement have shaped every stage of this journey. This thesis would not exist without his mentorship. I am also grateful to my co-supervisors: Dr Kathy Evans for her insight and kindness throughout my studies, and Prof Jackie Price for her invaluable support with the clinical aspects of my work and for her feedback that continually strengthened my research. I feel truly fortunate to have been guided by such a supportive supervisory team.

My sincere gratitude goes to the participants of Generation Scotland, the Lothian Birth Cohort, and the Avon Longitudinal Study of Parents and Children, whose contributions made this work possible. I am also thankful to my funding body, the Medical Research Council, for supporting me throughout my studentship. I would also like to thank my thesis committee chair, Prof Samantha Lycett, for her constructive comments and suggestions during our committee meetings. Finally, I would like to express my appreciation to Precision Medicine DTP team, especially Prof. Susan Farrington and Susan Mitchell, who were there when I needed a helping hand.

My heartfelt thanks to Dr Daniel McCartney for his constant support with every methods-related challenge and for all his help with my first paper. I am also grateful to Dr Yipeng Cheng for always looking out for me and to Dr Jure Mur for his honest feedback. Thank you to Josie for being the clinician we all need, and to Hannah for always cracking a great joke just when I need it most. Finally, thank you to all the other staff and students I had the privilege of working with for making this PhD such an enjoyable experience.

To Bart, thank you for *always* being by my side. Thank you for looking after Roza, cooking for me, and for being with me through every difficult moment and every moment of joy -- I could not have written this thesis without you. To my friends, Ola and Aneta, thank you for putting up with me throughout the years. Thank you to my parents for their unending support — I am truly privileged to be part of your family. Finally, to everyone who reads this thesis, I hope you find this research useful.

Funding statement:

A.D. Chybowska was supported by a Medical Research Council PhD Studentship in Precision Medicine with funding from the Medical Research Council Doctoral Training Program and the University of Edinburgh College of Medicine and Veterinary Medicine.

Table of Contents

Abbreviations

5hmC – 5-Hydroxymethylcytosine

5mC – 5-Methylcytosine

ACE – Angiotensin-Converting Enzyme

ADH – Antidiuretic Hormone

ALSPAC – Avon Longitudinal Study of Parents and Children

ARIC – Atherosclerosis Risk in Communities

ARIES – Accessible Resource for Integrated Epigenomics Studies

ASSIGN – Scottish cardiovascular risk score (ASsessing cardiovascular risk using SIGN guidelines)

AUC – Area Under the Curve

BMI – Body Mass Index

C-index – Concordance Index

CAD – Coronary Artery Disease

CHARGE – Cohorts for Heart and Aging Research in Genomic Epidemiology

CHD – Coronary Heart Disease

CKB – China Kadoorie Biobank

CNV – Copy Number Variant

CpG – Cytosine–Phosphate–Guanine dinucleotide

CRP – C-Reactive Protein

CV – Cardiovascular

CVD – Cardiovascular Disease

DALY – Disability-Adjusted Life Year

DIA – Data-Independent Acquisition

DNAm – DNA Methylation

ECG – Electrocardiogram

ELISA – Enzyme-Linked Immunosorbent Assay

ESC – European Society of Cardiology

EWAS – Epigenome-Wide Association Study

F17 – ALSPAC F17 (offspring assessed at ages 15–17 years)

F24 – ALSPAC F24 (offspring assessed at ages 15–17 years)

FDR – False Discovery Rate

FHS – Framingham Heart Study

FOF – ALSPAC Focus on Fathers

FOM – ALSPAC Focus on Mothers

FPG – Fasting Plasma Glucose

GENOA – Genetic Epidemiology Network of Arteriopathy

GP – General Practitioner

GPT – Generative Pre-trained Transformer

GS – Generation Scotland

GWAS – Genome-Wide Association Study

HDL – High-Density Lipoprotein

HF – Heart Failure

HOMAGE – Heart OMics in AGEing

HRC – Haplotype Reference Consortium

ICD-10 – International Classification of Diseases, 10th Revision

IHD – Ischaemic Heart Disease

IL6 – Interleukin-6

KORA – Cooperative Health Research in the Region of Augsburg (Germany)

LBC1936 – Lothian Birth Cohort 1936

LC – Liquid Chromatography

LD – Linkage Disequilibrium

LDL – Low-Density Lipoprotein

m/z – Mass-to-Charge Ratio

MAJA – Methylation Array Joint Analysis

MI – Myocardial Infarction

ML – Machine Learning

MR – Mendelian Randomization

MRI – Magnetic Resonance Imaging

MS – Mass Spectrometry

NGS – Next-Generation Sequencing

NHS – National Health Service

NICE – National Institute for Health and Care Excellence

NSTEMI – Non-ST-Segment Elevation Myocardial Infarction

ONT – Oxford Nanopore Technologies

PEA – Proximity Extension Assay

PG – Protein Group

PH – Proportional Hazards (Cox PH model)

PIP – Posterior Inclusion Probability

QC – Quality Control

RA – Rheumatoid Arthritis

RCT – Randomized Controlled Trial

REDCap – Research Electronic Data Capture

ROC – Receiver Operating Characteristic

SCORE2 – Systematic Coronary Risk Evaluation 2 (European CVD risk algorithm)

SCORE2-OP – SCORE2-Older Persons

SES – Socioeconomic Status

SHHEC – Scottish Heart Health Extended Cohort

SHS – Scottish Health Survey

SIGN – Scottish Intercollegiate Guidelines Network

SIMD – Scottish Index of Multiple Deprivation

SMuRFs – Standard Modifiable Risk Factors

SNP – Single Nucleotide Polymorphism

SNV – Single Nucleotide Variants

SOMAmers – Slow Off-Rate Modified Aptamers

SPRINT – Systolic Blood Pressure Intervention Trial

STEMI – ST-Segment Elevation Myocardial Infarction

STRADL – Stratifying Resilience and Depression Longitudinally

SV – Structural Variant

TIA – Transient Ischaemic Attack

UKB-PPP – UK Biobank Proteomics Platform Project

WBC – White Blood Cell

WMH – White Matter Hyperintensity

List of Equations

List of Figures

List of Tables

1. Cardiovascular Disease: Mechanisms, Epidemiology, and Prevention

Cardiovascular disease (CVD) remains one of the most common and life-threatening health conditions worldwide [2,3]. It encompasses a range of disorders affecting the heart and arteries. Although first vascular changes (such as fatty streaks, see **Section 1.2**) associated with the development of atherosclerosis (the primary pathological process in many forms of CVD) emerge during childhood, the disease typically progresses silently over decades, and may remain asymptomatic until it manifests as a sudden, life-threatening event. While there currently is not a cure for atherosclerosis that would allow for its complete reversal, the existence of a prolonged, subclinical phase highlights the importance of early detection and risk stratification.

In this thesis, I focus on the development of novel omics-based biomarkers, with CVD chosen as the primary application area due to its high incidence and profound impact on both public health and the global economy. The following chapters begin a brief overview of CVD and its subtypes, followed by a summary of the condition's epidemiology and a description of commonly used clinical risk prediction scores. This sets the stage for a more detailed discussion of the role of multi-omics in CVD risk prediction.

1.1. What is CVD?

CVD is defined as a group of conditions characterised by impaired blood flow, structural abnormalities, or dysfunction within the cardiovascular (CV) system [4]. According to the 10th Revision of the International Classification of Diseases (ICD-10) – a globally recognised system for coding diseases and medical conditions – all conditions categorised under codes I00-I99 are classified as CVD [5]. This categorisation is presented below, with only the subcategories relevant to this thesis listed and highlighted in bold; in some cases, only specific codes within a subcategory were analysed (see **Sections 1.3** – **1.5**):

- I00-I02 Acute rheumatic fever
- I05-I09 Chronic rheumatic heart diseases
- **I10-I15** Hypertensive diseases
    - **I11** Hypertensive heart disease
    - **I13** Hypertensive heart and renal disease
- **I20-I25** Ischaemic heart diseases
    - **I20** Angina pectoris
    - **I21** Acute myocardial infarction
    - **I22** Subsequent myocardial infarction
    - **I23** Certain current complications following acute myocardial infarction
    - **I24** Other acute ischaemic heart diseases
    - **I25** Chronic ischaemic heart disease
- I26-I28 Pulmonary heart disease and diseases of pulmonary circulation
- **I30-I52** Other forms of heart disease
    - **I50** Heart failure
- **I60-I69** Cerebrovascular diseases
    - **I63** Cerebral infarction
    - **I65** Occlusion and stenosis of precerebral arteries, not resulting in cerebral infarction
    - **I66** Occlusion and stenosis of cerebral arteries, not resulting in cerebral infarction
    - **I69** Sequelae of cerebrovascular disease
- I70-I79 Diseases of arteries, arterioles and capillaries
- I80-I89 Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified
- I95-I99 Other and unspecified disorders of the circulatory system

While individual CV conditions have distinct risk profiles, clinical presentations, and treatment strategies, they often share underlying pathophysiological mechanisms. Atherosclerosis, for instance, is commonly the pathogenic mechanism underlying ischaemic heart disease (IHD) – including coronary heart disease (CHD) and its manifestations such as angina pectoris and myocardial infarction (MI) [6] – and also underlies ischaemic stroke, transient ischaemic attack (TIA) [7,8], peripheral artery disease [9], and certain forms of heart failure (HF) [4].

Clinical CVD risk prediction tools (see **Section 1.7**) typically concentrate on:

- CHD
- TIA and Ischaemic stroke
- CVD death (including HF)

Therefore, I decided to focus on these conditions in this thesis. Other major CVDs – such as cardiomyopathies and arrhythmias – primarily result from structural or electrical disturbances and involve fundamentally different biological pathways [10]. As a result, these non-atherosclerotic conditions require separate predictive models that reflect their unique pathophysiology to support more effective prevention and clinical management.

1.2. Atherosclerosis

Atherosclerosis is defined as the accumulation of fatty and fibrous material called a plaque within the innermost layer of the arteries [11]. The term derives from the Greek words *athero* ("gruel" or "paste") and *sclerosis* ("hardness"), reflecting the plaque's evolution from soft lipid deposits to hardened, calcified structures [12]. Atherosclerosis most commonly affects the aorta, coronary arteries, and brain arteries [11]. It develops when an excess of low-density lipoprotein (LDL) in the blood triggers inflammation of the arterial walls (see **Figure 1**) [11].

Atherosclerotic plaques can be divided into two groups: stable and unstable [12]. Stable plaques are characterised by a thick fibrous cap and gradual lipid accumulation, leading to a progressive narrowing of the arterial lumen. This restriction reduces blood flow and oxygen delivery to critical organs, resulting in ischemia. In contrast, unstable plaques possess thin fibrous caps and large lipid cores, making them prone to rupture. When an unstable plaque ruptures, it can trigger the formation of a blood clot at the site of injury [12]. Such clots may further obstruct the artery, leading to life-threatening events [6].

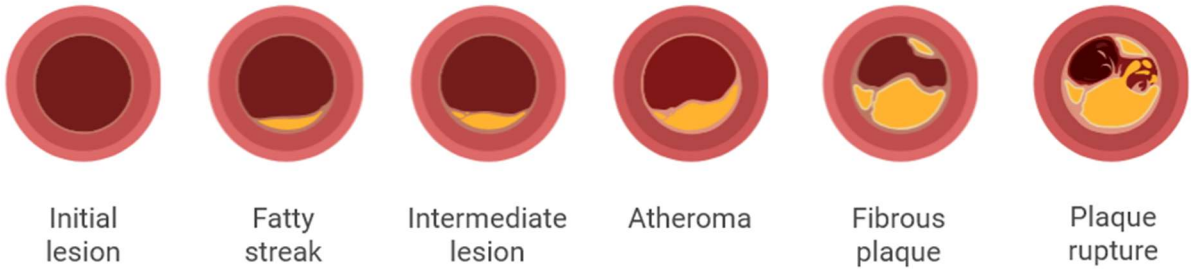


Figure 1. Progression of atherosclerosis. Atherosclerosis begins with endothelial dysfunction caused by factors such as disturbed flow, hypertension, dyslipidaemia, smoking, or diabetes [13]. This dysfunction increases endothelial permeability, allowing low-density lipoprotein (LDL) particles to enter the arterial wall [14], become oxidised [15], and trigger an inflammatory response [15]. Over time, initial lesions progress to fatty streaks and intermediate lesions. As disease develops, an atheroma forms – a structured plaque of lipids, inflammatory cells, and cellular debris that can weaken the vessel wall [16]. A fibrous cap eventually develops over the plaque (fibrous plaque) [12]. In severe cases, plaque rupture can occur, leading to thrombosis and potentially blocking blood flow [11]. Created with BioRender.com

Despite significant advances in pharmacological and interventional therapies, atherosclerotic plaques cannot be fully reversed [17,18]. Current treatments can slow disease progression, stabilise existing plaques, and lower the risk of acute CV events; however, they do not eliminate the underlying disease process. This highlights an urgent need for a deeper and more comprehensive understanding of the molecular and cellular mechanisms that drive atherogenesis.

1.3. Ischaemic Heart Disease

Studied ICD10 codes: I21, I22, I23, I24.1, I25.0, I25.1, I25.2, I25.3, I25.4, I25.5, I25.6, I25.8, I25.9

IHD refers to all conditions in which reduced blood flow limits oxygen supply to the heart muscle [19]. A major subtype of IHD is CHD, which encompasses the clinical manifestations of coronary artery disease (CAD) [6]. CAD arises when atherosclerotic plaque gradually narrows the coronary arteries, often developing silently over many years. When blood flow becomes sufficiently restricted or a plaque ruptures, CHD presents clinically, most commonly as stable angina or as acute coronary syndromes, which include unstable angina and MI [6].

Stable angina typically arises due to the presence of an atherosclerotic plaque that restricts coronary blood flow [20]. It typically presents as chest heaviness, tightness, pressure, or shortness of breath [19], often triggered by physical exertion or emotional stress and relieved by rest or medication [21]. Symptoms can differ between sexes: women more commonly report shortness of breath, fatigue, sleep disturbances, anxiety, and digestive discomfort [22,23].

Acute coronary syndromes represent a more severe, and life-threatening progression of CHD [19]. They typically occur when a plaque ruptures or fissures [19]. This triggers activation of the coagulation cascade, forming a clot that acutely obstructs the coronary artery. Acute coronary syndromes encompass the following conditions:

- Unstable Angina [19]: New or worsening chest pain without myocardial injury, considered a pre-MI state.
- MI [24]:
    - Non-ST-Segment Elevation Myocardial Infarction (NSTEMI): An MI without ST-segment elevation, diagnosed by elevated cardiac biomarkers.
    - ST-Segment Elevation Myocardial Infarction (STEMI): A severe form of MI caused by complete obstruction of a major coronary artery, leading to ST-segment elevation on an electrocardiogram (ECG) and significant myocardial damage

While unstable angina, NSTEMI, and STEMI share overlapping symptoms, severity and duration help differentiate them [19]. Pain is usually crushing or tight, may radiate to the left arm, neck, or jaw, and as in the case of stable angina, can present differently in men and women [22,23,25]. Timely identification and treatment are critical; mortality for STEMI is ~4.6% for patients reaching hospital, with ~30% dying before medical care [19,26,27]. Cardiac troponins are central to diagnosis [24], and high-sensitivity assays aid risk stratification [28].

1.4. Cerebrovascular Disease, Including Stroke

Studied ICD10 codes: I63.0, I63.1, I63.2, I63.3, I63.4, I63.5, I63.8, I63.9, I69.3, G45.0, G45.1, G45.2, G45.3, G45.4, G45.8, G45.9, G46.0, G46.1, G46.2, I65, I66

Cerebrovascular disease encompasses disorders affecting blood vessels that supply the brain and central nervous system [29]. The most common types are TIAs, ischaemic strokes, and haemorrhagic strokes [30].

TIAs, often called “mini-strokes,” are episodes of neurological dysfunction that last less than 24 hours (usually less than one hour), without evidence of acute infarction [8,31]. They occur when a temporary blockage briefly disrupts blood flow to the brain, spinal cord, or retina [7,8]. This blockage is usually caused by a thromboembolus – a blood clot that forms elsewhere in the body, such as the heart during atrial fibrillation, and travels through the bloodstream to the affected area [32]. While TIAs do not cause permanent brain damage, they serve as critical warning signs for more severe cerebrovascular events [7].

Strokes are characterised by neurological dysfunction lasting at least 24 hours or until death with no apparent nonvascular cause [33] . Unlike TIAs, strokes cause permanent damage to the affected brain areas. Ischaemic strokes are caused by the disruption of blood flow to the brain, spinal cord, or retina, resulting in infarction [33]. In contrast, haemorrhagic strokes occur when weakened blood vessels rupture, often due to hypertension or aneurysms (abnormal bulges in blood vessel walls that can break and cause bleeding) [4].

Stroke and TIA symptoms vary by the brain region affected and may include sudden weakness or numbness (often one-sided), speech or visual disturbances, facial drooping, loss of coordination, severe headache, confusion, or altered consciousness [34–41].

As the time window for treatment of stroke is limited, rapid diagnosis and medical intervention are critical for improving outcomes [42]. It is essential to distinguish ischaemic strokes and TIAs from haemorrhagic strokes and stroke mimics, as the use of thrombolytic agents can worsen outcomes in patients with haemorrhagic stroke due to promoting bleeding [43].

GFAP and S100B are promising blood-based biomarkers for early stroke diagnosis and for distinguishing between stroke subtypes, particularly in challenging diagnostic scenarios [44]. However, their use in routine clinical practice remains limited due to variability in diagnostic accuracy, lack of standardised protocols, and the need for further validation in larger, diverse patient populations.

1.5. Heart Failure

Studied ICD10 codes: I11.0, I13.0, I13.2, I50

HF is a combination of symptoms and signs arising from the heart's impaired ability to pump blood effectively. Rather than being limited to a single-organ disease, HF is a complex syndrome that leads to dysfunction across multiple systems, including the musculoskeletal, endothelial, pulmonary, endocrine, hepatic and renal systems [45]. Its diagnosis is further supported by elevated natriuretic peptide levels and objective evidence of pulmonary or systemic congestion [45]. HF is a major public health concern, associated with high morbidity and a mortality rate of 50% within five years of diagnosis [46].

The term "congestive heart failure" is often used, highlighting the accumulation of fluid in tissues and veins and swelling (edema) in the legs and feet [4]. Fluid may also accumulate in the lungs (pulmonary edema) [4]. HF can be classified as either chronic or acute. Chronic HF develops gradually over time and is typically caused by long-standing conditions such as hypertension or CAD. In contrast, acute HF refers to a sudden or severe deterioration in heart function, often triggered by events such as a MI, which damages the heart muscle and leads to the formation of scar tissue [4]. Unlike healthy muscle, scar tissue cannot contribute to the pumping action, resulting in a reduced volume of blood being ejected with each heartbeat.

In HF, cardiac output is often reduced – or fails to rise adequately with demand. In response, the body activates the sympathetic nervous system (which increases heart rate and strengthens heart contractions) and the renin – angiotensin - aldosterone system (a hormone system regulating blood pressure and fluid balance; see **Figure 2**) [4,47]. This activation temporarily increases cardiac output and blood pressure to preserve tissue perfusion [47]. Although these compensatory mechanisms initially support heart function, they ultimately accelerate the progression of heart failure [47].

HF symptoms across all types commonly include shortness of breath (especially during activity or when lying down), fatigue, nighttime coughing or breathlessness, and swelling in the legs or feet [48]. Diagnosis begins with a thorough review of symptoms, medical history, and physical examination [45]. Core investigations include ECG, chest X-ray, blood tests, and echocardiography. Measurement of natriuretic peptides (BNP or NT-proBNP) is particularly valuable, as elevated levels reflect cardiac strain, support the diagnosis, and help gauge disease severity [45].

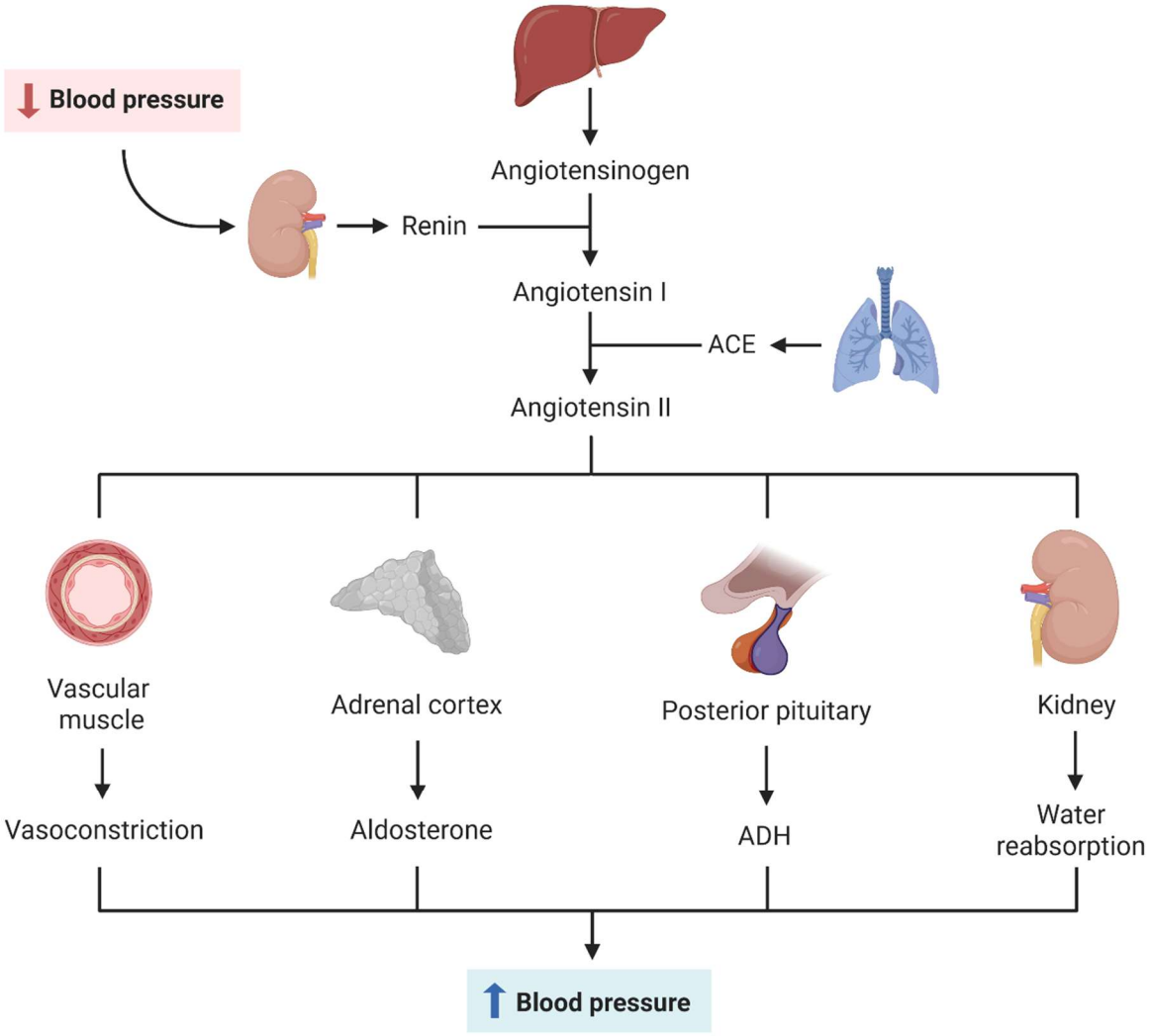


Figure 2. The renin–angiotensin–aldosterone system [47]**.** In response to reduced blood flow to the kidneys – a common consequence of heart failure – the kidneys release an enzyme called renin. Renin initiates a hormonal cascade by converting angiotensinogen, a protein produced by the liver, into angiotensin I. This is then converted into angiotensin II by angiotensin-converting enzyme (ACE), primarily in the lungs. Angiotensin II: a) constricts blood vessels, raising blood pressure, b) stimulates the adrenal cortex to release aldosterone, which causes the kidneys to retain sodium and water, c) acts on the posterior pituitary to release antidiuretic hormone (ADH) and triggers thirst, both of which further promote water retention. Collectively, these responses increase blood volume and pressure, helping to temporarily maintain organ perfusion. However, chronic activation of this system in heart failure can lead to fluid overload and further stress on the heart. Created with BioRender.com.

1.6. Epidemiology of CVD

1.6.1. Mortality and Morbidity

For decades, CVD has been the leading cause of global mortality, accounting for nearly one-third of all deaths in 2021, with approximately 20 million lives lost [3,49] . While age-standardised CVD death rates have generally declined since the mid-20th century, this progress is now stalling, and, in some regions, reversing [50].

According to the World Heart Federation's 2023 report, the global decline in CVD mortality has begun to plateau particularly in low- and middle-income countries, where over 80% of CVD deaths occur [51]. This stagnation is attributed to factors such as inadequate healthcare infrastructure, rising prevalence of risk factors (discussed below), and limited access to preventive and therapeutic interventions.

In high-income countries, the rate of decline in CVD mortality has slowed considerably. For instance, the UK recorded only an 11% reduction in premature CVD deaths between 2012 and 2019 – a sharp decline in progress compared to the 33% decrease achieved between 2005 and 2012 [52]. Since 2019, this downward trend has reversed, with CVD mortality steadily rising for the first time in nearly 60 years [53]. In Scotland alone, 10,792 deaths in 2023 were attributed to either IHD or cerebrovascular disease as the underlying cause [54]. Contributing factors include an increasingly unhealthy population, widening health inequalities, and pressures on the National Health Service (NHS) [53].

.

While mortality is a critical outcome, it does not fully reflect the worldwide burden of CVD. A more comprehensive metric is the disability-adjusted life year (DALY), which combines years of life lost due to premature death with years lived with disability [55]. A single DALY corresponds to the loss of one year of life lived in full health. **Figure 3** illustrates the global distribution of DALYs attributed to CVD per 100,000 population. Data from 2021 [56] were adjusted to a standard age distribution, ensuring that observed differences in disease burden reflect true differences in health outcomes rather than differences in population age structure. This process is known as age-standardisation.

The map reveals stark regional disparities, with particularly high burdens observed across Eastern Europe, Central Asia, and parts of Africa. In contrast, many high-income regions, including Western Europe, North America, and Oceania, exhibit substantially lower CVD-related DALY rates. This distribution underscores the growing toll of chronic, non-fatal CVDs, especially in settings where healthcare access and preventive strategies are limited.

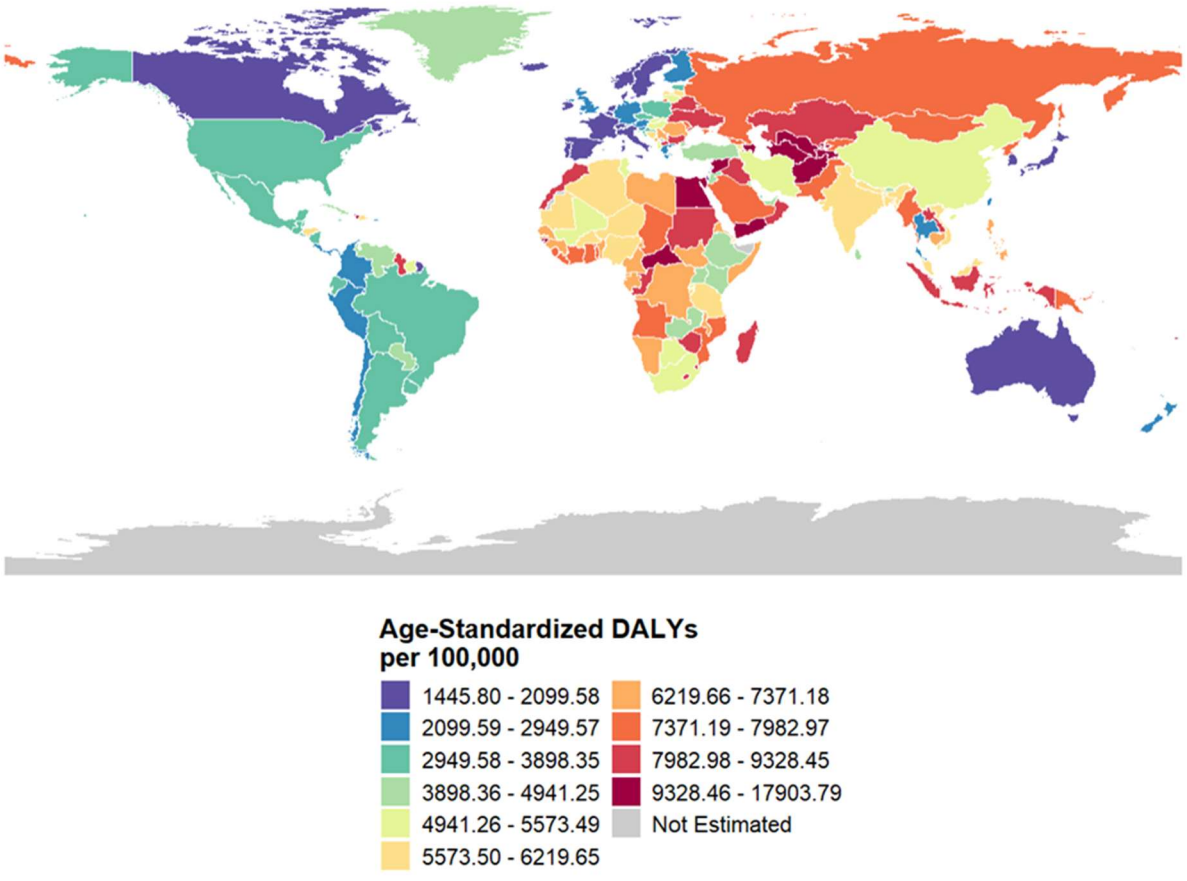


Figure 3. Global burden of CVD. The map shows the geographic distribution of age-standardised disability-adjusted life years (DALYs) per 100,000 population in 2021. DALYs represent the total burden of disease by combining years of life lost due to premature mortality and years lived with disability. Data were sourced from Global Burden of Cardiovascular Diseases and Risks Collaboration, 1990-2021 [56].

1.6.2. CVD Deaths by Cause

Among CVDs, IHD is the leading cause of death worldwide, responsible for approximately nine million deaths annually (**Figure 4**) [56]. The second major cause is stroke, which accounts for around six to seven million deaths per year, with ischaemic stroke being more common than haemorrhagic forms. Hypertensive heart disease (defined here as symptomatic HF due to the direct and long-term effects of hypertension) also contributes significantly to CVD mortality, particularly among older adults [56]. In low-income regions, rheumatic heart disease – a preventable condition stemming from untreated streptococcal infections – remains a major cause of CVD – related deaths [56]. Together, these conditions form the core of the global CV mortality burden.

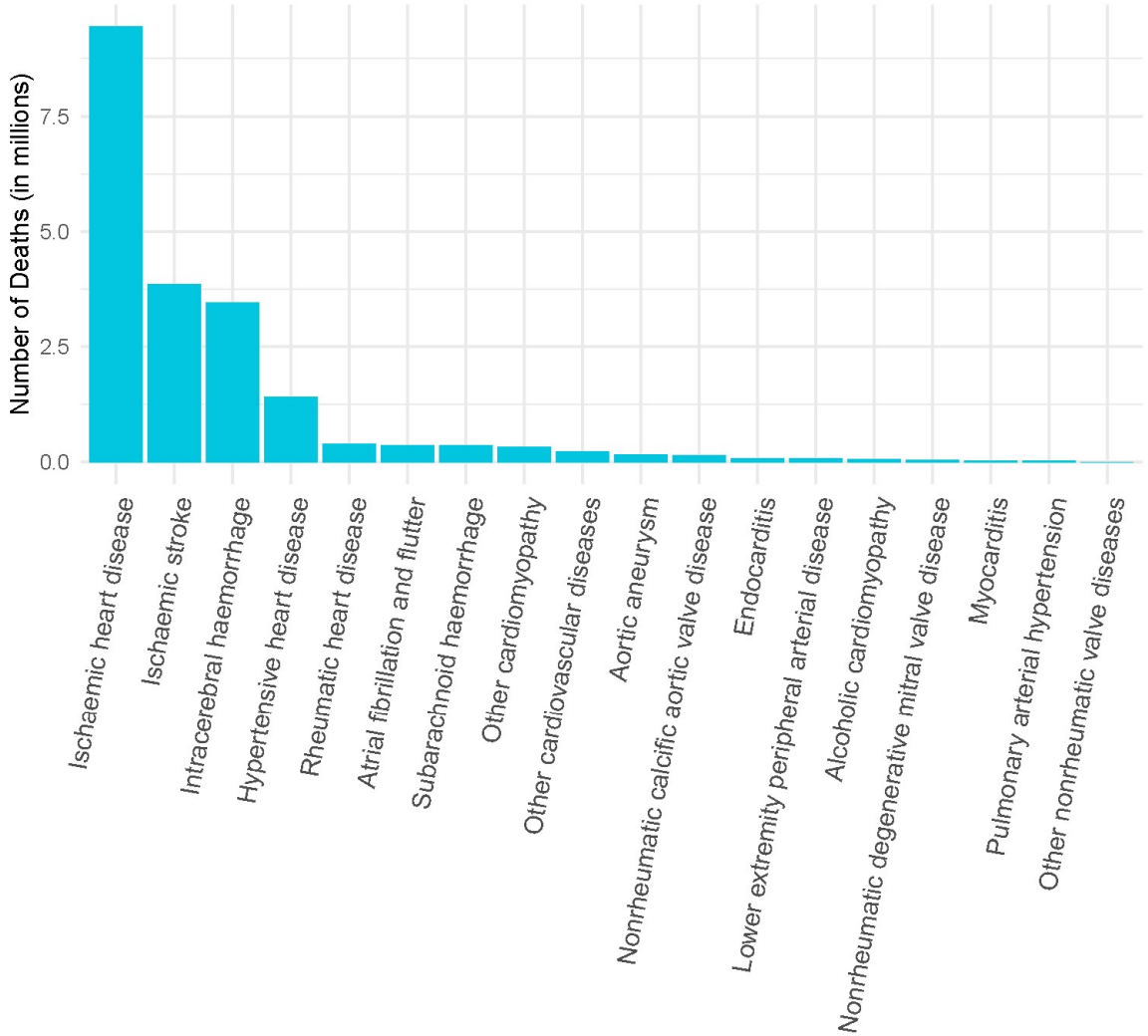


Figure 4. Leading causes of CVD death in 2021. Data were sourced from Global Burden of Cardiovascular Diseases and Risks Collaboration, 1990-2021 [56]

1.6.3. Risk Factors

The development and progression of CVD are influenced by a complex interplay of non-modifiable and modifiable risk factors [57]. Non-modifiable factors define an individual's baseline susceptibility and cannot be altered [57]. In contrast, modifiable risk factors include lifestyle and clinical factors that have the potential to be improved or controlled through behavioural changes and medical interventions [58]. Importantly, the effects of exposure to the modifiable risk factors accumulate over the life course [59].

Non-modifiable risk factors for CVD include [58]:

- **Age**: CVD risk increases with age, particularly in individuals over 50.
- **Sex**: Men have a higher risk of developing CVD and tend to do so about a decade earlier than women. However, women with CVD may experience worse outcomes.
- **Family history**: A family history of premature CVD reflects both genetic susceptibility and shared environmental factors, and is associated with a higher lifetime risk.
- **Ethnicity**: Individuals of South Asian and sub-Saharan African descent are at higher risk, while those of Chinese or South American origin generally have a lower risk compared to those of European ancestry.

Numerous modifiable factors have been implicated in CVD pathogenesis. They include, but are not limited to [58]:

- **Metabolic and clinical factors**, for example hypertension, dyslipidaemia (particularly elevated LDL cholesterol), diabetes mellitus, obesity, chronic kidney disease, and chronic inflammatory disorders such as rheumatoid arthritis and lupus.
- **Lifestyle factors** such as smoking, poor diet, physical inactivity, and excessive alcohol consumption influence CVD risk in a dose-dependent manner, but the magnitude and even direction of these effects differ depending on the pattern and context of exposure – for example, the intensity and duration of smoking or the distinction between regular moderate and episodic heavy alcohol intake.
- **Environmental and psychosocial factors**, for example social deprivation, low educational attainment, air pollution, and chronic stress.

It is important to acknowledge that this list is not exhaustive and that there is variability in both the evidence for, and the strength of, associations between individual risk factors and CVD. For example, as discussed in subsequent sub-sections of **Section 1.6.3**, large-scale meta-analyses consistently report strong, approximately log-linear associations between systolic blood pressure or total cholesterol and CVD incidence, whereas the evidence for the effects of social deprivation or rheumatoid arthritis is generally weaker. These variations highlight the underlying uncertainty and complexity of CVD aetiology.

Among these modifiable risk factors, four are classified as Standard Modifiable Risk Factors (SMuRFs), which are well-established contributors to IHD[60]:

- Dyslipidaemia
- Hypertension
- Diabetes mellitus
- Smoking

These SMuRFs have been the cornerstone of CV prevention strategies for over 50 years and are recognised as major contributors to the population burden of IHD [60]. A significant proportion of patients with acute coronary syndrome present with these risk factors. A meta-analysis of 14 clinical trials involving 122,458 patients with CAD found that 83% had at least one SMuRF [61]. Similarly, in the Swedish myocardial infarction registry, which included 62,048 patients with STEMI, 85% had one or more SMuRFs [62].

In line with the variables included in the ASsessing cardiovascular risk using SIGN guidelines to assign preventive treatment (ASSIGN) risk score, which was developed and validated for the Scottish population (see **Section 1.7.3**), my models were adjusted for age, sex, and the following modifiable risk factors: SMuRFs (using variables such as average systolic blood pressure, total and high-density lipoprotein (HDL) cholesterol, smoking, diabetes), rheumatoid arthritis, and social deprivation measures (Scottish Index of Multiple Deprivation (SIMD) score and years of education, with the latter not included in ASSIGN but complementing SIMD). Although family history of CVD is part of the ASSIGN score, it was intentionally excluded. Family history is a non-modifiable factor that largely reflects inherited genetic predisposition, which may also influence circulating protein profiles. Adjusting for it could therefore attenuate true biological associations between protein levels and disease risk.

The selected covariates represent a pragmatic subset of modifiable factors, allowing effective control for established confounders while preserving the ability to detect multi-omic signals relevant to disease risk. In the following sections, I will summarise the evidence linking these covariates to CVD.

1.6.3.1. High Blood Pressure

Systolic blood pressure (the higher reading in a standard blood pressure measurement) is the pressure exerted on arterial walls during the contraction of the heart and serves as a key indicator of CV strain. Diastolic blood pressure (the lower reading) is the pressure in the arteries when the heart is at rest between beats. Elevations in either systolic or diastolic pressure cause atherosclerotic CVD, though their relative importance varies with age, with raised diastolic pressure often more predictive in younger adults and raised systolic pressure predominating in older populations [63,64].

The National Institute for Health and Care Excellence (NICE) classifies brachial blood pressure readings between 120/80 mmHg and 140/90 mmHg as "high normal," while hypertension is defined as a clinic blood pressure of 140/90 mmHg or higher [65].

Elevated systolic pressure is a leading contributor to premature CVD death [49]. In England, more than one in four adults are affected by hypertension [65]. Epidemiological evidence consistently shows a strong association between systolic blood pressure and CV mortality. The Prospective Studies Collaboration demonstrated that, in adults aged 40 – 69 years, each 20 mmHg increment in usual systolic blood pressure (or approximately 10 mmHg in diastolic pressure) was associated with more than a twofold increase in stroke mortality and approximately a twofold increase in mortality from IHD and other vascular causes [66].

Randomised controlled trials (RCTs) have demonstrated that pharmacologic lowering of systolic blood pressure significantly reduces the incidence of major CV events. Trials such as Systolic Blood Pressure Intervention Trial and the Heart Outcomes Prevention Evaluation–3 have shown that more intensive systolic blood pressure control (<120 mmHg) compared to standard targets (<140 mmHg) results in better CV outcomes, particularly in high-risk individuals [67,68]. Meta-analyses of antihypertensive trials indicate that each 10 mmHg reduction in systolic blood pressure is associated with a relative risk reduction of approximately 20% for major CV events, reinforcing the concept of a continuous and dose-dependent benefit [69].

As discussed in **Section 1.7**, systolic blood pressure is a core component of virtually all major CV risk prediction algorithms, including the Framingham Risk Score, Systematic Coronary Risk Evaluation 2 (SCORE2), and QRISK3.

1.6.3.2. High LDL Cholesterol

Raised LDL cholesterol causes atherosclerotic CVD [12,70]. Individuals with genetic variants that confer lifelong lower LDL cholesterol levels exhibit substantially reduced rates of CAD – by approximately 50% - compared to those without such variants [71]. The relationship between LDL cholesterol levels and CVD risk is log-linear and dose-dependent [70]. Each 1 mmol/L reduction in LDL cholesterol is associated with a 20% to 25% relative reduction in the risk of major CVD events, including MI and ischaemic stroke [72]. This relationship holds true across different populations, age groups, and baseline levels of CV risk. Moreover, evidence suggests that the cumulative burden of LDL cholesterol exposure over time is a stronger determinant of CVD risk than a single-time measurement, underscoring the importance of early and sustained LDL cholesterol management [71].

Given its strong and modifiable nature, LDL cholesterol is a central component of most CV risk prediction algorithms, including the Framingham Risk Score and the SCORE2 model. From a public health perspective, modest reductions in average LDL cholesterol levels across a population can lead to significant declines in CVD incidence and mortality. This makes LDL cholesterol not only a target for individual clinical intervention but also a critical focus for broader preventive strategies.

1.6.3.3. Smoking

Tobacco use is one of the most significant contributors to CVD and premature death worldwide. The causal relationship between cigarette smoking and CVD is well-established through decades of epidemiological research [73–77]. Smokers have approximately 2 to 4 times higher risk of CHD compared to non-smokers, although in certain populations or subgroups (e.g., young women, heavy smokers), the risk may approach 6-fold [78]. Quitting smoking rapidly reduces CV risk, with substantial benefits (Hazard Ratio (HR) = 0.61) observed within the first five years of cessation [79].

Smoking contributes to CVD through interconnected processes including endothelial dysfunction, inflammation, and thrombosis [80]. Cigarette smoke reduces the bioavailability of nitric oxide, impairing vascular relaxation and promoting arterial stiffness [80]. It also induces systemic inflammation, marked by elevated cytokines and adhesion molecules that facilitate immune cell infiltration into vascular tissue [80]. This pro-inflammatory state enhances platelet adhesion and aggregation, while disrupting the balance between coagulation and fibrinolysis [80]. Collectively, these effects promote a pro-thrombotic environment, accelerating the development of atherosclerosis and increasing the risk of CV events. Smoking also induces long-term consequences at the molecular level, including alterations to DNA (in the form of somatic mutations, epigenetic alternations and DNA adducts) [81].

Smoking is a fundamental variable included in nearly all major CV risk prediction algorithms, such as the Framingham Risk Score, SCORE2, and QRISK3, due to its strong and well-documented association with CV morbidity and mortality.

Smoking and brain health

Cigarette smoking contributes significantly to cognitive decline, dementia, and accelerated brain aging [82,83]. Chronic exposure to tobacco smoke induces a cascade of systemic and neurovascular alterations [84]. These mechanisms interfere with the brain's ability to regulate and deliver blood where and when it is needed [85]. This leads to inadequate support for neurons, making them more vulnerable to injury and death, which contributes to cognitive decline and dementia [86].

A growing body of neuroimaging evidence indicates that smokers exhibit an increased burden of white matter hyperintensities (WMHs) – areas of demyelination typically seen in magnetic resonance imaging (MRI) imaging data [87]. WMHs are markers of cerebral small vessel disease and are strongly associated with cognitive deficits, particularly in executive function and processing speed [88]. Smoking is also linked with reduced total brain volume, accelerated cortical thinning, and subcortical atrophy, particularly in regions such as the hippocampus, thalamus, and prefrontal cortex – areas critical for memory formation and higher-order cognitive processing [89,90].

Smoking-related brain changes are at least partially reversible after cessation [91]. The brain cortex gradually recovers with each year of abstinence, although it is a long process – complete recovery in affected regions may take 25 years [91]. Similarly, long-term abstinence (>20 years) allows white matter integrity to approach levels seen in never-smokers [92]. Former smokers also experience more favourable cognitive trajectories than current smokers [93]. Together, these findings highlight smoking cessation as a critical neuroprotective strategy, with earlier quitting providing the greatest long-term benefits [94].

1.6.3.4. High Fasting Plasma Glucose

Elevated fasting plasma glucose (FPG) is a well-established, causal risk factor for CVD. Chronic hyperglycaemia contributes to atherosclerosis and vascular dysfunction through several biological mechanisms, including endothelial dysfunction and the promotion of a proinflammatory and prothrombotic environment [95]. High glucose levels also worsen lipid profiles by promoting insulin resistance, leading to elevated triglycerides, reduced HDL cholesterol, and the formation of small, dense LDL particles – all of which are associated with increased atherogenicity [95].

Multiple large cohort studies demonstrate that higher FPG, even within the normal or prediabetic range, is independently associated with increased risk of CVD. For example, a study of over 10,000 adults found a linear increase in CVD risk starting from FPG levels as low as 90 mg/dL, with no clear threshold below which risk does not rise [95]. Another cohort found that individuals with FPG in the high-normal range (95–99 mg/dL) had significantly higher CVD risk compared to those with FPG <80 mg/dL (HR 1.53; 95% CI 1.22–1.91), and risk continued to climb with higher FPG [96].

Adults with diabetes have a two to four times higher risk of developing CVD compared to those without diabetes, with risk rising further as glycaemic control worsens [97]. In recognition of this elevated risk, most CVD risk prediction tools – including SCORE2 – automatically classify individuals with diabetes as high or very high risk for CVD, regardless of other risk factors.

1.6.3.5. Socioeconomic Factors

Socioeconomic status (SES) is a composite measure of an individual's economic position within society, typically assessed using indicators such as income, educational attainment, and occupation [98]. Numerous large-scale studies have shown a strong, graded association between low SES and increased incidence of CHD [99–101], stroke [102], and HF [103]. For instance, a meta-analysis of over 1.7 million individuals found that those in the lowest SES group had a 50% higher risk of developing CVD compared to those in the highest, even after adjusting for traditional risk factors [104]. Low SES influences CVD risk through multiple mechanisms [102], including higher prevalence of smoking, poor diet, psychosocial stress, physical inactivity, and limited access to healthcare – all of which contribute to the development of hypertension, diabetes, and atherosclerosis.

Despite its significant impact, SES is not included in many well-known CVD risk prediction models, such as SCORE2 or the Framingham Risk Score. However, ASSIGN explicitly incorporates a deprivation index (SIMD) to account for socioeconomic disparities in risk [105]. This inclusion improves risk stratification, particularly in socioeconomically deprived populations where traditional models may underestimate risk. Thus, while SES is often indirectly reflected through intermediate clinical factors, its direct inclusion – as in ASSIGN – can enhance the equity and accuracy of CVD risk assessment in populations where deprivation strongly influences health outcomes.

1.6.3.6. Chronic Inflammatory Diseases

Chronic inflammatory diseases, such as rheumatoid arthritis (RA), are now recognised as risk factors for CVD [58]. Individuals with RA face an approximately two-fold increased risk of developing CAD and MI compared to the general population [106]. Notably, the magnitude of this excess risk is comparable to that observed in individuals with diabetes [106]. The excess CVD risk in RA is largely attributed to persistent systemic inflammation, which accelerates atherosclerosis and promotes endothelial dysfunction [106]. Additional mechanisms include increased prevalence of traditional risk factors (e.g., smoking, physical inactivity) [106] and shared genetic predispositions [107].

Despite this growing body of evidence, the inclusion of RA in CVD risk prediction models remains inconsistent. RA is accounted for in some tools, such as QRISK3 and the original ASSIGN (v1.0) score, both of which were developed using large, contemporary UK population datasets. In contrast, other widely used models – such as SCORE2 and the Framingham Risk Score – do not include RA as an explicit risk factor. This variation may reflect differences in model development timelines, underlying populations, and philosophies regarding the inclusion of disease-specific risk factors.

1.7. Prevention of CVD: Risk Scores

The pathological processes of CVD often progress silently for decades before any clinical symptoms emerge [6]. As a result, the first manifestation of disease may be a major and life-threatening event, such as a MI or stroke. Consequently, early detection and prevention strategies are critical to reducing its burden.

Preventive cardiology has increasingly emphasised the importance of identifying individuals at elevated risk before clinical disease manifests [108]. This requires tools that can assess risk accurately and early in the disease trajectory. Total CVD risk is the product of multiple interacting risk factors [108,109]. While a single major risk factor (such as very high blood pressure or familial hypercholesterolemia) may clearly indicate elevated risk, it is more common for individuals to present with several moderately elevated factors (such as borderline cholesterol levels, slightly elevated blood pressure, a family history of CVD, and smoking) which collectively confer a substantial risk. This multiplicative nature of risk means that relying on individual risk factors alone may lead to underestimation of true disease risk in many people.

To address this, CVD risk prediction tools have been developed to assist clinicians in identifying apparently healthy individuals who may benefit from preventive interventions [109]. These models combine data from large epidemiological studies and clinical cohorts to estimate the probability of a future CV event within a specified time frame, often 10 years. By quantifying overall risk, they allow for more informed clinical decision-making regarding lifestyle interventions, medications, and monitoring.

Effective prevention strategies rely on addressing modifiable factors while accounting for the cumulative impact of non-modifiable risks. In practice, this often involves providing preventive treatment tailored to elevated modifiable factors – for example, prescribing statins to individuals with high cholesterol or antihypertensive therapy to those with elevated blood pressure. Individuals at highest overall risk may benefit from more intensive interventions, even if measured levels of modifiable risk factors are within the normal range. For instance, a patient with LDL cholesterol levels within the normal range may still be prescribed a statin if their overall risk – based on a validated risk score – is sufficiently high to justify preventive therapy.

Several well-established CVD risk prediction tools are currently in use [110]. The Framingham Risk Score, one of the earliest tools, was derived from the Framingham Heart Study in the United States [109]. The original version of the calculator incorporates age, sex, systolic blood pressure, cholesterol levels, diabetes status, and smoking to estimate 10-year CVD risk. The QRISK3 model, based on UK general practice data, incorporates a broader range of factors, including ethnicity and chronic conditions such as atrial fibrillation and RA [111], however, it has not been validated in Scotland, as opposed to ASSIGN [105,112], which has been tailored to the Scottish population. Finally, the SCORE2 algorithm, developed by the European Society of Cardiology (ESC), adapts to regional variations in CV mortality and is used widely across Europe [113]. More details about individual models can be found in subsequent sections. All of the aforementioned risk prediction tools share several key features: they are typically derived from large observational cohorts, rely on conventional and routinely collected clinical risk factors, and employ statistical regression models to estimate CVD risk. Their respective strengths and limitations are summarised in **Table 1**.

All models demonstrate moderate to good discriminatory performance, with Area Under the Curve (AUC) values for the Receiver Operating Characteristic (ROC) generally ranging from 0.73 to 0.82 [114]. However, their predictions may vary in how closely the estimated risk aligns with the actual risk. They may be miscalibrated due to differences in baseline CVD incidence across regions or reliance on historical cohorts. Many models exclude social, environmental, and mental health factors – such as air pollution, education, chronic stress, or depression – that independently influence risk. Representation of ethnic minorities is often limited, reducing predictive accuracy in diverse populations. Additionally, most tools provide static 10-year risk estimates, do not fully capture changes in risk factor profiles (e.g., weight gain, smoking history, or improved cholesterol control), and omit emerging predictors such as omic biomarkers and gene–environment interactions, which could improve early detection and individualised risk assessment.

As scientific understanding of disease mechanisms advances – particularly through genomics, epigenomics, and other omics technologies – there is growing interest in integrating novel biomarkers into prediction tools. The goal is to enhance precision, improve risk stratification, and identify individuals who might otherwise be overlooked by conventional algorithms. Future models may incorporate molecular markers, such as DNA methylation (DNAm) signatures or polygenic risk scores, to better reflect cumulative exposures and individual susceptibility.

Table 1. Advantages and disadvantages of selected CVD risk prediction tools currently used in clinical settings.

Score	*Advantages*	*Limitations*	*Ref.*
Framingham	• Closely phenotyped cohort, high validity for component parts (risk factors), • Widely published and utilised • Externally validated • Multi-generational • Can be adjusted for individual country CVD data	• Based on a small US community (largely white, middle income) • Minimal risk factor set • Does not take into account deprivation level, family history of CVD or ethnicity • As it is based on historical data when CVD risk in the population was higher, it may overestimate risk	109
QRISK3	• Huge, contemporary dataset • Frequently updated algorithm • Large risk factor set (e.g. social deprivation, history of CVD and the effect of existing antihypertensive treatment)	• Based on population from England and Wales (may underestimate risk in European populations) • For some risk factors the dataset was incomplete and therefore imputation and statistical modelling were required • Patients with diabetes excluded	111
ASSIGN v.1.0	• Takes into account social deprivation and family history of CVD • Uses a quantitative measure of smoking	• Based on historical population data from Scotland • Does not take into account variables included in more recent calculators.	105
SCORE2	• Easy to use • Based on 45 European cohort studies covering a wide geographic spread of countries • Available for low, moderate, high and very high European risk regions. • Dedicated version for Older Persons (SCORE2-OP) enables risk prediction for patients 70 years old and above. • Externally validated	• Minimal risk factor set • May underestimate risk in patients with diabetes, central obesity, family history of premature CVD	115

37

1.7.1. Framingham Risk Score

Link: https://www.framinghamheartstudy.org/fhs-risk-functions/cardiovascular-disease-10-year-risk/ (newest version)

The Framingham Risk Score is one of the earliest and most widely known CV risk prediction tools, developed from the long-running Framingham Heart Study in the United States [109]. It was designed to estimate an individual's 10-year risk of developing a first major CV event, based on epidemiological data collected over several decades. The Framingham Risk Score has been widely adopted globally and served as a foundation for many later risk models, including region-specific adaptations such as QRISK3 and ASSIGN.

The original version of the Framingham Risk Score predicts the following CV outcomes [109]:

- CHD (including angina and MI)
- Stroke
- Peripheral arterial disease
- HF

The score was initially developed for adults aged 30 to 74 years who were free of CVD at baseline. Later versions have been adapted for broader age ranges and different outcomes, including stroke-specific or general CVD risk.

The input variables of the classic Framingham Risk Score include [109]:

- Age
- Sex
- Total cholesterol
- HDL cholesterol
- Systolic blood pressure
- Use of antihypertensive treatment
- Smoking status (smoker, non-smoker)
- Diabetes

The Framingham algorithm calculates a percentage 10-year risk of a CV event. A score of ≥20% is typically considered high risk, indicating the need for more aggressive risk-reducing interventions, including pharmacological treatment (e.g., statins or antihypertensives). Intermediate and low-risk thresholds are used to guide lifestyle advice and surveillance frequency.

Although the Framingham Risk Score was ground-breaking, it has limitations when applied outside the original study population. The Framingham cohorts were predominantly white, middle-class Americans, which raises concerns about

generalizability to other ethnicities or more socioeconomically diverse populations. Moreover, risk overestimation has been observed when the Framingham Risk Score is applied to European or Asian populations without recalibration.

The Framingham Risk Score remains historically significant and methodologically influential. It provided the blueprint for population-level CV risk prediction and continues to be referenced in epidemiological studies and international guidelines. However, in contemporary UK clinical settings, tools like QRISK3 or ASSIGN are preferred due to their greater relevance to the local population and health system context.

1.7.2. QRISK3

Link: https://qrisk.org/

QRISK3 is a CVD risk calculator developed using real-world electronic health record data from general practices in England [111]. It is currently recommended by NICE for routine use in England and Wales, where it helps estimate an individual's 10-year risk of developing a first major CV event. The tool predicts the risk of the following CV outcomes [111]:

- CHD (including angina and MI)
- TIA
- Ischaemic stroke

QRISK3 applies to adults aged 25 to 84 years, capturing a broader age range than many other tools. This extended range is particularly relevant for addressing early-onset risk factors in younger individuals, as well as the complex comorbidities common in older adults.

The input variables of QRISK3 are [111]:

- Age
- Sex
- Ethnicity
- Deprivation score (UK Postcode)
- Smoking status (non-smoker, ex-smoker, light/moderate/heavy smoker)
- Diabetes (type 1 or type 2)
- Family history of premature CHD (angina in first-degree relative under age 60)
- Chronic kidney disease (stage 3, 4, or 5)
- Atrial fibrillation
- Hypertension treatment
- Migraine
- Rheumatoid arthritis
- Systemic lupus erythematosus
- Severe mental illness (schizophrenia, bipolar disorder, or severe depression)
- Atypical antipsychotic medication use
- Corticosteroid use
- Erectile dysfunction (in men)
- Total cholesterol/HDL cholesterol ratio
- Systolic blood pressure
- Systolic blood pressure variability (if multiple readings are available)
- Body mass index (BMI)

The tool calculates a percentage CVD risk [111]. A QRISK3 score ≥10% is generally used as a threshold for recommending preventive interventions, such as initiating statin therapy or intensifying lifestyle modifications. Scores below 10% may still prompt intervention depending on clinical judgment, comorbidities, or patient preference.

QRISK3 is not validated for use in Scottish populations. Its risk models are based on English primary care data, and population-specific differences – such as baseline risk, socioeconomic variation, and healthcare access – may limit accuracy when applied elsewhere. In contrast, the ASSIGN score, developed in Scotland, incorporates measures of social deprivation and family history and is calibrated to the Scottish population. While ASSIGN uses fewer variables than QRISK3, it may offer improved population-specific accuracy in regions where QRISK3 has not been validated.

1.7.3. ASSIGN

Link: https://rightdecisions.scot.nhs.uk/assign-v20/assign-cardiovascular-risk-score-calculator/ (newest version)

ASSIGN version 1.0. is a CVD risk calculator developed specifically for use in Scotland, based on data from the general population (Scottish Heart Health Extended Cohort; SHHEC) [105,116] . ASSIGN is selected by the Scottish Intercollegiate Guidelines Network (SIGN) and the Scottish Government Health Directorates for use in clinical practice [112]. Unlike QRISK3, which is derived from English primary care records, ASSIGN was designed to reflect the demographic, clinical, and socioeconomic characteristics of the Scottish population [105]. It estimates the 10-year risk of a first CV event, aiding in primary prevention decisions.

The tool predicts the risk of the following CV outcomes [105]:

- CHD (including angina and MI)
- TIA
- Ischaemic stroke

All analyses in this thesis are based on the original version of ASSIGN, developed in 2006 by Professor Hugh Tunstall-Pedoe and Professor Mark Woodward in collaboration with SIGN [105]. The score was updated in 2024 to version 2.0. [116]. Although this update reflects valuable refinements to CVD risk prediction in Scotland, it was released after the analysis period of this work. Nevertheless, the principles underpinning both versions are consistent: region-specific calibration, a focus on first-event prevention, and inclusion of deprivation and family history as key contributors to CV risk.

The 2006 ASSIGN model estimated the 10-year CVD risk for individuals aged 30 to 74 years (which is slightly narrower than the range used in QRISK3), incorporating the following risk factors [105,112]:

- Age
- Sex
- Smoking (cigarettes per day)
- Systolic blood pressure
- Total cholesterol
- HDL cholesterol
- Family history of premature CVD (first-degree relative <60 years)
- Diabetes (type 1 or type 2)
- Rheumatoid arthritis
- SIMD deprivation score

This version was notable for its early emphasis on socioeconomic factors and family history, which were not commonly included in other UK risk models at the time. Its outputs were expressed as a percentage risk, with ≥20% generally used as the threshold for initiating preventive interventions [105].

A revised version of the ASSIGN tool incorporates several important updates [116]:

- Age range expanded to 25 to 90 years, improving applicability to both younger and older adults.
- Recalibration based on more recent cohort data, reflecting current CVD incidence trends and improved population health.
- Updated SIMD values from the 2012 version to the 2020 SIMD tables, ensuring more accurate representation of current socioeconomic deprivation.
- Rheumatoid arthritis was removed from the input variable list.

ASSIGN is well-suited to the Scottish population, addressing regional differences in risk factors, healthcare access, and disease burden. It may be less applicable to non-Scottish or ethnically diverse populations, as it does not explicitly include ethnicity as a variable. Moreover, its limited variable set may reduce its sensitivity in capturing complex, individual-level risk profiles compared to newer models like QRISK3.

Nevertheless, ASSIGN remains an essential tool for population-specific risk prediction in Scotland, particularly in settings where QRISK3 has not been validated.

1.7.4. SCORE2

Link: https://www.heartscore.org/en_GB

SCORE2 is a CV risk prediction model developed by the ESC [113]. Introduced in 2021 as an update to the original SCORE model, SCORE2 estimates the 10-year risk of developing a first fatal or non-fatal CV event in individuals without prior CVD. It is recommended in current ESC guidelines for use in European countries, with risk models calibrated to specific European regions based on population-level event rates.

SCORE2 predicts the following CV outcomes [113]:

- Non-fatal MI
- Non-fatal stroke
- CVD death

Unlike its predecessor, which estimated only CVD mortality, SCORE2 incorporates both fatal and non-fatal events, offering a more comprehensive and clinically relevant risk estimate.

The tool is intended for use in adults aged 40–69 years. For individuals aged 70 and older, a separate version called SCORE2-OP (Older Persons) is used, which accounts for the different risk dynamics in older populations.

The input variables of SCORE2 include:

- Age
- Sex
- Smoking status (smoker, non-smoker)
- Systolic blood pressure
- Total cholesterol
- HDL cholesterol

SCORE2 stratifies countries into four CV risk regions (low, moderate, high, and very high) based on observed event rates. This geographic calibration allows for more accurate, region-specific predictions across diverse European populations.

The tool expresses risk as a percentage likelihood of a major CV event within 10 years. SCORE2 uses age-dependent thresholds to define high-risk status. For example, a 10-year risk ≥7.5% may be considered high in a 50-year-old, while a lower threshold may apply to younger individuals. These thresholds support early preventive intervention before the onset of symptomatic disease.

Compared to older models such as the original SCORE or Framingham Risk Score, SCORE2 offers several improvements:

- Inclusion of non-fatal events, improving clinical relevance
- Updated European population data, increasing applicability and accuracy
- Age-specific and region-specific risk thresholds, improving personalisation
- A companion tool (SCORE2-OP) tailored for the elderly

However, SCORE2 does not include variables such as ethnicity, socioeconomic status, chronic inflammatory conditions, or mental illness, which are considered in more granular tools like QRISK3. Additionally, it does not account for family history of premature CVD, a notable risk factor in some individuals.

2. Multi-omic Biomarkers

Despite extensive research, predicting who will develop CVD remains difficult. As outlined in **Chapter 1**, current risk prediction tools are largely based on clinical and demographic risk factors. Yet, up to 20% of patients with CHD present without traditional risk factors [60], underscoring the limitations of existing models. To improve risk stratification and guide more effective prevention, there is growing interest in integrating multi-omic biomarkers into clinical models.

Multi-omics includes data from genomics, epigenomics, transcriptomics, proteomics, and metabolomics. While genomics reveals inherited susceptibility, the other layers reflect dynamic, context-dependent changes in gene regulation, cellular activity, and systemic physiology. By leveraging these complementary data types, it may be possible to move beyond conventional risk scores toward more precise, biology-driven models of CVD prediction and prevention. In this chapter, I introduce the central dogma of molecular biology, describe each omic layer in detail and examine the potential of multi-omic biomarkers to improve CVD risk prediction.

2.1. The Central Dogma of Molecular Biology

Genetic information in living organisms is stored in DNA, a molecule composed of nucleotide sequences. Each nucleotide consists of three components: a phosphate group, a nitrogenous base, and a five-carbon sugar called deoxyribose. The nitrogenous bases in DNA are adenine (A), thymine (T), guanine (G), and cytosine (C). These bases pair specifically – A with T and G with C – through hydrogen bonding, forming the double-stranded helical structure of DNA. Thanks to complementary base pairing, genetic information can be faithfully copied from an existing DNA molecule during DNA replication. This process ensures that each daughter cell receives an identical copy of the genome during cell division.

To produce functional molecules such as proteins, the genetic information encoded in DNA must first be converted into RNA through a process called transcription. During transcription, one strand of DNA serves as a template for synthesising a complementary RNA molecule. In RNA, the base uracil (U) replaces thymine; thus, adenine in the DNA pairs with uracil in the RNA. In the next step, called translation, the sequence of nucleotides in the RNA is read in sets of three bases (codons), each corresponding to a specific amino acid. Ribosomes facilitate the assembly of these amino acids into a polypeptide chain, which folds into a functional protein. This directional flow of genetic information – from DNA to RNA, and from RNA to protein – is known as the central dogma of molecular biology [117] (**Figure 5**). It describes the fundamental pathway by which the instructions encoded in the genome are ultimately expressed as cellular structure and function.

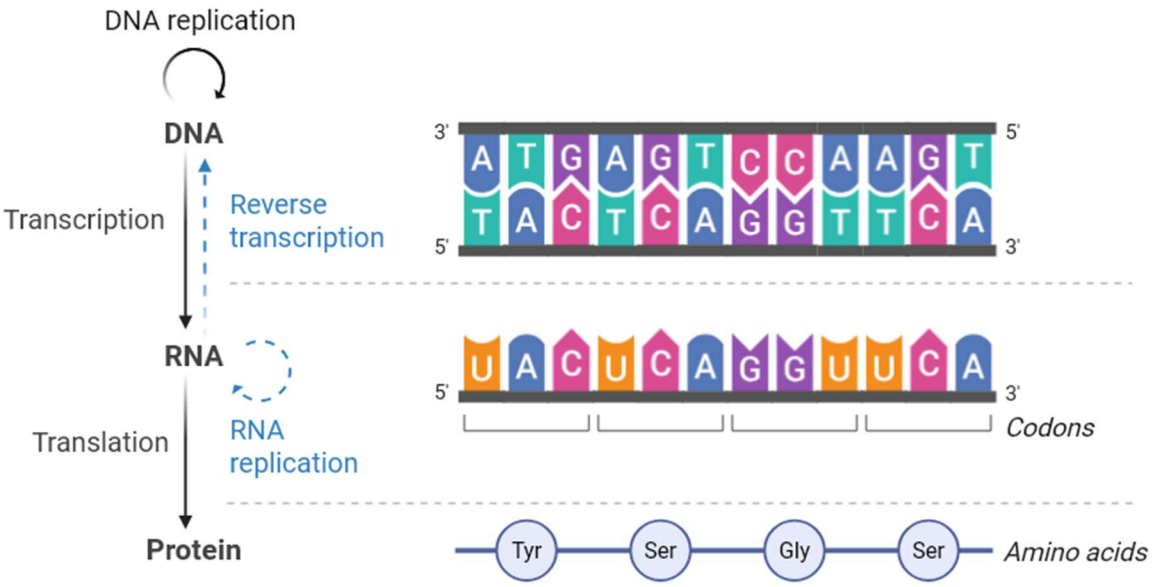


Figure 5. The central dogma of molecular biology and additional information transfers. Genetic information primarily flows from DNA to RNA through transcription, and from RNA to protein via translation, as indicated by solid black arrows. This canonical pathway underpins gene expression in all living organisms. During DNA replication, each strand of the double helix serves as a template for synthesising a new complementary strand through specific base pairing: adenine (A) pairs with thymine (T), and cytosine (C) pairs with guanine (G). In transcription, one DNA strand is used as a template to synthesise messenger RNA, where A pairs with uracil (U) instead of T, while C still pairs with G. The resulting RNA carries the genetic code to the cytoplasm, where it is read in triplets of nucleotides, or codons, during translation. Each codon specifies a particular amino acid. Ribosomes facilitate the sequential addition of amino acids into a growing polypeptide chain, which then folds into a functional protein. Additional, non-canonical routes of information transfer – RNA replication (RNA → RNA), and reverse transcription (RNA → DNA) – are shown as blue dashed arrows. The latter two processes are mainly utilised by certain viruses and mobile genetic elements. Created with BioRender.com.

2.2. DNA

DNA encodes stable, lifelong information that underpins human development, cellular function, and responses to environmental cues. This information is organised into genes – segments of DNA that serve as instructions for building proteins or regulating biological processes. The genome refers to the complete set of genetic material in an organism. In humans, it consists of approximately 3 billion nucleotides, organised into 23 pairs of chromosomes, including 22 pairs of autosomes and one pair of sex chromosomes (XX in females, XY in males). Despite high similarity among individuals, human genomes differ by approximately 0.4%, accounting for both single-nucleotide changes and larger differences involving multiple bases [118,119]. These differences, collectively known as genomic variants, help shape individual traits and disease risk, although most have no functional impact.

There are several types of genomic variants [118,119]:

- Single-nucleotide variants (SNVs) are the most common. They involve a change in a single base. When present in at least 1% of the population, they are referred to as single-nucleotide polymorphisms (SNPs).
- Insertions and deletions (indels) involve the addition or loss of a small number of nucleotides (typically <50 bases). Though less frequent than SNVs, indels can disrupt gene function.
- Tandem repeats are short sequences repeated multiple times in a row. When they involve more than 50 bases, they are classified as structural variants (SVs).
- SVs refer to larger genomic alterations (≥50 bases), including large insertions, deletions, inversions, duplications, and translocations. A subtype, copy-number variants (CNVs), involves changes in the number of copies of a genomic region.

To date, over 80 million SNPs have been identified in the human genome [120]. Due to their abundance and biological relevance, they serve as a primary source of genetic information in the empirical analyses presented in this thesis.

Genomics – the comprehensive study of the structure, function, variation, and regulation of the genome – seeks to characterise the complete DNA sequence and its variants across individuals and populations. A variety of technologies have been developed to study the genome, with genotyping arrays and next-generation sequencing (NGS) being the most commonly used in population-based and epidemiological research. The principles of these approaches are outlined in the sections below. They form the conceptual and technical foundation for subsequent methods discussed in this thesis.

2.2.1. DNA Measurement Methods

Microarrays

DNA microarrays are widely used to detect known genetic variants across the genome. Platforms such as the Illumina Infinium HumanOmniExpressExome BeadChip (>950,000 SNPs)[121] and the Global Screening Array (>650,000 SNPs, depending on the version)[122] typically genotype hundreds of thousands of variants. However, as this represents only a subset of common genomic variation, a statistical method known as genotype imputation is routinely applied to infer untyped variants.

Imputation leverages the haplotype structure of the genome – that is, the correlated inheritance of neighbouring variants along a chromosome – to infer genotypes for millions of SNPs not directly assayed by the array [123]. Commonly used imputation panels include the 1000 Genomes Project, the Haplotype Reference Consortium (HRC), and the TOPMed reference panel. After imputation, the number of available SNPs increases dramatically, often to over 20 million variants per individual, depending on the reference panel and filtering criteria used.

Illumina Infinium assay chemistry [124] begins with whole-genome amplification of genomic DNA to generate sufficient material for analysis (**Figure *6***). The amplified DNA is then enzymatically fragmented and denatured to produce single-stranded molecules. These fragments are applied to a silicon BeadChip, where each microscopic bead is coated with a 50-mer oligonucleotide probe. Each probe is designed to hybridise immediately adjacent to the SNP of interest, stopping exactly one base before the interrogated site. This is a key feature of the Infinium design, allowing precise targeting of SNP loci with a single probe per variant.

After hybridization, a single-base extension step is performed. A DNA polymerase performs single-base extension of the probe, adding one fluorescently labelled nucleotide that is complementary to the base on the sample DNA at the SNP site. Each base (A, T, C, G) is tagged with a dye. The fluorescence signals are captured using high-resolution laser scanning.

- Homozygotes produce a single-colour signal (e.g., red/red or green/green).
- Heterozygotes show a mixture of signals (e.g., red/green), which appears as yellow in the composite image.

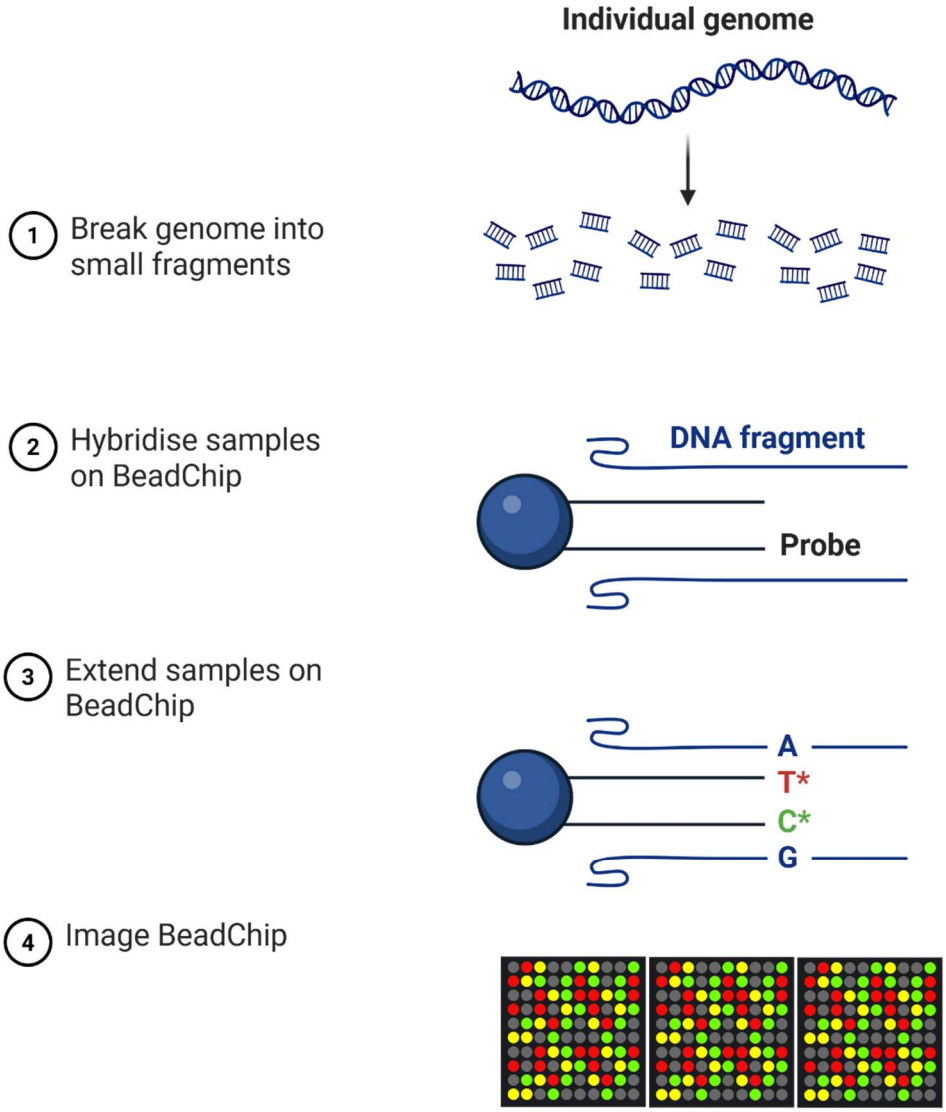


Figure 6. Overview of the Illumina Infinium genotyping workflow. The assay involves four main steps: (1) Genomic DNA is fragmented, (2) Single-stranded DNA fragments hybridise to bead-bound probes that stop one base before the SNP, (3) A single fluorescently labelled nucleotide is added to the probe by DNA polymerase, complementary to the SNP base in the sample DNA, (4) The array is scanned, and fluorescent signals (green, red, or yellow) indicate homozygous or heterozygous genotypes. Created with BioRender.com.

Sequencing-Based Approaches

NGS provides a comprehensive approach for detecting genetic variation across the genome. Unlike microarrays, which assess predefined variants, NGS can identify both known and novel mutations, including rare variants and structural changes. However, SNP detection may be less precise due to base-calling and alignment challenges [125].

There are three main types of NGS [126]. Whole-genome sequencing captures the entire genome, enabling analysis of both coding and non-coding regions. Whole-exome sequencing focuses on the protein-coding regions where many disease-related variants occur. Targeted sequencing covers specific genes or loci and is often used in clinical testing for inherited disorders or cancers.

NGS involves both lab-based and computational steps, typically divided into three main stages (**Figure 7**):

1. **Fragmentation** – DNA is first extracted and fragmented. Short-read platforms (such as Illumina or Ion Torrent) produce ~150–500 base pair (bp) fragments, while long-read technologies (e.g. PacBio, Oxford Nanopore) generate much longer reads – often over 10,000 bp – facilitating detection of structural variants and repetitive regions [127].

2. **Sequencing** – DNA fragments are sequenced in parallel. Each read corresponds to a short DNA segment. High read redundancy improves accuracy and reduces random error [125].

3. **Read alignment and variant calling** – Sequencing reads are aligned to a reference genome using tools such as Burrows-Wheeler Aligner or Bowtie2. These algorithms map reads to their most likely genomic location, allowing for mismatches and small insertions or deletions. Aligned reads are then sorted and indexed. The depth of coverage (the number of reads overlapping at a given position in the genome) is also evaluated, as high-confidence variant calls require sufficient read support [125]. Finally, variants are identified using callers such as Genome Analysis Toolkit HaplotypeCaller or DeepVariant, which detect SNPs, insertions, deletions, and larger structural changes.

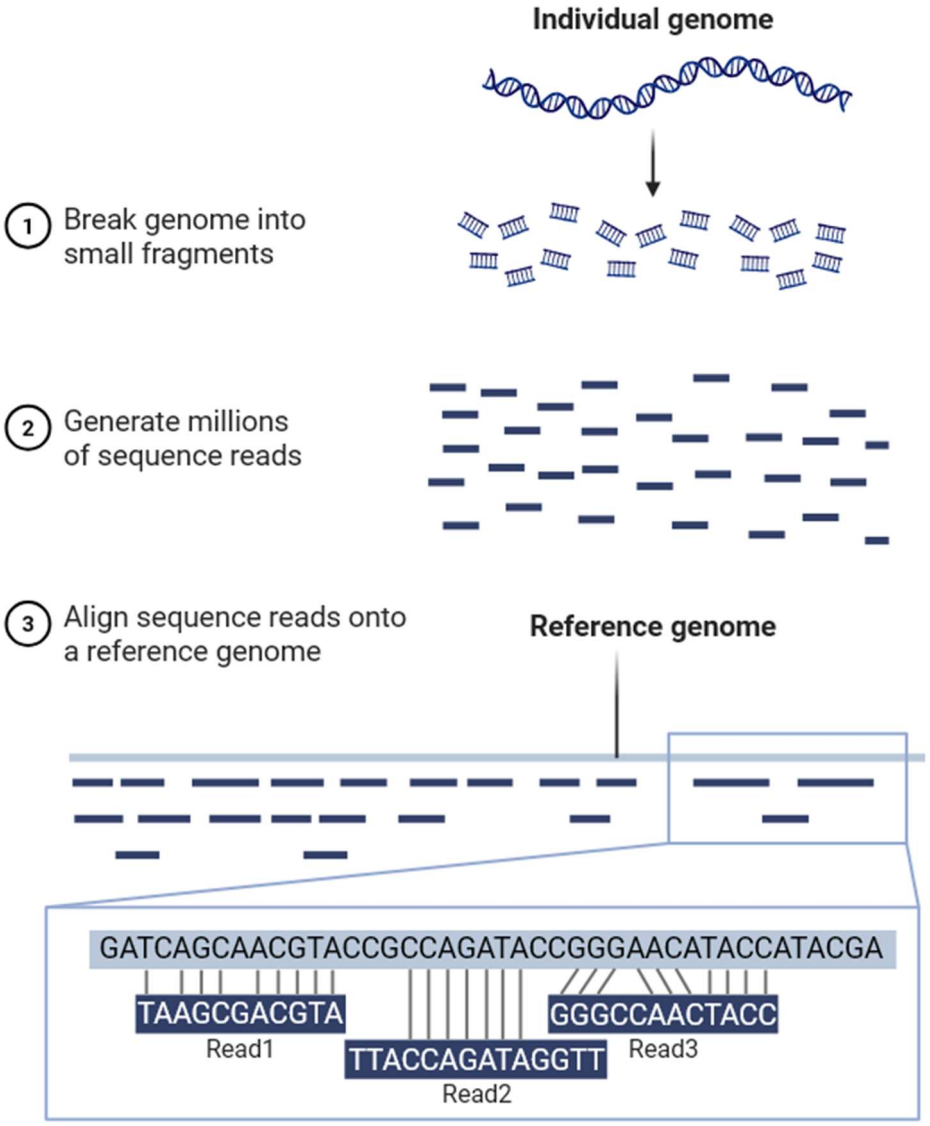


Figure 7. Overview of the next-generation sequencing and read alignment workflow. (1) Genomic DNA is extracted and fragmented into smaller pieces, (2) these fragments are sequenced to produce millions of short sequence reads, (3) the reads are then aligned to a reference genome using bioinformatic algorithms. The zoomed-in view shows individual reads (Read1, Read2, Read3) mapped to their corresponding locations on the reference genome, enabling the detection of sequence variants. Created with BioRender.com.

2.2.2. Genome-Wide Association Studies (GWAS)

Over the past two decades, GWASs have been extensively used to elucidate the relationship between common genetic variation and CVD or its risk factors [128]. As of 2025, this methodology has identified over 900,000 SNP-trait associations across more than 7,000 GWASs curated in the GWAS Catalog, and this number continues to grow steadily as sample sizes and phenotypic resolution increase [129].

In GWAS, linear or logistic regression models are typically used to test for associations, depending on the nature of the phenotype - continuous (e.g., height, blood pressure, BMI) or binary (e.g., disease status), respectively. GWAS often use tag SNPs, which serve as proxies for nearby variants due to linkage disequilibrium (LD) – the non-random association of alleles at different loci. LD enables the detection of genomic regions associated with a trait, even if the causal variant itself is not directly genotyped. Covariates such as age, sex, and ancestry-informative variables are included to reduce confounding and account for population stratification, which arises when allele frequencies differ systematically between subgroups due to ancestry rather than true association with the phenotype.

GWAS are typically conducted in a “one-SNP-at-a-time” framework. In this approach, each genetic variant is tested individually for its association with the phenotype, rather than simultaneously modelling all variants. This is necessary for several reasons. Firstly, the number of variants far exceeds the number of samples, making it statistically impractical to estimate all effects jointly. Secondly, variants in LD are correlated, which can reduce the reliability of coefficient estimates in linear regression models.

For any given variant, the following formula is solved (**Eq. 1**):

$$y = \mu + W\alpha + v\beta + \varepsilon$$

Eq. 1

Here, y is the phenotype vector, μ is the intercept, W is the matrix of covariates and α is the corresponding vector of their effect sizes; v is the vector of genotypes for the variant being tested (typically coded as 0, 1, or 2 for the number of minor alleles), and β is the effect size of that variant; ε is the error term.

Using this approach, millions of individual variants are tested for association with a phenotype of interest, making it necessary to apply a stringent multiple-testing threshold to limit false positives. The International HapMap Project and related studies have estimated roughly 1 million independent common variants across the human genome, yielding a Bonferroni significance threshold of $P < 5 \times 10^{-8}$ (representing a false discovery rate of approximately $0.05/10^6$) [130,131].

2.2.3. GWASs of CVD and Smoking

According to the GWAS catalog, the link between genetics and various subtypes of CVD was examined in more than 2600 publications [129] (status for September 2025). For example, a recent comprehensive genome-wide association meta-analysis of CAD (185,000 individuals; 9.4 million variants) linked 56 variants to CVD at $P<5 \times 10^{-8}$ [132]. This included 48 known and ten new loci. One of the most consistently replicated genetic risk loci for CAD lies on chromosome 9p21 [133]. This locus was initially defined through several strongly associated SNPs, which highlighted a shared risk haplotype inherited across large segments in European and East Asian populations [134]. The haplotype lies adjacent to the *CDKN2A/B* gene cluster and encompasses multiple non-coding regulatory elements, including distal enhancers and the long non-coding RNA *ANRIL* [134]. Carriers of the risk haplotype have an approximately 30-40% increased risk of CAD [134].

Although functional studies suggest that the 9p21 haplotype influences expression of *CDKN2A/B* and *ANRIL* in vascular cells, the precise causative variant(s) remain unresolved [135,136]. This reflects a broader challenge in GWAS: association signals typically identify statistical markers rather than the functional nucleotide changes themselves. In regions such as 9p21, where many variants are inherited together on a common haplotype and lie entirely within non-coding regulatory DNA, it is difficult to pinpoint which specific variant - or combination of variants - drives disease susceptibility. Consequently, the 9p21 region remains one of the best-characterised yet mechanistically complex examples of the limitations of GWAS resolution in defining causal variants within associated loci. Sequence variation at 9p21 is associated with an approximately 30 - 40% increased risk of CAD in individuals of European and East Asian ancestry [134].

Variants in the *PCSK9* gene, which lies in a another region associated with CVD through GWAS, exemplify the translational potential of genetic discoveries [137]. This gene plays a key role in regulating cholesterol levels in the blood and harbours both common and rare mutations with a range of effects – from modest reductions in LDL cholesterol to monogenic hypercholesterolemia. For instance, nonsense mutations in *PCSK9* were associated with a 28% reduction in mean LDL cholesterol levels and an 88% lower risk of CHD among 3,363 Black participants followed for 15 years as part of the Atherosclerosis Risk In Communities (ARIC) study [138]. In 9,524 White participants of the same study, sequence variation in *PCSK9* was linked to a 15% reduction in LDL cholesterol and an 47% reduction in CHD risk [138]. These genetic insights directly informed the development of *PCSK9* inhibitors now used as effective therapies for individuals with elevated CVD risk.

Less than 25% of CAD loci can be attributed to established clinical risk factors, including lipid levels, blood pressure, glycaemic traits and, to a much smaller extent, smoking behaviour [133]. In the largest exome-wide genetic association study of smoking behaviour to date, encompassing traits such as smoking initiation, cigarettes per day, pack-years, and smoking cessation (up to n = 622,409; $P < 5 \times 10^{-8}$) [139], more than 40 SNVs were associated with the studied phenotypes. Four variants reached genome-wide significance specifically for smoking pack-years (n = 131,892) – a composite trait central to this thesis, calculated as: pack-years = (cigarettes per day ÷ 20) × years smoked. These variants, linked to smoking behaviour in previous GWASs [140–142], were located in or near *PDE1C, DBH, CHRNA3,* and *RAB4B*. *PDE1C* and *RAB4B* have been implicated in cancer development and progression [143,144], whereas *DBH* and *CHRNA3* play roles in neurochemical pathways underlying nicotine dependence [145,146].

2.3. DNA Methylation

DNAm is a tissue- and cell-specific type of an epigenetic modification that regulates gene expression without altering the underlying DNA sequence [147]. It involves the covalent attachment of a methyl group to the fifth carbon of cytosine residues (5-methylcytosine; 5mC), most commonly at cytosine-phosphate-guanine (CpG) dinucleotides. These CpG sites are enriched in promoter regions but are also present in gene bodies and intergenic regions (**Figure 8**) [148]. Methylation in promoter regions or the first exon is typically associated with transcriptional repression [149]. In contrast, methylation within gene bodies can promote transcriptional activity by modulating alternative splicing and silencing alternative intragenic promoters [149].

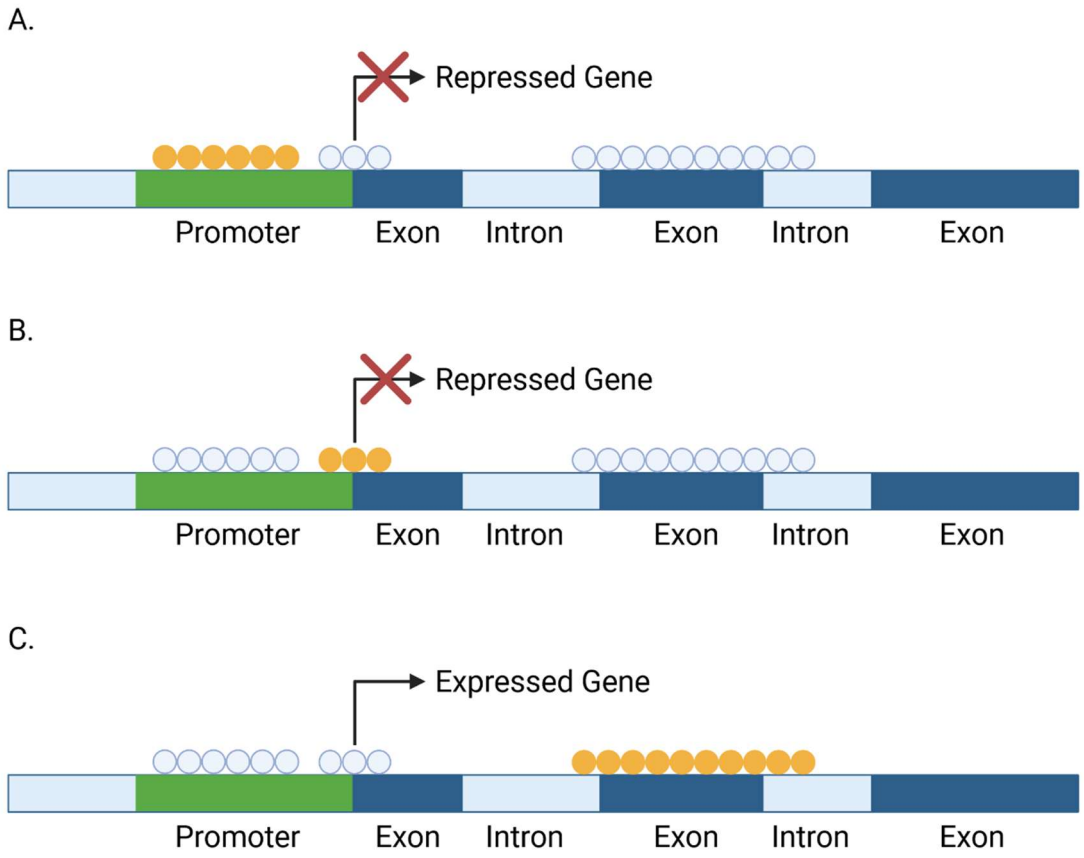


Figure 8. The relationship between DNA methylation and gene expression. Empty circles represent unmethylated cytosines, while filled circles indicate methylated cytosines. Methylation is usually (but not always) linked with transcriptional repression. While increased methylation in promoter regions (A) and the first exon (B) is typically associated with gene silencing, methylation within gene bodies (C) can promote transcriptional activity by influencing alternative splicing (exon inclusion or exclusion) and silencing alternative intragenic promoters. Figure adapted from Russo *et al*. [149]. Created with BioRender.com.

The human genome contains approximately 28 million CpG sites, of which an estimated 60 - 80% are methylated in mammalian cells. This extensive methylation landscape is maintained by DNA methyltransferases, the enzymes responsible for catalysing the addition of methyl groups to DNA. The absence of DNA methyltransferases is incompatible with life, as it results in embryonic lethality, underscoring the essential role of DNAm in normal development and cellular function.

DNAm is influenced by both genetic sequence variation and environmental exposures [150]. Genetic variants can influence DNAm by creating or abolishing CpG sites, altering transcription factor binding – which can block or recruit methylation enzymes – and modifying local chromatin structure, thereby affecting the accessibility of DNA to methylation enzymes [150]. These effects, known as methylation quantitative trait loci, can act both locally (cis) and at distant sites (trans). Among environmental factors, smoking is the strongest correlate of DNAm changes [151], while diet, alcohol consumption, and exposure to pollutants are also known to associate with methylation profiles [152–154]. Unlike fixed genomic variants, DNAm is dynamic and can change over the life course [155], enabling fine-tuned and temporally responsive regulation of gene activity.

Epigenomics is the large-scale study of epigenetic mechanisms – such as DNAm, histone modifications, and chromatin architecture – that govern gene regulation without altering the DNA sequence. In population-scale and epidemiological studies, DNAm profiling has become a key tool for assessing epigenetic variation, using both array-based and NGS technologies. The following sections outline the primary methods used to generate and analyse epigenomic data, and review recent advances in epigenetic biomarkers of CVD.

2.3.1. DNAm Measurement Methods

Microarrays

DNAm arrays operate on principles similar to DNA microarrays, relying on sequence-specific hybridization between fluorescently labelled nucleic acid targets and complementary probes immobilised on a solid surface. For Illumina Infinium Methylation BeadChip arrays, genomic DNA is first treated with sodium bisulfite [156], which converts unmethylated cytosines to uracil, while methylated cytosines remain unchanged [156]. During amplification by polymerase chain reaction, uracils are converted to thymines, introducing sequence differences that reflect the methylation status of individual cytosines [156]. Array probes are designed to hybridise to specific 50bp regions of this bisulfite-converted DNA, enabling quantification of methylation at individual CpG sites.

Illumina Infinium arrays incorporate two probe types (see **Figure 9** for a schematic comparison) [157,158]:

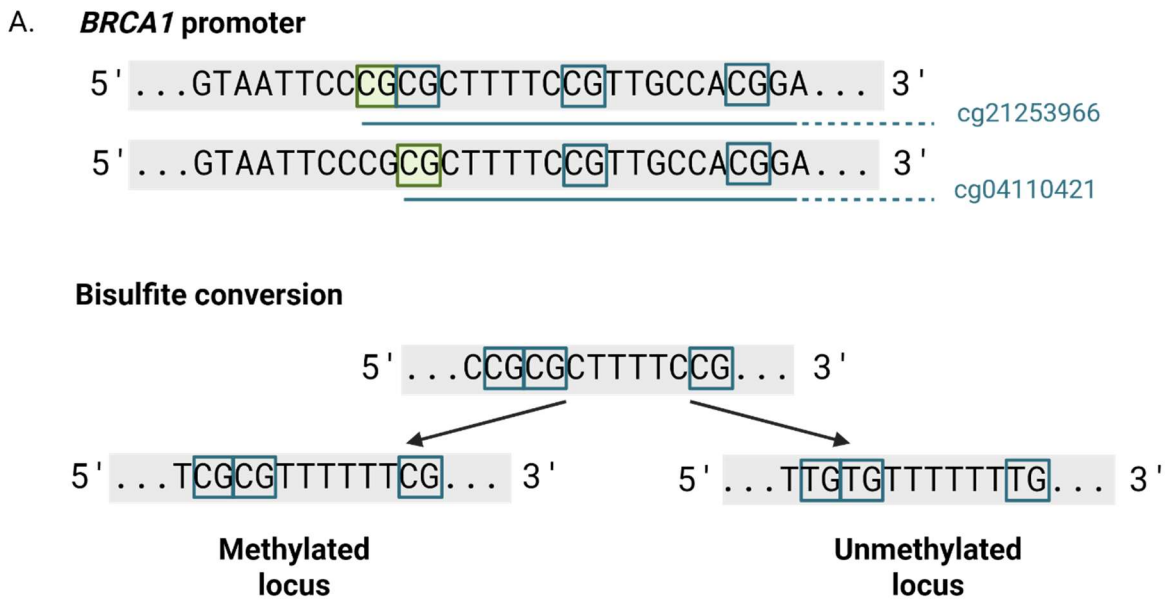
- Type I probes use two separate probes per CpG site (one for the methylated form, one for the unmethylated form), each detected via a separate colour channel.
- Type II probes use a single probe per CpG site and infer methylation levels by comparing signal intensities across two dye channels.

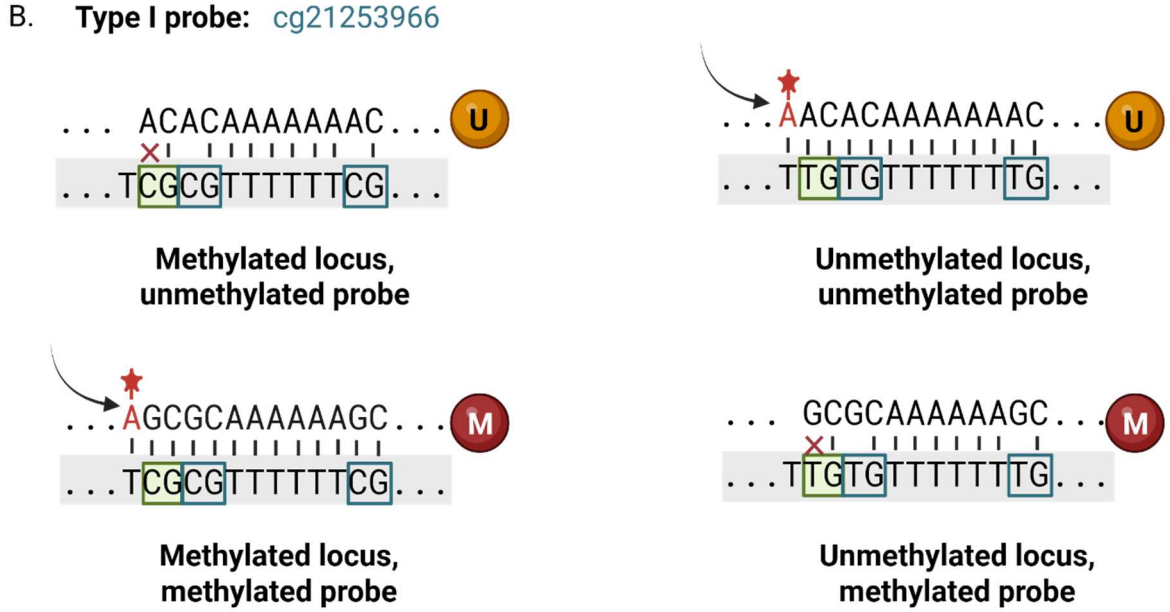
While Type II probes occupy only half the physical space on the array compared to Type I probes – and are therefore used preferentially when possible – they can tolerate up to three underlying CpG sites within the 50-mer probe sequence without compromising data quality [157]. Type I probes, by contrast, offer higher specificity and accuracy in CpG-dense regions such as CpG islands - stretches of DNA, typically >200 bp, with a high frequency of CpG sites often located near gene promoters – and CpG shores, which are regions up to ~2 kilobases flanking islands [157].

For downstream processing of array-based methylation data, the raw intensities of methylated and unmethylated probes are transformed into either beta values or M-values [159]:

- Beta values range from 0 to 1 and represent the proportion of methylation at a CpG site, calculated as the ratio of methylated probe intensity to the total probe intensity (methylated + unmethylated).
- M-values are the log2-transformed ratio of methylated to unmethylated intensities.

In my projects, I used data generated using the Illumina Infinium HumanMethylation450 BeadChip array (~485,000 CpG sites, “Illumina 450k array”) [160] and the Illumina Infinium MethylationEPIC BeadChip array (>850,000 CpG sites, “Illumina EPIC array”) [161]. More recent versions, such as the Illumina Infinium MethylationEPIC v2.0 BeadChip (released in June 2023 [162]), now cover over 935,000 CpG sites [163].


A.
BRCA1 promoter
5' ...GTAATTCCCGCGCTTTTCCGTTGCCACGGA... 3'
cg21253966
5' ...GTAATTCCCGCGCTTTTCCGTTGCCACGGA... 3'
cg04110421
Bisulfite conversion
5' ...CCGCGCTTTTCCG... 3'
5' ...TCGCGTTTTTTCG... 3'
5' ...TTGTGTTTTTTTG... 3'
Methylated locus
Unmethylated locus


B.
Type I probe: cg21253966
... ACACAAAAAAAC... U
...TCGCGTTTTTTCG...
Methylated locus, unmethylated probe
...AACACAAAAAAAC... U
...TTGTGTTTTTTTG...
Unmethylated locus, unmethylated probe
...AGCGCAAAAAAGC... M
...TCGCGTTTTTTCG...
Methylated locus, methylated probe
... GCGCAAAAAAGC... M
...TTGTGTTTTTTTG...
Unmethylated locus, methylated probe

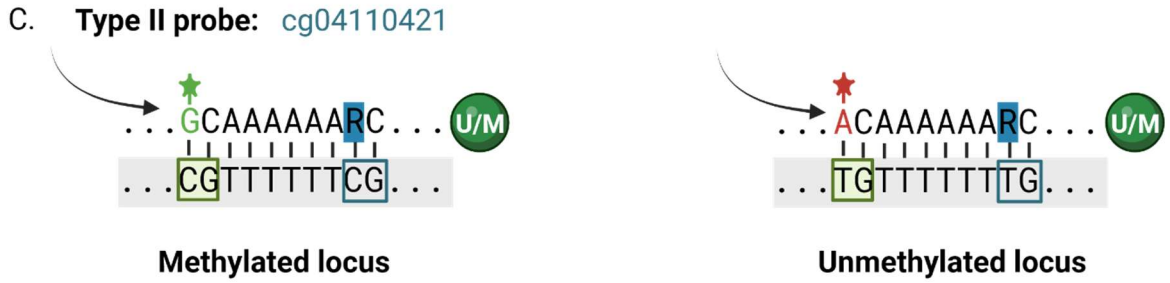

C.
Type II probe: cg04110421
...GCAAAAAARC... U/M
...CGTTTTTTCG...
Methylated locus
...ACAAAAAARC... U/M
...TGTTTTTTTG...
Unmethylated locus

Figure 9. The differences between Illumina Infinium type I and type II probes. A. Bisulfite conversion of two probes targeting adjacent CpG sites in the *BRCA1* promoter, both present on the EPIC and HM450 platforms. The Type I probe (cg21253966) and Type II probe (cg04110421) each hybridise to a 50 bp DNA sequence downstream of the targeted CpG site (highlighted in green; probe sequence underlined in blue). B. Type I probes use two separate beads to measure methylated (M) and unmethylated (U) signals. The unmethylated bead sequence is designed to match the bisulfite-converted DNA sequence assuming the locus is unmethylated, allowing single-base extension and incorporation of a labelled nucleotide immediately upstream of the target CpG (RED channel). The methylated bead is complementary to the methylated CpG, enabling detection in the GREEN channel. C. Type II probes use a single bead to detect both methylated and unmethylated signals. The cytosine at the target CpG site serves as the single-base extension locus, while all other cytosines in the probe sequence are replaced with degenerate R bases, which can hybridise to either T (unmethylated, bisulfite-converted) or C (methylated) nucleotides. Hybridisation of bisulfite-converted DNA allows single-base extension to incorporate a labelled A nucleotide for unmethylated CpGs (RED channel) or a labelled G nucleotide for methylated CpGs (GREEN channel), enabling simultaneous detection of both states on the same bead. Adapted from Pidsley *et al.* [158]. Created with BioRender.com.

Next generation sequencing

Array-based methylation assays, such as the Illumina EPIC v2.0 array, measure less than 5% of CpG sites in the human genome [163]. While these platforms are highly reproducible, cost-effective, and suitable for large epidemiological studies [158], they are designed with a fixed CpG content that can hinder discovery of new methylation signals for complex traits. By contrast, sequencing-based approaches can provide single-base resolution across the entire genome (approximately 28 million CpG sites [164]) enabling the detection of both known and previously uncharacterised methylation sites. Although arrays generally offer greater precision at interrogated loci [165], NGS expands the breadth of methylome coverage by reading millions of bases directly.

In my projects, I measured DNAm levels using two NGS approaches: Twist Bioscience sequencing (targeted short-read sequencing of 4 million CpG sites using Human Methylome Panel) and Oxford Nanopore sequencing (long-read sequencing of 28 million CpG sites). Both methods avoid bisulfite conversion, which, although long considered the gold standard for DNAm analysis, causes extensive DNA fragmentation and is therefore suboptimal for sequencing applications.

Similar to arrays, the Human Methylation Panel from Twist Bioscience measures methylation at a predefined subset of CpG sites. It relies on targeted capture, where probes selectively bind and enrich specific genomic regions for sequencing. The overlap between sites covered by the EPIC array and those targeted by the Twist Human Methylome Panel is shown in **Figure 10**. In the Twist DNA processing pipeline, DNA undergoes enzymatic conversion rather than bisulfite treatment. This approach generates a sequence pattern analogous to bisulfite conversion but preserves DNA integrity and yields higher-quality reads.

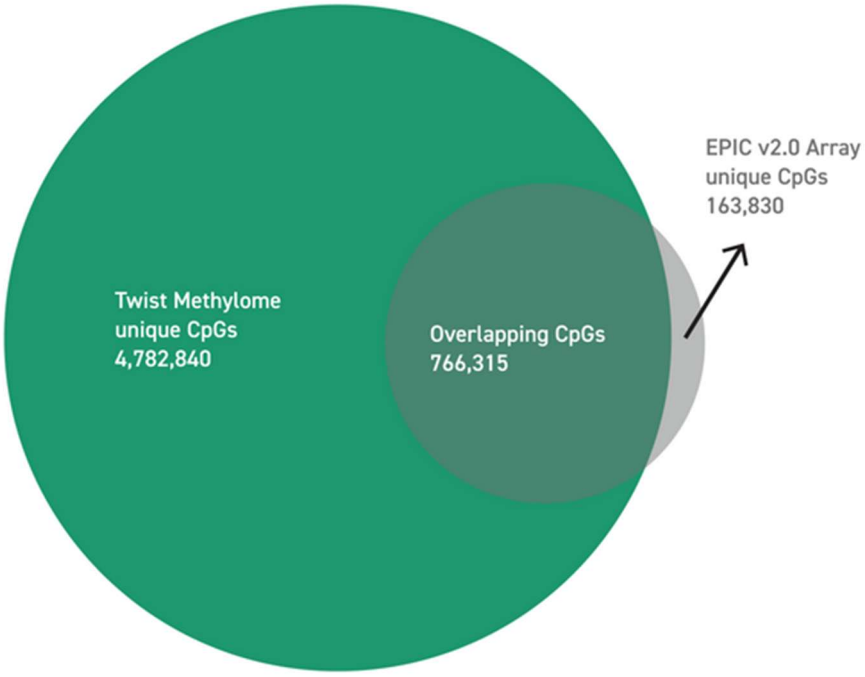


Figure 10. The overlap between sites profiled with Illumina Infinium EPIC v 2.0 BeadChip array and measured with Twist Human Methylome Kit. Sourced from Twist Bioscience [166].

Oxford Nanopore sequencing detects methylation directly in native DNA molecules (**Figure 11**). As DNA strands pass through nanopores, characteristic changes in electrical current are generated. These signals are decoded into corresponding bases, enabling simultaneous identification of DNA sequence and methylation state. The long read lengths (often >10,000 bases) facilitate measuring methylation levels across repetitive elements and haplotypes. Computational tools such as Dorado translate current patterns into methylation calls using neural networks.

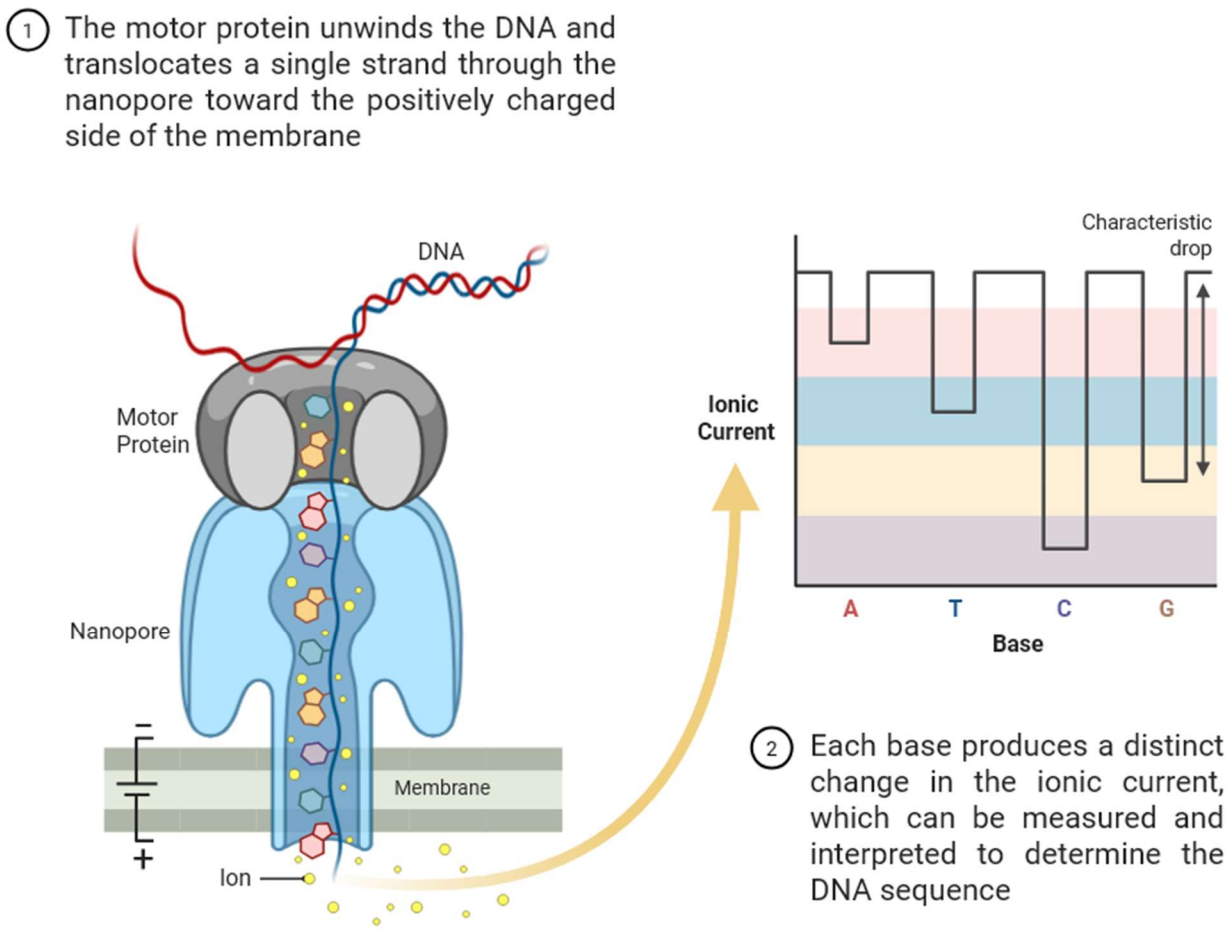


Figure 11. Basecalling with Oxford Nanopore. Basecalling is the process of converting the electrical signals generated by DNA strand passing through the nanopore into the corresponding base sequence of the strand. Glencross F., Khan D. A. (2025). Adapted from template: "Nanopore Sequencing". Retrieved from https://app.biorender.com/biorender-templates. Created with BioRender.com.

Computational processing of methylation sequencing resembles that of genetic variant detection, with reads basecalled, quality-filtered, and aligned to a reference genome before methylation states are inferred. The main distinction lies in alignment: while conventional sequencing maps only canonical bases, methylation analysis must account for cytosine modifications (Nanopore) or cytosine-to-uracil transitions (enzymatic conversion), requiring specialised tools for methylation quantification. Output files (bedMethyl or bedGraph) record the counts of methylated and unmethylated bases at each locus, which can then be used to calculate beta and M-values.

2.3.2. Epigenome-Wide Association Studies (EWAS)

EWAS examine how DNAm levels at individual CpG sites are associated with phenotypic traits or disease outcomes across the genome. In my analyses, I applied both frequentist and Bayesian EWAS frameworks.

In frequentist EWAS, linear regression is used to test associations with the outcome of interest. As in GWAS, these models are applied one site at a time. The mathematical form of the EWAS model closely resembles that used in GWAS (see **Eq. 1**), with methylation values replacing genotype dosages.

Because hundreds of thousands of CpG sites are tested simultaneously, stringent multiple testing correction is required. Saffari *et al*. used a permutation method to estimate an Illumina 450K array-specific significance threshold of $P=2.4\times10^{-7}$, and a simulation-extrapolation approach to derive a more stringent genome-wide threshold of $P=3.6\times10^{-8}$, reflecting all CpG sites across the genome [167]. In addition to multiple testing, frequentist EWAS must contend with confounding factors such as cell type heterogeneity, technical variation, and the spatial correlation of methylation marks - factors that can lead to biased or unstable estimates if not properly accounted for.

A recently developed Bayesian EWAS approach jointly models all CpG sites [168]. This captures the correlation structure across the methylome and shares information between sites, improving effect size estimation and implicitly adjusting for confounders such as age, sex, and cell composition [168]. Known covariates can also be included directly in the model.

A defining feature of Bayesian EWAS is the use of prior distributions, which encode assumptions or beliefs about the distribution of effect sizes. To manage the high dimensionality of methylation data, these models often employ sparsity-inducing priors, which shrink most effects toward zero while allowing some to remain large. The priors can be also used to classify CpG sites into those with small, moderate, or strong effects.

Posterior distributions, derived by updating the prior beliefs in light of the observed data, represent the updated uncertainty about each parameter after considering the evidence. Inference is typically performed using Markov Chain Monte Carlo algorithms [169], which generate samples from the posterior distribution. These samples are then used to estimate quantities such as credible intervals – the Bayesian analogue of confidence intervals – and posterior inclusion probabilities (PIPs), which quantify the probability that a given CpG site has a non-zero effect. Together, these metrics provide a coherent framework for uncertainty quantification and for identifying meaningful biological associations.

2.3.3. EWASs of SMuRFs

As described in **Section 1.6.3**, four common risk factors (SMuRFs: dyslipidaemia, diabetes, hypertension, and smoking) contribute to CVD development. Investigating molecular correlates of these traits can provide insights into the biological processes underlying CVD. To explore the relationship between altered DNAm and these risk factors, I queried the EWAS catalogue, a resource that curates results from published EWAS studies [170]. As no studies of dyslipidaemia were available, I instead searched for EWASs of total cholesterol and HDL cholesterol. These traits are commonly assessed in cohort studies and are often incorporated into CVD risk calculators. **Table 2** lists the search terms used for each trait. To reduce the number of search terms, all data (including the catalogue) were converted to lower case letters prior to analysis.

Table 2. Search terms used in the EWAS Catalog for Standard Modifiable Risk Factors (SMuRFs) of CVD.

Risk factor	*Search terms*
Total cholesterol	Total cholesterol, Serum total cholesterol
HDL cholesterol	HDL cholesterol, High-density lipoprotein cholesterol, Serum high-density lipoprotein cholesterol, Cholesterol esters in small HDL, Cholesterol esters in medium HDL, Cholesterol esters in large HDL, Cholesterol esters in very large HDL, Concentration of small HDL particles, Concentration of medium HDL particles, Concentration of large HDL particles, Concentration of very large HDL particles, Free cholesterol in small HDL, Free cholesterol in medium HDL, Free cholesterol in large HDL, Free cholesterol in very large HDL, Total cholesterol in HDL, Total cholesterol in HDL2, Total cholesterol in HDL3, Total cholesterol in large HDL
Hypertension	Hypertension, Blood pressure
Diabetes	Type 1 diabetes, Type I diabetes, Type 2 diabetes, Type II diabetes
Smoking	Smoking, Smoking pack-years, Tobacco smoking, Tobacco use

EWASs were then filtered to sample size greater than 1000 and to DNAm measured in either whole blood or in white blood cells. Nineteen unique publications matched this search criterion: four for both total and HDL cholesterol, four for type 2 diabetes, two for blood pressure, and nine for smoking (**Table 3**). I added to these results one smoking EWAS that was not listed in the EWAS catalog but was identified based on my knowledge of the literature [171]. No studies matched the search criteria for type 1 diabetes.

Consistent associations were observed for both HDL and total cholesterol [172–175]. The most robust DNAm signal was at cg06500161 in *ABCG1*, an ATP-binding cassette gene involved in reverse cholesterol transport. This CpG was strongly and repeatedly linked to reduced *ABCG1* transcript levels, lower HDL cholesterol , and increased risk of MI-related hospitalisation [172–174]. A second CpG in *ABCG1* (cg27243685) was also associated with incident CVD risk (HR per SD=1.38, $P=6.9x10^{-4}$) [175]. Cg16000331 in *SREBF2* (a transcription factor that controls cholesterol synthesis) was associated with total cholesterol [174]. Meta-analyses of the Registre Gironí del COR and Framingham Heart Study (FHS) Offspring cohorts showed that DNAm explained nearly 11% of the variance in HDL cholesterol (with approximately 5% attributable to cg06500161 alone) and up to 4% for total cholesterol [174].

CpGs near *ABCG1* and *SREBF2* (among others) were also consistently associated with future type 2 diabetes incidence across individuals of multiple ancestries, independent of traditional risk factors such as age, sex and BMI [176–178]. Meta-analyses in European cohorts identified a panel of six CpGs annotated to genes mainly implicated in glucose and lipid metabolism (*TXNIP, ABCG1, CPT1A, HDAC4, SYNM,* and *MIR23A*), which together explained 11% of the variance in type 2 diabetes risk [179]. Longitudinal studies further demonstrated that methylation at these loci precedes disease onset and may influence risk through glucose- and obesity-related pathways, underscoring their potential role in type 2 diabetes aetiology [177,178]. Overall, large EWASs show that DNAm at a limited number of CpG sites can serve as robust, reproducible markers of type 2 diabetes risk.

Two large cohort studies investigated DNAm in relation to blood pressure [180,181]. The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium meta-analysis of 17,010 individuals of diverse ancestries identified 16 replicable loci associated with systolic and diastolic blood pressure. These CpGs explained 1.4% and 2.0% of the variance in systolic and diastolic blood pressure, respectively, beyond age, sex, and BMI [180]. A combined meta-analysis of the discovery and replication sets identified 126 CpG sites associated with blood pressure. Mendelian randomisation (MR) suggested a causal role for cg08035323 (*TAF1B-YWHAQ*, a genomic region involved in endothelial function and vascular tone) in blood pressure regulation, while blood pressure itself appeared to influence methylation at CpGs annotated to *ZMIZ1*, *CPT1A*, and *SLC1A*. A subsequent meta-analysis of 4,820 individuals of European and African ancestry identified 39 CpGs associated with blood pressure, 16 of which replicated in CHARGE [181]. Conversely, 21 of the 126 sites reported by CHARGE were validated in this study. In total, 34 CpGs were cross-validated, several of which showed links to gene expression.

In my projects, I focus in particular on smoking, a trait that leaves a profound mark on the epigenome. In 2013, Zeilinger *et al*. conducted one of the first large-scale EWAS of current, former, and never smokers, analysing whole-blood DNAm using the Illumina 450K BeadChip in 1,793 participants of the Cooperative Health Research in the Region of Augsburg (KORA) F4 panel, with replication in 479 individuals from KORA F3 [182]. Comparison of current versus never smokers revealed widespread hypomethylation, with the strongest signal at cg05575921 in *AHRR*, a key repressor in the aryl hydrocarbon receptor pathway involved in detoxifying polycyclic aromatic hydrocarbons from tobacco smoke. Additional significant loci included cg21566642 near *ALPPL2* (part of the alkaline phosphatase gene cluster *ALPPL2, ALPP, ALPI*), cg03636183 in *F2RL3* (linked to CV function and mortality), cg19859270 in *GPR15* (immune regulation), and multiple CpGs in the 6p21.33 region (cg06126421, cg14753356, cg24859433, cg15342087), as well as *GNG12, GFI1, MYO1G, CNTNAP2, LRP5,* and *RARA*, highlighting smoking's broad epigenetic footprint across carcinogen metabolism, immune response, and CV pathways.

In 2017 Joehanes *et al*. conducted the largest adult EWAS of smoking, assessing blood-derived DNAm measured with Illumina 450k BeadChip in 15,907 participants from 16 cohorts (2,433 current, 6,518 former, 6,956 never smokers) [183]. Studied phenotypes included smoking status and, in a subset of three cohorts (n = 1,827), pack-years. Comparison of current versus never smokers identified 18,760 significant CpGs at False Discovery Rate (FDR)<0.05, largely replicating previous studies. Of these 18,760 CpGs, 11,267 (60.1%) showed significant dose-response relationships in pack-years analysis. Comparing former versus never smokers revealed 2,568 significant CpGs, with 185 overlapping current-never results, indicating that many smoking-induced methylation changes persist after cessation, albeit generally attenuated. Most CpGs reverted toward never-smoker levels within five years of quitting; however, 36 CpGs annotated to 19 genes – including *AHRR, F2RL3*, and *PRSS23* – remained altered even after 30 years of cessation, demonstrating the long-lasting epigenetic impact of tobacco exposure. The reversibility of DNAm changes associated with smoking was further studied by Dugué *et al.* using data from Melbourne Collaborative Cohort Study (n = 5,044 adults, with repeat measures in 1,032 participants after a median of 11 years) [184]. Their longitudinal analyses replicated persistent methylation at the 36 loci reported by Joehanes *et al*. [183] and revealed 368 CpGs with dynamic changes after cessation.

Several studies explored smoking-associated DNAm across diverse ages, populations and tissues [185–187]. Sikdar *et al*. compared adult methylation signatures from the Joehanes EWAS (n=15,907) with those in 5,648 newborns from nine cohorts exposed to maternal smoking *in utero* (897 exposed), identifying both shared and newborn-specific CpGs, with the latter enriched for xenobiotic metabolism pathways [185]. Marzi *et al*. performed an EWAS of smoking pack-years in 18-year-olds and identified 83 significant associations, largely consistent with the findings of Joehanes *et al*.[186]. Barcelona *et al*. examined smoking in African Americans using the Illumina EPIC 850K BeadChip, with DNAm measured in saliva in the InterGEN discovery cohort (n=156 females) and in blood in the Genetic Epidemiology Network of Arteriopathy (GENOA) replication sample (n=1100 individuals) [187]. Despite the limited discovery sample size, after adjusting for age, BMI, population structure, and cell composition, 26 significant associations were identified. Of these, six novel sites replicated in GENOA blood samples. These sites mapped to *RARA, FSIP1, ALPP, PIK3R5, KIAA0087,* and *MGAT3*, highlighting disease-relevant epigenetic effects of smoking on CV and cancer-related pathways.

Findings from Dogan *et al.* highlighted the importance of genomic context in assessing smoking-associated DNAm changes [188]. They conducted two EWASs using data from the FHS (n = 2,406, DNAm measured with Illumina 450k array): in the first, methylation at each CpG was regressed on smoking status, controlling for age, sex, and batch; in the second, an interaction term (SNP × smoking status) was included. Cis effects were defined as SNPs within 1 Mb of the CpG, and trans effects as SNPs beyond 1 Mb. The first EWAS identified 525 CpGs mapping to 310 unique genes, whereas the second revealed cis- and trans-interaction effects at CpGs mapping to 266 and 4,353 genes, respectively. Many loci with significant interaction effects were previously associated with complex diseases.

The first large-scale EWAS of smoking using the Illumina EPIC array was conducted in 2020 by Domingo-Relloso *et al*. using data of 2,325 Strong Heart Study (SHS) participants (n=790,026 CpG sites) [189]. This study replicated top findings of the Joehanes EWAS and uncovered numerous associations with CpGs not measured by Illumina 450k array. The authors identified 288 CpGs associated with current smoking (149 not measured by 450k array), 17 with former smoking (5 not measured by 450k array), and 77 with pack-years (29 not measured by 450k array). Novel associations included CpGs near *ZNF83* (a zinc finger protein), *PTPN1* (implicated in oncogenic transformations), and *RAB32* (a RAS family member strongly overexpressed in pancreatic cancer). A year later, Christiansen *et al*. extended this work by performing an EPIC-array-based meta-analysis across four UK cohorts (n = 1,407) with replication in 3,425 trans-ethnic samples, including American Indian and African-American participants from SHS and GENOA studies [190]. They reported 952 CpGs differentially methylated between smokers and never-smokers, of which 526 were EPIC-exclusive, with 92% replicating in independent cohorts.

To my knowledge, the largest EWAS of smoking to date using blood DNAm measured with the Illumina EPIC array is a meta-analysis of 15,014 individuals across five cohorts, conducted by Hoang *et al.* [171]. The study investigated current smoking (n=2,560), recent quitting within the past year (n=500), *in utero* exposure (n=286), and environmental tobacco smoke (n=676), and additionally evaluated interactions of current smoking with sex and dietary intake (fibre, folate, and vitamin C). The analysis identified 65,857 CpGs associated with current smoking, 4,025 with recent quitting, 594 with *in utero* exposure, and 6 with environmental tobacco smoke (all significant at FDR<0.05). Most CpGs linked to current smoking reverted within a year of quitting, whereas those associated with *in utero* exposure persisted into adulthood and were enriched for sites previously observed in newborns. A subset of CpGs (4 – 71) showed modification by sex or diet. Notably, CpGs related to current and *in utero* smoking mapped to 3,049 and 1,067 druggable targets, respectively, including chemotherapy drugs, highlighting potential translational insights into cancer treatment response and shared mechanisms across smoking-related diseases.

Table 3. Epigenome-Wide Association Studies of Standard Modifiable Risk Factors (SMuRFs) of CVD.

Ages are given in years.

SMuRF	Author (year)	N	Cohort	Age (% female)	DNAm platform	Association number
HDL cholesterol, Total cholesterol	Pfeiffer *et al.*, (2015) [172]	Discovery: 1,776 Replication: 1,827	Discovery: KORA F4 Replication: KORA F3, InCHIANTI, MuTHER	Discovery: mean age: 61 (51%) Replication: mean age range: 53 – 71 (46% - 100%)	Illumina 450k array	HDL cholesterol: 1 (1 replicated)
HDL cholesterol, Total cholesterol	Sayols-Baixeras *et al.*, (2016) [174]	Discovery: 645 Replication: 2,542	Discovery: REGICOR Replication: FHS Offspring	Discovery: mean age: 63 (51%) Replication: mean age: 66 (54%)	Illumina 450k array	HDL cholesterol: 39 (3 replicated), Total cholesterol: 16 (1 replicated) ($P<1x10^{-5}$)
HDL cholesterol, Total cholesterol	Braun *et al.*, (2017) [173]	Discovery: 725 Replication: 760 Meta-analysis: 1,485	Discovery: RS Replication: RS	Discovery: mean age: 60 (54%) Replication: mean age: 68 (58%)	Illumina 450k array	HDL cholesterol: 3 (2 replicated) Meta-analysis: HDL cholesterol: 55, Total cholesterol: 4

HDL cholesterol, Total cholesterol	Hedman *et al.*, (2017) [175]	Discovery: max 2,306 Replication: max 2,025	Discovery: FHS, PIVUS Replication: LBC1921, LBC1936, GOLDN	Discovery: mean age range: 66 – 70 Replication: mean age range: 49 – 79	Illumina 450k array	HDL cholesterol: 14 (11 replicated), Total cholesterol: 32 (5 replicated)
Type 2 diabetes	Chambers *et al.*, (2015) [176]	Discovery: 2,664 Replication: 1,141	Discovery: LOLIPOP Replication: LOLIPOP, KORA S3, KORA S4	Discovery: mean age: 51 (32%) Replication: mean age range: 58 – 61 (26% - 46%)	Discovery: Illumina 450k array Replication: Pyrosequencing (LOLIPOP), Illumina 450k array (KORA)	7 (5 replicated)
Type 2 diabetes	Cardona *et al.*, (2019) [177]	Discovery: 1,264 Replication: 5,271	Discovery: EPIC-Norfolk Replication: LOLIPOP, FHS	Discovery: mean age: 60 (70%) Replication: mean age range: 50 – 69 (32% - 54%)	Illumina 450k array	18

Type 2 diabetes	Juvinao-Quintero *et al.*, (2021) [179]	3,428	ALSPAC, LBC1936, RSIII-1, RS-Bios	Mean age: 62 (56%)	Illumina 450k array	6
Type 2 diabetes	Hillary *et al.*, (2023) [178]	max 18,413	GS	Mean age: 48 (59%)	Illumina EPIC array	58
Blood pressure	Richard *et al.*, (2017) [180]	Discovery: 9,828 Replication: 7,182	Discovery: CHARGE Replication: CHARGE	Discovery: mean age range: 49 – 76 Replication: mean age range: 46 – 68	Illumina 450k array	31 (13 replicated)
Blood pressure	Huang *et al.*, (2020) [181]	Discovery: 4,820 Replication: 17,010	Discovery: BHS, GSH, DILGOM, ETS, EGCUT (Asthma, Young_Old), FTC, HBCS, JHS, Lifelines, NTR, PREVEND, YFS, EpiGO, LACHY Replication: CHARGE	Discovery: mean age range: 14 – 69 (0% – 66%) Ages not reported for replication cohort	Illumina 450k array	39 (16 replicated)

Smoking	Zeilinger *et al.*, (2013) [182]	Discovery: 1,793 Replication: 479	Discovery: KORA F4 Replication: KORA F3	Discovery: mean age range: 57 – 62 (40% - 65%) Replication: mean age: 53 (50%)	Illumina 450k array	972 (187 replicated)
Smoking	Joehanes *et al.*, (2017) [183]	15,907	CHARGE	Mean age range: 58 – 65 (44% - 68%)	Illumina 450k array	Current vs. never: 2,623 (Bonferroni), 18,760 (FDR) Former vs. never: 185 (Bonferroni), 2,568 (FDR)
Smoking	Dogan *et al.*, (2017) [188]	1,597	FHS Offspring	Mean age: 67 (55%)	Illumina 450k array	525
Smoking	Marzi *et al.*, (2018) [186]	1,658	E-Risk	Mean age: 18	Illumina 450k array	83
Smoking	Barcelona *et al.*, (2019) [187]	Discovery: 156 Replication: 1,100	Discovery: InterGEN Replication: GENOA	Discovery: mean age: 32 (100%) Ages not reported for replication cohort	Illumina EPIC array	26 (6 novel loci replicated)

Smoking	Sikdar *et al.*, (2019) [185]	Adults: 15,907, Newborns: 5,648	CHARGE, PACE	Adults: as in Joehanes *et al.*, (2017)	Illumina 450k array	Adults: 34,541 Newborns: 5,547
Smoking	Dugué *et al.*, (2020) [184]	5,044	MCCS	Mean age: 61 (32%)	Illumina 450k array	Current vs. never: 1,851, Former vs. never: 156 Smoking index: 4,496
Smoking	Domingo-Relloso *et al.*, (2020) [189]	2,325	SHS	Mean age: 55 (59%)	Illumina EPIC array	Current vs. never: 288, Former vs. never: 17, Pack-years: 77
Smoking	Christiansen *et al.*, (2021) [190]	Discovery: 1,407 Replication: 3,425	Discovery: TwinsUK, NSHD, NCDS1, NCDS2, BCS70 Replication: SHS, GENOA	Discovery: mean age range: 45 – 64 (51% - 100%) Replication: mean age range: 55 – 56 (59% - 71%)	Illumina EPIC array	Current vs. never: 952 (389 replicated)

Smoking	Hoang *et al.,* (2024) [171]	max 15,014	START, ALHS, GS, SHS	Median age range: 33 – 62 (47% – 62%)	Illumina EPIC array	Current vs. never: 65,857, Former vs. never: 4,025, *In utero*: 594, Environmental exposure: 6 (all at FDR<0.05)

ALHS indicates Agricultural Lung Health Study; ALSPAC, Avon Longitudinal Study of Parents and Children; BCS70, 1970 British Cohort Study; BHS, Bogalusa Heart Study; CHARGE, Cohorts for Heart and Aging Research in Genomic Epidemiology; DILGOM, the Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome Study; E-Risk, Environmental Risk Longitudinal Twin Study; EGCUT, Estonian Genome Center of the University of Tartu; EPIC, European Prospective Investigation into Cancer; EpiGO, the Epigenetic basis of Obesity induced cardiovascular disease and type 2 diabetes study; ETS, the Emory Twin Study; FHS, Framingham Heart Study; FTC, the Finnish Twin Cohort; GENOA, Genetic Epidemiology Network of Arteriopathy; GOLDN, Genetics of Lipid Lowering Drugs and Diet Network; GS, Generation Scotland, GSH, the Georgia Stress and Heart study; HBCS, the Helsinki Birth Cohort Study; InCHIANTI, Invecchiare in Chianti, Aging in the Chianti Area; InterGEN, the Intergenerational Impact of Genetic and Psychological Factors on Blood Pressure Study; JHS, Jackson Heart Study; KORA, Cooperative Health Research in the Region of Augsburg; LACHY, the Lifestyle, Adiposity, and Cardiovascular Health in Youth study; LBC1921, Lothian Birth Cohort 1921; LBC1936, Lothian Birth Cohort 1936; Lifelines, the Lifelines Cohort Study; LOLIPOP, London Life Sciences Prospective Population Study; MCCS, Melbourne Collaborative Cohort Study; MuTHER, Multiple Tissue Human Expression Resource; NCDS, National Child Development Study (NCDS1 = selected to minimise data missingness, but not selected for specific exposures and trait outcomes. NCDS2 = selected for extremes of child and adulthood adversity); NSHD, MRC National Survey of Health and Development; NTR, the Netherlands Twin Register; PACE, the Pregnancy And Childhood Epigenetics; PIVUS, Prospective Investigation of the Vasculature in Uppsala Seniors; PREVEND, Prevention of Renal and Vascular End stage Disease study; REGICOR, Registre Gironí del COR; RS, Rotterdam Study (with sub-cohorts: RS-Bios and RSIII-1); SHS, Strong Heart Study; START, Study of Assisted Reproductive Technology (a sub-study of the Norwegian Mother, Father, and Child Cohort Study); TwinsUK, TwinsUK Study; YFS, the Young Finns Study.

2.3.4. EpiScores

Information from multiple CpG sites can be combined into composite risk scores, known as epigenetic scores (EpiScores), which estimate phenotypes or disease risk based on the additive, weighted contribution of each site [1]. Analogous to polygenic risk scores, EpiScores are typically derived using EWAS and penalised regression approaches [1,191–193], and can be developed for a wide range of traits [155]. Because DNAm at many CpG sites changes gradually, EpiScores are often more stable over time than conventional biomarkers and can capture health information extending years into the past [151]. They are particularly valuable for exposures well captured by DNAm, such as cigarette smoking, where existing biomarkers are limited.

Smoking EpiScore

While the information about tobacco use is routinely incorporated into clinical risk prediction tools [114], these calculators typically rely on self-reported smoking status or the number of cigarettes smoked per day. However, self-reported data are susceptible to recall bias and underreporting due to social desirability, and may not accurately reflect actual exposure [194]. They also do not capture information about second-hand smoking. Serum cotinine, a metabolite of nicotine, offers a more objective measure of recent tobacco use [194]. Nevertheless, its utility is limited by a short half-life (approximately 15-19 hours in plasma) [195], making it unsuitable for assessing long-term exposure or time since cessation. As a result, cotinine levels cannot reliably distinguish former smokers from never smokers – an important limitation when evaluating risk for chronic conditions such as CVD, which develop over extended periods.

The limitations of traditional methods for assessing smoking exposure have motivated the development of novel molecular biomarkers. A summary of both established and novel biomarkers of smoking is presented in **Table 4**. Among these, blood-based DNAm biomarkers have emerged as particularly promising tools for capturing both current and historical tobacco exposure. Smoking EpiScores provide nuanced estimates of exposure intensity, duration, and cumulative burden. These include:

- BayesR-based scores, which use Bayesian regression to model complex methylation patterns [168];
- EpiSmokEr, a smoking status estimator that categorises individuals into current, former, or never smokers based on a panel of CpGs [196];
- The pack-years DNAm score developed by McCartney *et al.*, which quantifies lifetime exposure using elastic net regression [155];
- And the DNAm pack-years estimate embedded in the GrimAge epigenetic clock, which was trained to predict smoking pack-years and contributes to accurate mortality prediction [197].

Table 4. Biomarkers of smoking. DNAm – DNA methylation. CpG site – cytosine-phosphate-guanine site.

Biomarker	*Type*	*Advantages*	*Limitations*	*Predictive Performance*	*Ref.*
Cotinine	Metabolite	- Widely used and validated - Detectable in blood, urine, saliva - Reflects recent exposure (1–3 days)	- Short half-life limits detection window - Cannot distinguish smoking from nicotine replacement therapy - Not useful for long-term exposure	High accuracy for distinguishing current and never smokers; cannot distinguish former from never smokers	198
AHRR methylation	DNAm	- Reflects current and past smoking - Changes persist after cessation - Potential link to disease risk	- Not routinely measured in clinic - Cost may be higher than metabolite testing	High accuracy for distinguishing current and never smokers; lower accuracy for former vs. never smokers	199
EpiScores (e.g., EpiSmokEr, McCartney et al).	DNAm composite	- Aggregate multiple CpG sites - All advantages of DNAm biomarkers	- Not routinely measured in clinic - Some scores do not measure intensity of smoking - Limited use in non-Europeans	- Near perfect discrimination of current vs. never smokers; lower accuracy for former vs. never smokers	155,196
miRNAs (e.g., miR-21)	Non-coding RNA	- May reflect gene regulation changes - Potential for early disease signals	- Not routinely measured in clinic - Limited validation - Limited use for long-term exposure	Promising, but performance data limited	200
Self-report	Survey-based	- Low-cost and easy to obtain - Can provide behavioural context	- Subject to recall bias and underreporting	Variable accuracy, data prone to bias; best when supported by biomarkers	201

It is important to mention that DNAm is inherently tissue-specific, and signals identified in one tissue may not generalise to others. While CpG sites in genes such as *AHRR* and *F2RL3* are among the most robust and reproducible markers of smoking exposure in blood, these same loci may not display comparable methylation patterns in other tissues, including the brain. This tissue specificity presents a significant challenge to the universal application of DNAm biomarkers, particularly in conditions that involve the central nervous system, where relevant epigenetic changes may not be captured through blood-based profiling alone.

Nonetheless, accumulating evidence suggests that blood-based DNAm signatures can serve as accessible proxies for systemic effects of smoking, including vascular damage that impacts both the heart and brain. As such, they hold promise for enhancing risk prediction in both CV (and neurodegenerative) diseases. Continued research is needed to understand the tissue specificity of these markers and to validate their clinical utility in diverse populations and disease contexts. In the next subsection, I will discuss EpiScores for protein levels. Proteomic biomarkers of CVD have been discussed in **Section 2.4.2**.

Protein EpiScores

As outlined in **Section 2.4**, most established biomarkers of incident disease are proteins. However, the concentrations of several widely used protein biomarkers – such as lipoproteins [202] and C-reactive protein (CRP) [203] – can fluctuate in response to transient influences including circadian rhythm, dietary intake, and short-term inflammatory responses. Such variability may limit their reliability for long-term risk prediction. Protein EpiScores address these limitations. Similar to EpiScores of environmental exposures, they aim to capture more stable, cumulative biological signals [191]. Conceptually, they resemble glycated haemoglobin (HbA1c), which reflects average blood glucose levels over time and is widely used in diabetes management [204].

The relationship between CRP and its EpiScore provides a clear illustration. In 2020, Stevenson *et al*. constructed a CRP EpiScore using weights from a previously published EWAS [205] in two cohorts: Generation Scotland (GS, n=7028) and the Lothian Birth Cohort 1936 (LBC1936, n=889) [191]. While serum CRP showed no consistent age-related trajectories across cohorts, the CRP EpiScore increased with age in both (standardised β=0.07 in LBC1936; β=0.01 in GS). The EpiScore also demonstrated greater temporal stability than serum CRP, with inter-wave correlations in LBC1936 ranging from 0.53 to 0.75 compared with 0.30 to 0.40 for the measured protein.

Two years later, Wielscher *et al.* published a multi-ethnic EWAS study of serum CRP (n=22,774 participants, 30 independent cohorts), in which they investigated the causal role of DNAm on CRP levels [206]. Their results suggested that altered CpG methylation is more likely a consequence of elevated CRP rather than a causal factor, with smoking and obesity being underlying driving factors for altered methylation signature. They also developed a CRP EpiScore, which was associated with increased cardiometabolic risk: a one percent increase in the score corresponded to a 2.9% higher risk of MI ($P=1.7x10^{-3}$), 4.3% higher risk of CAD ($P=9.8x10^{-3}$), and 0.2% higher risk of hypertension ($P=9.8x10^{-22}$).

Building on these findings, Hillary *et al.* trained an enhanced CRP EpiScore in GS (n=17,936) using three machine-learning approaches and evaluated it across five independent cohorts of European and Asian ancestry spanning different life stages [207]. In LBC1936, the best-performing score explained 18% of the variance in serum CRP (n=756) after adjusting for age and sex – approximately doubling the predictive performance of earlier scores. Importantly, the EpiScore performed well across the adult life course and outperformed both assay-measured CRP and a genetic score in associations with 26 cardiometabolic outcomes. For example, the EpiScore was associated with a history of CVD (odds ratio=1.28, $P_{FDR} < 0.05$), whereas serum CRP was not (odds ratio=1.00, $P_{FDR}=0.97$).

EpiScores have also been developed for other inflammatory proteins and CVD biomarkers [193,208,209]. For example, an EpiScore for interleukin-6 (IL6) was trained in LBC1936 (n=875) and tested in GS ($n_{measured}$=417, n_{DNAm}=7028) [193]. Both the serum IL6 (n=417, standardised β=0.02, P=1.3x10^{-7}) and the EpiScore (n=7028, β=0.02, P<2x10^{-16}) increased with chronological age and were significantly associated with CVD risk factors, such as BMI (β=0.09, P_{FDR}=1.3x10^{-7}), current smoking (odds per SD of the EpiScore=1.2, P_{FDR}<2x10^{-16}) and SIMD (β=−0.24, P_{FDR}=<2x10^{-16}). In another study, which I contributed to the analysis, Gadd *et al.* analysed epigenetic signals underlying increased concentrations of two CV biomarkers: GDF15 and NT-proBNP [209]. We conducted an EWAS of the studied proteins and trained EpiScores in more than 16,963 individuals from GS. Both scores associated with a range of traits affecting body and the brain. While the GDF15 EpiScore replicated protein associations with type 2 diabetes and ischaemic stroke in the GS test set (n ≥ 2808, HR range 1.36-1.41, P_{FDR} < 0.05). The EpiScore for NT-proBNP replicated the protein association with type 2 diabetes (HR=0.73, P=4.7x10^{-5}), but failed to replicate an association with ischaemic stroke.

In 2022, Gadd *et al.* developed multiple protein EpiScores in a single study using affinity protein panels (Olink and SomaScan, see **Section 2.4.1.**) [1]. They trained EpiScores for 953 plasma proteins (706≤ $n_{training}$ ≤ 944 individuals) and validated them in two independent cohorts (162≤$n_{validation}$ ≤778 individuals). A total of 109 EpiScores demonstrating robust performance (Pearson's r>0.1 and P<0.05) were selected, projected into GS, and examined for incident associations with 12 diseases over 14 years of follow-up. In fully adjusted Cox proportional hazards (PH) models, eight EpiScores were significantly associated with stroke and 13 with IHD. EpiScores associated with stroke included CD209 ($HR_{per\ SD\ of\ the\ EpiScore}$=0.82, P=3.7x10^{-3}), SELE (HR=1.24, P=6.7x10^{-3}) and FGF21 (HR=1.33, P= 2.3x10^{-4}), while associations with IHD included SELL (HR=0.80, P=6.9x10^{-3}), ENPP7 (HR=1.26, P= 3.2x10^{-4}), and SELE (HR=1.27, P= 1.9x10^{-3}). The consistency of disease associations between EpiScores and measured proteins was assessed using diabetes as an exemplar. This analysis identified 34 EpiScore-disease associations, 28 of which had been previously reported for measured protein concentrations, suggesting that EpiScores are able to capture links between proteins and disease. Six associations were novel and may represent previously unreported protein–diabetes relationships or associations not captured by protein measurements alone.

Several studies used Gadd EpiScores as proxies for protein biomarkers [208,210,211]. For example, Waterfield *et al.* evaluated a subset of these scores in the ALSPAC cohort of white European participants, assessing their ability to predict Olink-measured protein abundances at ages nine (n = 222), 24 (n = 763), and midlife (n = 622) [211]. Predictive performance was highest in midlife, where 10 of 14 EpiScores were significantly associated with protein abundance (Pearson's r > 0.1, P < 0.05), likely reflecting the demographic similarity to the training data. At age 24, nine EpiScores remained correlated, whereas at age nine none showed significant associations. The

authors subsequently trained novel EpiScores, which transferred across age groups, and assessed the performance of both newly derived and Gadd's scores in models adjusted for genetic effects. In adulthood, most EpiScores explained additional variance beyond polygenic risk scores. For the Gadd EpiScores, this outcome is expected, as the authors' training strategy involved regressing out SNPs known to be associated with protein levels before constructing the EpiScores. Among the strongest predictors was the OSM EpiScore, which explained >5% of protein variance beyond genetics and had previously been linked to IHD in the Gadd study [1].

Taken together, current evidence shows that protein EpiScores can replicate the associations between measured proteins and CVD outcomes. In some cases, they have a greater predictive performance than measured protein concentrations, likely due to a higher signal-to-noise ratio of the EpiScore and greater temporal stability. This stability is particularly important for CVD, where risk develops gradually over many years and short-term fluctuations in circulating protein levels may obscure long-term disease signals. A further advantage of EpiScores is their portability – they can be projected into any cohort with DNAm data. In cohorts with both DNAm and proteomic data, EpiScores can be combined with measured proteins to enhance CVD risk prediction.

2.4. Proteins

Proteins serve as key integrators of genetic, environmental, and lifestyle influences on disease risk. Their levels not only reflect the downstream effects of genomic variation but also capture dynamic physiological responses to external exposures. As such, they provide direct insight into the biological processes driving disease. Therefore, it is perhaps not surprising that most clinically established biomarkers belong to this category.

Proteomics – the large-scale study of proteins – aims to quantify and characterise the full complement of proteins expressed under specific physiological or pathological conditions. The human genome is estimated to encode just under 20,000 protein-coding genes (19,370 in GENCODE Release 41) [212], which give rise to a growing number of distinct protein isoforms – alternative versions of a protein generated from the same gene, typically through processes such as alternative splicing or post-translational modifications. Human blood plasma and serum offer accessible windows into the circulating proteome. Plasma is the clear, yellow fluid that remains after the removal of blood cells and platelets, while serum is derived from plasma but lacks clotting factors such as fibrinogen.

Both plasma and serum pose significant analytical challenges: a small number of highly abundant proteins can obscure the detection of lower-abundance species. Of the 4,608 canonical proteins (most common or most studied or functionally representative form) identified to date in plasma (according to the Human Plasma PeptideAtlas [213] – a compendium of uniformly processed mass spectrometry datasets; build 2023-04 based on 35,801 runs), just 22 proteins account for 99% of the total protein mass in both plasma and serum [214]. The plasma proteome spans an exceptionally wide dynamic range – between 9 and 13 orders of magnitude [215]. **Table 5** outlines the approaches used in proteomic studies along with their respective detection ranges. The technologies used for protein quantification are further discussed in the following section.

Table 5. Three common proteomics approaches: discovery (shotgun), targeted discovery, and targeted proteomics. Table adapted from Lam *et al*. [216]

	Discovery Approach	***Targeted Discovery Approach***	***Targeted Approach***
Proteome Coverage	High (up to ~10,000 proteins analysed in a single experiment)	Medium (depends on panel/array size)	Low (typically 10s of proteins)
Example Technology	Mass spectrometry	Antibody/aptamer arrays	Enzyme-linked immunosorbent assay (ELISA)
Advantages	Most unbiased	Middle ground between sensitivity and scope	Excellent quantification precision
Disadvantages	Quantification may be less precise than in targeted approaches	Target may be unavailable in commercial panels	Low discovery capability
Sample Throughput	Low – Medium	High	Low
Sensitivity Limit	µg/ml – ng/ml	ng/ml – pg/ml	ng/ml – pg/ml

2.4.1. Protein Quantification Methods

Protein levels are commonly quantified using immunoassays, affinity-based platforms, or mass spectrometry (MS). In the following sections, I outline the principles of each approach, with particular emphasis on Olink and SomaScan – two affinity-based platforms widely applied in large-scale cohort studies – and on MS, which is used for protein quantification in the empirical analyses presented in this thesis.

Enzyme-Linked Immunosorbent Assay (ELISA)

The accepted gold standard for single-protein quantification is ELISA, an immunoassay which exploits the high specificity of antibodies for their target antigens [217]. In sandwich ELISA (the most common type of the assay) a plate is first coated with a capture antibody specific to the target protein. Next, the sample is added, allowing any present antigen to bind to the immobilised capture antibody. After washing, a detecting antibody is introduced, which binds to a different site on the captured antigen. An enzyme-linked secondary antibody is then added to recognise the detecting antibody. Finally, the substrate is added and is converted by the enzyme into a measurable colour or light signal.

ELISA assays are highly specific, relatively affordable, reproducible, and straightforward to perform. They require only standard laboratory equipment and are capable of detecting low concentrations of proteins – often in the picogram to nanogram range – depending on the assay design and antibody quality. This makes ELISAs a reliable choice for both research and clinical diagnostics. However, the method has notable limitations. ELISAs typically measure a single protein per assay, limiting their efficiency in large-scale proteomic studies. Furthermore, ELISAs are not well-suited for biomarker discovery, as they require prior knowledge of the target protein.

Affinity-Based Platforms: Olink and SomaScan

Affinity-based proteomic platforms enable the high-throughput quantification of proteins by leveraging the specific binding interactions between target proteins and affinity reagents, such as antibodies or aptamers. These technologies offer a scalable alternative to traditional immunoassays, with the capacity to simultaneously measure thousands of proteins. Two leading affinity-based platforms used in large-scale proteomic studies are Olink (Olink Proteomics, Uppsala, Sweden) and SomaScan (SomaLogic Operating Co., Inc), each employing a distinct detection principle.

Olink is based on proximity extension assay (PEA) technology, which uses pairs of antibodies labelled with unique DNA oligonucleotides [218]. When both antibodies bind to their specific epitopes on a target protein, the attached oligonucleotides are brought into close proximity, allowing them to hybridise and be enzymatically extended to form a new DNA sequence unique for each protein (a barcode). This barcode is then amplified and quantified using real-time polymerase chain reaction or NGS. Because amplification only occurs when both antibodies are correctly bound, PEAs are highly specific. Olink's panels currently enable the multiplexed measurement of up to 5,420 proteins (Olink Explore HT, status for July 2025) [219,220].

SomaScan, developed by SomaLogic, uses a different approach based on slow off-rate modified aptamers (SOMAmers) – short, chemically modified single-stranded DNA molecules engineered to bind target proteins with high specificity [221]. These modifications enhance binding strength and slow the rate at which the aptamers dissociate from their targets, reducing nonspecific interactions. In the SomaScan assay [222], each protein in the sample binds to its corresponding SOMAmer in a highly controlled environment. Unbound proteins and non-specific interactions are removed through a series of stringent washing steps that exploit the slow off-rate properties of the aptamers. The bound SOMAmers are then released and quantified (by hybridisation to a complementary DNA microarray or via NGS). SomaScan 11K Assay v 5.0 supports the detection of over 11,037 proteins across a wide dynamic range [223].

Mass Spectrometry

MS is widely regarded as the gold standard for large-scale, unbiased protein identification and quantification [224]. It works by measuring the mass-to-charge ratio (m/z) of ionised molecules. In a typical workflow, before the MS analysis begins, proteins extracted from biological samples (plasma, tissue) cells are denatured and digested – most often using the enzyme trypsin – into peptides with predictable properties. These peptides are then separated by liquid chromatography (LC), which helps reduce sample complexity and spreads peptides out over time before they enter the mass spectrometer [224].

MS analysis involves four main steps: ionisation, acceleration, deflection, and detection [225–227]. During ionisation, peptides are converted into charged particles in the gas phase. The ions are then accelerated so they all reach the same kinetic energy. When they enter a magnetic or electric field, ions are deflected based on their mass-to-charge ratio – lighter or more highly charged ions are deflected more than heavier ones. By adjusting the field strength, the instrument directs ions to the detector in a controlled way. When ions hit the detector, they are neutralised, causing a flow of electrons that produces an electrical signal. This signal is recorded as a mass spectrum – a chart where each peak represents a specific m/z value and its intensity reflects the number of ions detected.

To identify proteins, MS instruments often operate in tandem MS (MS/MS) mode [227,228]. In this approach, a first scan (MS1) detects all intact peptide ions and measures their mass-to-charge (m/z) ratios. Selected peptide ions – referred to as precursor ions – are then isolated based on their m/z values. These isolated ions are directed into a collision cell, where they are fragmented into smaller, sequence-specific product ions. A second scan (MS2) then measures the m/z values of these fragment ions, generating a fragmentation spectrum that reveals information about the peptide's amino acid sequence. The resulting MS/MS spectra are computationally matched to theoretical spectra from protein sequence databases, allowing peptide identification. Each peptide is then mapped back to one or more parent proteins. While some peptides are unique to a single protein, many are shared across proteins, such as isoforms or members of protein families. If all identified peptides are common to multiple proteins, the software cannot determine which specific protein is present. These indistinguishable proteins are then grouped together into a protein group.

Mass spectrometry enables multiplexed detection of thousands of proteins in a single experiment without prior knowledge of the analytes. It quantifies protein abundance and detects post-translational modifications, isoforms, and cleavage products. However, it requires costly instrumentation, specialised expertise, and has lower sensitivity for low-abundance proteins compared to immunoassays [229]. Despite these limitations, MS remains the most comprehensive and versatile platform for proteome-wide analysis.

2.4.2. Proteomic Biomarkers of CVD

This section reviews recent studies that employed high-throughput proteomic technologies to identify CVD biomarkers in large, well-characterised population cohorts.

Several studies have examined associations between Olink-measured plasma protein concentrations and CV outcomes. Among the most comprehensive are analyses based on the UK Biobank Pharma Proteomics Project (UKB-PPP; n = 54,219). Using data from 47,600 individuals, Gadd *et al*. tested associations between 1,468 proteins and 23 age-related diseases, including CVD [230]. They identified 405 and 186 proteins significantly associated with the future onset of IHD and ischaemic stroke, respectively – up to 15 years before diagnosis and after adjustment for a wide range of demographic, lifestyle, and clinical factors. They also developed ProteinScores (proteomic analogues of EpiScores), which improved 10-year IHD risk prediction beyond an extended set of risk factors (ΔAUC = 0.027, P < 0.001). In a related study, Royer *et al*. applied extreme gradient boosting (n = 38,380; $n_{proteins}$ = 2,919) to derive a 114-protein panel that enhanced prediction of 10-year risk of a composite CVD outcome (defined as MI, ischaemic and haemorrhagic stroke, and a revascularisation procedure – a proxy for CAD) [231]. This panel significantly outperformed both SCORE2 (ΔAUC = 0.029, P < 0.001) and a refitted SCORE2 model (ΔAUC = 0.016, P = 0.031).

Additional large-scale analyses have confirmed the predictive utility of plasma proteomics across ancestries. Using data from the UKB-PPP (n = 52,164; $n_{proteins}$ = 2,919), Lind *et al.* applied Cox PH regression to identify 126 proteins associated with a composite CVD outcome, defined as incident MI, ischaemic stroke, or HF [232]. Of these, 118 associations were successfully replicated in the China Kadoorie Biobank (CKB), supporting their generalisability across populations. MR and colocalisation analyses further suggested likely causal roles for a subset of these proteins, including FGF5, PROCR, and FURIN. In parallel, Mazidi *et al*. conducted a nested case–cohort study in CKB (n ≈ 4,000; including 1,976 IHD cases) to assess the relationship between Olink-measured proteins and incident IHD over a 12-year follow-up [233]. After adjusting for demographic, lifestyle, and clinical factors, they identified 446 proteins significantly associated with IHD. Adding these proteins to conventional risk models improved discrimination (ΔC-statistic=0.021). These findings were replicated in UKB-PPP, demonstrating robust and cross-population predictive performance.

In addition to Olink-based studies, several large-scale investigations have employed SomaLogic's SomaScan platform to study CVD risk. For example, Corlin *et al.* used SomaScan to measure 1,305 plasma proteins in 897 individuals from FHS Generation 3 (discovery sample) and 1121 individuals from FHS Offspring study (validation sample) to identify proteomic signatures of CVD risk factors [234]. After adjusting for age, sex, BMI, and family structure, and validating findings in FHS Offspring, they identified 37 proteins associated with smoking, 23 proteins associated with alcohol consumption, and 2 proteins associated with physical activity.

There is growing interest in modelling the risk of HF and stroke as distinct CV outcomes, as clinical risk stratification remains challenging. While some individuals with HF may ultimately require advanced interventions such as mechanical support or transplantation, others can be managed effectively with guideline-directed medical therapy. A community-based study in Southeast Minnesota used the SomaScan platform to measure 7,289 plasma proteins in 1,351 patients with established HF [235]. A protein risk score for 5-year mortality was developed using LASSO regression (see **Section 4.4.1**). The final 38-protein score demonstrated strong predictive performance, with good calibration and improved clinical utility compared to standard clinical risk scores, including the Meta-Analysis Global Group in Chronic Heart Failure score and NT-proBNP. Notably, the model showed particular value at the extremes of the risk spectrum, where clinical decision-making is often most uncertain.

Complementing this work, Girerd *et al.* integrated data from the Heart OMics in AGEing Study (HOMAGE), ARIC Study, and the FHS to identify plasma proteins predictive of new-onset HF [236]. Risk was modelled using logistic regression in a nested case-control design. Of the 276 Olink-measured proteins, 62 were associated with incident HF in ARIC, 16 in FHS, and 116 in HOMAGE, with eight proteins – BNP, NT-proBNP, 4E-BP1, HGF, Gal-9, TGF-α, THBS2, and uPAR – consistently associated across all three cohorts. Importantly, their multimarker model improved HF risk prediction beyond clinical risk factors and NT-proBNP, with C-index increases of 11.1% in ARIC, 5.9% in FHS, and 7.5% in HOMAGE (all $P < 0.001$), outperforming NT-proBNP alone.

Similarly, stroke-focused proteomic studies have demonstrated the potential to differentiate ischaemic from haemorrhagic stroke and to predict disease progression. A recent systematic review and meta-analysis of 112 studies (42 included in the meta-analysis) evaluated the diagnostic accuracy of non-coding RNA and protein biomarkers in over 11,000 ischaemic stroke patients, 2,100 haemorrhagic stroke patients, and nearly 7,000 controls [237]. Proteins such as IL-6 and S100B showed comparable diagnostic performance for ischaemic stroke. For differentiating stroke subtypes, GFAP and NR2aAb performed best, while no biomarkers reliably distinguished strokes from mimics.

Finally, MS studies aimed at identifying proteins associated with incident CVD remain relatively limited. This scarcity may be attributed to challenges such as high costs and technical complexity. Although conceptual papers have outlined the potential of MS for investigating CVD risk in large population cohorts, its application has so far been largely confined to the discovery of novel biomarkers in small datasets [238]. However, ongoing advancements in more affordable and scalable MS technologies [239] hold promise for expanding its use in large-scale epidemiological studies in the near future.

Taken together, these large-scale proteomic studies highlight the potential of circulating proteins as early, dynamic, and mechanistically informative biomarkers of CVD. To provide a clear overview of the current landscape, I cross-referenced findings from the above-mentioned analyses with review articles [216,240]. **Table 6** summarises clinically established biomarkers, while **Table 7** presents emerging candidates.

Table 6. Established CVD biomarkers. Table adapted from Lam *et al*. [216]

Biomarker	*Abbreviation*	*Disease*	*Assay Sensitivity*	*Discovery Period*	*Ref.*
Apolipoprotein A-I	APOA	CVD	~1 mg/ml	1980s	[241]
Apolipoprotein B	APOB	CVD	~1 mg/ml	1980s	[242]
C-reactive protein	CRP	CVD	~10 µg/ml	1990s	[243]
Creatine kinase-myocardial band	CKMB	MI	~1 ng/ml	1960s	[244]
Cystatin-C	CST3	CVD	~1 µg/ml	2000s	[245]
D-Dimer	D-Dimer	DVT, PE	~1 µg/ml	1980s	[246]
Fibrinogen	FBN	CVD	~1 mg/ml	1980s	[243]
Lipoprotein-associated phospholipase A2	Lp-PLA2	CHD	~100 ng/ml	2000s	[247]
Myeloperoxidase	MPO	IHD, ACS	~10 ng/ml	1980s	[248]
Myoglobin	MYO	MI	~10 ng/ml	1970s	[249]
N-terminal pro-B-type natriuretic peptide	NTproBNP	HF, ACS	~100 pg/ml	2000s	[250]
Serum amyloid A	SAA	CAD	~10 µg/ml	1990s	[251]
Troponin I	cTnI	MI	~10 pg/ml	1970s	[252]
Troponin T	cTnT	MI	~10 pg/ml	1970s	[253]

ACS – Acute Coronary Syndrome, CAD – Coronary Artery Disease, CHD – Coronary Heart Disease, CVD – Cardiovascular Disease, DVT – Deep Vein Thrombosis, HF – Heart Failure, IHD – Ischaemic Heart Disease, MI – Myocardial Infarction, PE – Pulmonary Embolism

Table 7. Selected emerging CVD biomarkers. These biomarkers are not yet standard in clinical practice but have shown strong and replicated associations with CVD and risk stratification in multiple studies.

Biomarker	*Overview*	*Ref*
H-FABP	Studies show H-FABP is either superior to, or adds value to, troponins in the early diagnosis of ACS. It may help identify high-risk patients earlier.	254–256
GDF-15	GDF-15 is a strong predictor of cardiovascular events and all-cause mortality. Clinical trials suggest utility for risk stratification.	257–259
PAPP-A	Circulating PAPP-A has been proposed as a promising biomarker for ACS risk stratification.	260,261
MMPs	MMP-2, MMP-8, and MMP-9 are implicated in plaque rupture and cardiovascular events.	262–264
sPLA2	Elevated sPLA2-IIA and total sPLA2 levels have been linked to increased CV risk, but clinical utility remains uncertain.	265,266
sCD40L	Some prospective studies suggest prognostic value of sCD40L, though findings are inconsistent.	267–269
Copeptin	Copeptin may predict CAD and cardiovascular death, though its tissue origin remains unclear.	270–272

MR-proADM	MR-proADM shows promise in HF risk prediction and early atherosclerotic changes including subclinical CAD.	273–275
ST2	ST2 has been validated for CV risk stratification.	276–279
ET-1 (CT-proET-1)	CT-proET-1 is associated with HF and cardiovascular death independent of clinical variables.	280–282
Gal-3	FDA-approved in 2010 for HF risk stratification; Gal-3 reflects fibrosis and inflammation.	240,283
NRG-1	Elevated NRG-1 is linked to HF and CAD, though clinical use requires further validation.	284,285
GFAP	Astrocyte-specific marker elevated after brain injury; helps differentiate ischaemic vs. haemorrhagic stroke and is associated with infarct volume, severity, and outcome. Best used with other markers.	286,287
S100B	Astrocyte-derived protein linked to neuroinflammation. Elevated in both ischaemic and haemorrhagic stroke but lacks specificity to aid in diagnosis.	288–290

3. Thesis Aims

The overarching aim of this work was to identify multi-omic biomarkers of CVD that provide insights into its biology and improve risk prediction beyond traditional risk factors.

To address this, I first focused on smoking, an important but challenging-to-measure component of CVD risk prediction. Previous EWASs have examined its biological effects and developed objective biomarkers of tobacco exposure. However, no study has investigated associations between DNAm and smoking pack-years in a single-cohort sample exceeding 10,000 individuals. Furthermore, it remains unclear whether DNAm changes observed in blood translate to the brain. Therefore, the first aim of **Chapter 5** is:

Aim 1: To identify CpG sites associated with smoking pack-years in blood and brain using array- and sequencing-based DNA methylation data (max n = 17,865).

Although DNAm biomarkers of smoking have been developed, none were trained in a sample exceeding 10,000 individuals. Therefore, the second aim of **Chapter 5** is:

Aim 2: To train a smoking EpiScore in >10,000 blood DNAm samples and compare its predictive performance with existing smoking EpiScores.

Finally, no study has directly compared the DNA signal (GWAS) from self-reported smoking with that captured by epigenetic predictors. Therefore, the third aim of **Chapter 5** is:

Aim 3: To compare the DNA signal of self-reported smoking with that of epigenetic smoking (based on a smoking EpiScore) via GWAS.

Next, I analysed protein EpiScores as potential biomarkers of CVD. No studies have systematically examined the relationship between protein EpiScores and incident CVD risk. It remains unclear whether protein EpiScores can improve risk prediction beyond established clinical scores. Therefore, the aim of **Chapter 6** is:

Aim 4: To test the added predictive value of 109 protein EpiScores for CVD in comparison with the clinical risk scores ASSIGN and SCORE2.

Finally, I explored proteins in the context of CVD risk prediction. No large-scale mass spectrometry study has been conducted to study CVD risk. Therefore, the aim of **Chapter 7** is:

Aim 5: To determine associations between the abundances of 439 proteins measured by mass spectrometry and CVD in 8,343 individuals from the GS cohort.

In the next chapter, I will describe the cohorts used in this work and outline the main methods applied to evaluate EpiScores and measured protein levels as potential biomarkers of CVD.

4. Cohort Description and Key Methods

This chapter introduces the large-scale cohort studies used in this thesis: GS, the LBC1936, and the Avon Longitudinal Study of Parents and Children (ALSPAC). It also presents a concise overview of the main methods applied in the empirical analyses.

4.1. Generation Scotland

GS is a large, family-based cohort designed to investigate the genetic, environmental, and lifestyle determinants of health and disease in the Scottish population. The full cohort profile [291] and study protocol [292] were published previously The study comprises 24,084 individuals from 5,501 families, with recruitment conducted in two phases taking place between 2006 and 2011 (baseline) [293]. Phase 1 focused on Glasgow and Tayside, primarily targeting individuals aged 35-65 who were randomly selected through a network of general medical practices across Scotland. Participants were encouraged to invite their relatives (at least one sibling and any other first-degree adult relatives). Phase 2 expanded recruitment across Ayrshire, Arran, and the North East of Scotland. Overall, 59% of participants were female, with an age range of 18 – 99 years (mean = 47.7, SD = 15.4). At baseline, participants completed detailed questionnaires covering medical and family history, lifestyle, socioeconomic factors, and smoking behaviour. Blood (or saliva) samples were collected for DNA extraction, and in-person assessments captured biochemical and clinical measurements, including standard anthropometric and physiological data (e.g., height, weight, blood pressure). Participants provided consent for re-contact for future research and gave permission for linkage to NHS health records via the Community Health Index number. Linked data include hospital admissions (SMR01), outpatient visits (SMR00), mortality records, prescription data, and, for a subset, primary care records.

In the following sections, I will provide an overview of multi-omic data from GS that I used in this thesis, including genetic (DNA), epigenetic (DNAm) and proteomic data. I will also describe available risk factor measures and CVD diagnosis information.

4.1.1. Ethics and Funding

All components of GS received ethical approval from the National Health Service Scotland Tayside Committee on Medical Research Ethics (05/S1401/89 and 14/SS/0039). Research Tissue Bank status was granted by the Tayside Committee on Medical Research Ethics (20-ES-0021). GS is supported by core funding from the Chief Scientist Office of the Scottish Government Health Directorates (CZD/16/6) and the Scottish Funding Council (HR03006). Genotyping was funded through the Medical Research Council and the Wellcome Strategic Award (104036/Z/14/Z). Funding from the Medical Research Council and Wellcome also supported DNA methylation typing (104036/Z/14/Z), in addition to support from the Brain and Behaviour Research Foundation (27404).

4.1.2. Genetic Data

Blood samples were collected, processed, and stored according to standard operating procedures at the Wellcome Clinical Research Facility, Western General Hospital, Edinburgh. DNA concentrations were quantified using the Invitrogen PicoGreen assay kit and diluted to 50 ng/μL. Aliquots of 4μL were used for genotyping [294]. Genotyping was performed using Illumina HumanOmniExpressExome-8 BeadChips (versions 1.0 and 1.2) [121].

Quality control (QC) was carried out in PLINK v1.9b2c [295] as described previously [294]. Briefly, SNPs with missingness > 2% across all individuals or Hardy–Weinberg equilibrium $P < 1\times10^{-6}$ were excluded. Samples showing sex discrepancies relative to the phenotypic database or with > 2% missing genotypes were also removed. To identify population outliers, GS genotypes were merged with data from 1,092 individuals in the 1000 Genomes Project. Individuals located more than six SDs from the mean of the first two principal components of the merged dataset were classified as outliers and excluded.

After QC, the dataset comprised 20,032 individuals (11,805 females and 8,227 males) and 604,858 autosomal SNPs [294]. Following imputation with the Haplotype Reference Consortium reference panel, 24,161,581 variants with INFO > 0.4 were available.

4.1.3. Epigenetic Data

DNAm was profiled using two approaches: array-based profiling and, for a subset of samples, NGS.

Array-based Profiling

Blood samples collected at baseline were used for DNA extraction with Nucleon BACC3 kits. Aliquots of 500 ng DNA were bisulphite-converted using the EZ-96 DNA Methylation Kit (Zymo Research, Irvine, California) according to the manufacturer's instructions. DNAm levels were measured using the Illumina Infinium MethylationEPIC BeadChip (Illumina Inc., CA), scanned on a HiScan system (Illumina Inc., San Diego, California), and quality control was performed using GenomeStudio (version 2011.1). DNAm data were generated in multiple sets, referred to as "waves" [296]: Wave 1 (5,087 baseline samples, 860,925 probes) prepared between 2016 and 2017; Wave 2 (459 baseline, 501 longitudinal samples, 859,730 probes) prepared in 2017; Wave 3 (4,450 baseline, 295 longitudinal samples, 856,671 probes; unrelated individuals) prepared between 2018 and 2019; Wave 4 (8,873 baseline samples, 854,862 probes) prepared in 2021. The 796 longitudinal samples from Waves 2 and 3 were collected as part of the sub-study Stratifying Resilience and Depression Longitudinally (STRADL) [297]. QC steps were performed separately for each wave [296]. Methylation profiling was performed in multiple batches within each wave.

Before QC of Wave 1, 10 saliva-derived samples mistakenly submitted for whole-blood methylation profiling were excluded, along with three individuals reporting "Yes" to all health conditions and one sample likely to be XXY (identified through genotype data). Wave 1 QC was performed using shinyMethyl v1.10.0 [298]. Technical outliers were identified by visual inspection of plots showing the log median methylated signal intensity against the unmethylated signal, as well as via control probe and multidimensional scaling plots. Samples with mismatched methylation-predicted and self-reported sex were removed. Using the pfilter function from wateRmelon v1.18.0 [299], the following were excluded: (i) samples with >1% of CpGs having detection P > 0.05; (ii) probes with beadcount < 3 in > 5% of samples; and (iii) probes with detection P > 0.05 in > 0.5% of samples. White blood cell (WBC) proportions (monocytes, granulocytes, CD4+, CD8+ T cells, B cells, NK cells) were estimated using minfi's [300] Houseman algorithm [301].

QC for Waves 2–4 used meffil v1.1.0–1.1.2 [302] and shinyMethyl v1.14.0–1.30.0 [298]. Dye-bias and background correction was performed using meffil. The same package was used to exclude samples showing dye bias, poor bisulfite conversion, low median methylated signal intensity (>3 SDs below expectation), or sex discordance. Outliers were identified via control probe inspection in shinyMethyl. Poorly performing samples (>0.5% of CpGs with detection P > 0.01) were removed, followed by exclusion of probes with beadcount < 3 in > 5% of samples or detection P > 0.01 in > 1% of samples. WBC proportions were estimated as in Wave 1.

After QC, X-chromosome and suboptimally binding probes (Zhou *et al*. [303]; McCartney *et al*.[304]) were removed, along with Y-chromosome probes. Each wave was then normalised separately using dasen in wateRmelon v2.2.0 [299]. In addition, a jointly normalised dataset was prepared, in which the four QC'd waves were combined and normalised together using dasen [296]. The combined baseline DNAm dataset included 18,869 post-QC samples with 851,610 probes, representing the largest single-cohort DNAm dataset to date [296]. Participants with available DNAm data had a mean age of 47.1 years (SD=14.9), and 58.8% were female.

Next generation sequencing

DNAm was profiled in 48 unrelated individuals from Wave 3 (24 current smokers and 24 never smokers) using two NGS platforms: Twist Bioscience targeted short-read sequencing (Human Methylome Panel; ~4 million CpG sites) and Oxford Nanopore Technologies (ONT) long-read sequencing (~28 million CpG sites) [305]. To maximise contrast between groups, 12 male and 12 female heavy smokers (pack years of smoking ranging from 53.9 to 87.7) were selected as cases. Age- and sex-matched controls were identified using the MatchIt package in R [306].

TWIST sequencing was performed by the Genetics Core at the Edinburgh Clinical Research Facility according to the manufacturer's targeted methylation sequencing protocol [307]. Reads were aligned to the GRCh38 human reference genome (29.4 million CpG sites) using bwa-meth and processed with the MethylSeq v2.2.0 pipeline[308,309], yielding methylation and coverage estimates for 18,248,472 CpG sites.

Before ONT sequencing, sample quality was checked using the Fragment Analyzer (Agilent), Qubit 2.0, and Nanodrop-8000 [305]. Each 1.5 µg sample was purified, repaired, end-prepped, and barcoded using the ONT Native Barcoding Kit 24 with NEBNext ligation and repair modules. Libraries (unsheared for the first 24 samples, 10 kb sheared for the rest) were bead-purified, pooled by sex and age, adapter-ligated, and enriched for fragments >3 kb. Sequencing was performed on the Oxford Nanopore PromethION 24 (R10.4.1 flow cells, 72 h runs). The first 24 libraries were sequenced at Edinburgh Genomics without basecalling; later runs were processed at the Genetics Core using Dorado for high-accuracy basecalling, alignment, and CpG (5mC/5hmC) modified base detection. Data were basecalled with Dorado and analysed using the epi2me-labs/wf-human-variation v23.10.1 pipeline (GRCh38 reference), producing methylation estimates for 28,989,402 CpG sites.

QC of TWIST and ONT sequencing outputs was conducted in R v4.3.1 using the Methrix package [310]. Sites with very low (depth of coverage <2) or extremely high coverage (>0.99 quantile), or those overlapping known cytosine to thymine polymorphisms, were excluded. To ensure reliable estimates, only CpGs covered in at least 40 samples by ≥10 reads (TWIST) or ≥5 reads (ONT) were retained, resulting in 3,391,718 and 21,167,712 CpG sites, respectively. CpG sites were annotated using the Annotatr package [311]. As part of QC, one pair of individuals was filtered out, due to the age difference exceeding 12 months. After QC, NGS data were available for 46 individuals.

4.1.4. Proteomic Data

Serum samples from 15,818 individuals, collected during their baseline assessment, were analysed by mass spectrometry. Procedures for sample preparation, LC-MS, and MS result annotation are described below.

Serum sample preparation followed previously published protocols [239,312]. In brief, 5µL aliquots of serum were added to a solution of 0.1M ammonium bicarbonate (pH 8.0) and 50µL of 8M urea for protein denaturation. Proteins were reduced with 5µL of 50mM dithiothreitol for 1h at 30°C, alkylated with 5µL of 100mM iodoacetamide for 30min in the dark, and diluted with 340µL of 0.1M ammonium bicarbonate to reach 1.5M urea. Digestion was performed overnight at 37°C using trypsin (12.5µL, 1:40 enzyme-to-protein ratio). Reactions were quenched with 25µL of 0.1% v/v formic acid. Peptides were purified using C18 plates, eluted with 50% acetonitrile and dried under vacuum. The samples were prepared for LC-MS by redissolving the peptides in 50µL of 10% v/v formic acid.

Peptides were analysed using an Agilent 1290 Infinity II liquid chromatography system coupled to a SCIEX TripleTOF 6600 mass spectrometer. Two micrograms (2 µg) of peptides were injected onto a Luna Omega C18 column (1.6 µm particle size, 30 mm

× 2.1 mm), which separates molecules based on how strongly they interact with the column surface. Peptides were eluted over a 3-minute linear gradient, starting from 1% to 40% of Buffer B (Buffer A: 0.1% formic acid in water; Buffer B: 0.1% formic acid in acetonitrile) at a flow rate of 800 µL/min. After separation, the column was washed by increasing Buffer B from 40% to 80% over 0.5 min, held at 80% for 0.2 min, and then re-equilibrated at 3% Buffer B for 1 min in preparation for the next sample.
Data acquisition used a Scanning SWATH method [313] in high-sensitivity mode with the following parameters: total cycle time = 0.69 s, transmission window = 10 Da, precursor range = 450–850 Da, fragment range = 100–1500 Da, and accumulation time = 16.9 ms. Source settings were: gas 1 = 15 psi, gas 2 = 20 psi, curtain gas = 25 psi, temperature = 0 °C, IonSpray voltage = 5500 V, declustering potential = 80 V. Rolling collision energies were calculated using the equation CE = 0.034 × (m/z) + 2, where m/z represents the centre of the scanning quadrupole bin.

Output data were processed using DIA-NN v1.8.12, which applies deep neural networks for analysis of data-independent acquisition (DIA) proteomics [314]. The Robust LC quantification mode with default parameters was used. Protein identification was based on a previously generated spectral library [315], refined as described by Messner *et al*. [239]. A 1% FDR was applied at both precursor and gene-group levels, and only samples with ≥ 2000 identified precursors were retained. Precursors were required to have ≥ 80% prevalence in QC samples. Within-batch drift was corrected by fitting linear models to repeat injections of pooled QC samples [316], and between-batch effects were adjusted using limma v3.54.2 [317].

Protein identifiers were annotated using the UniProt ID mapping tool [318]. In total, 439 proteins were identified and classified as either individual proteins (n = 133) or protein groups (n = 306) (**Table 8**). Annotation processing involved removing any text enclosed in parentheses or brackets. Individual proteins were named using these standardised annotations. For protein groups, nomenclature depended on group type, defined as either gene-derived (G) or mixed (PG). Gene-derived groups comprised proteins sharing at least one common gene name and were sequentially labelled (e.g., Albumin [G1], Albumin [G2], etc.). Mixed groups contained proteins annotated to multiple, distinct genes and were named according to their constituent gene annotations. For groups containing up to four genes, all gene names were listed; for larger groups, three gene names representing the group diversity were reported followed by an ellipsis.

Table 8. Protein annotation details. For each protein category, the final three columns provide exemplar UniProt identifiers, gene symbols and the final annotations.

Protein Category	*Count (n)*	*Example UniProtID(s)*	*Gene Symbol(s)*	*Final annotation*
Unique Proteins	133	P00450	CP	Ceruloplasmin
Gene-Derived Protein Groups (G)	199	P02647.F8W696	APOA1, APOA1	Apolipoprotein A-I (G)
		P00736.F5H2D0.B4DPQ0	C1R, C1R, C1R	Complement C1r subcomponent (G1)
		P00736.F5H1V0.F5H6Y3.B4DPQ0	C1R, C1R, C1R, C1R	Complement C1r subcomponent (G2)
Mixed Protein Groups (PG)	107	P80748.A0A075B6K5	IGLV3-21, IGLV3-9	PG 70 (IGLV3-21, IGLV3-9)
		P00751.C9JRT3.B4E1Z4	BF, BFD, CFB	PG 18 (BF, BFD, CFB)
		A0A0B4J1V1.P01780.P01767. A0A0B4J1X5.A0A0J9YVY3.P0DP02. P01762.A0A0C4DH42.P01772. P01763	IGHV3-21, IGHV3-7, IGHV3-53, IGHV3-74, IGHV7-4-1, IGHV3-30-3, IGHV3-11, IGHV3-66, IGHV3-33, IGHV3-48	PG 7 (IGHV3-11, IGHV3-74, IGHV7-4-1, ...)

4.1.5. Risk Factor Measures

Age and sex were obtained from questionnaire data with the latter verified using genetic data (X/Y chromosome presence). Systolic blood pressure was recorded as the mean of two measurements. Total and HDL cholesterol were measured in blood samples, and non-HDL cholesterol (total cholesterol – HDL cholesterol) was used as a proxy for LDL cholesterol due to the absence of triglyceride data. Disease status and educational attainment were self-reported via a pre-clinic questionnaire. Educational attainment corresponds to years of full-time study and is an ordinal variable defined as follows: 0 (0 years), 1 (1–4 years), 2 (5–9 years), 3 (10–11 years), 4 (12–13 years), 5 (14–15 years), 6 (16–17 years), 7 (18–19 years), 8 (20–21 years), 9 (22–23 years), and 10 (24 or more years). Residential postcodes were used to obtain a SIMD deprivation rank. Smoking status (current, former and never smokers) and pack-years were derived from self-reported smoking history. Cotinine measurements were not available. Pack years were calculated using the following formula (**Eq. 2**):

$$pack\ years = \frac{(\text{age of cessation} - \text{age of initiation})\ \text{x cigarettes per day}}{20}$$

Eq. 2

Table 9 presents an overview of the risk factors described above.

Table 9. Selected CVD risk factors in Generation Scotland. Continuous variables with a normal distribution are reported as mean (± SD), while those with a skewed distribution are reported as median [Q1, Q3]. Categorical variables are presented as n (%). Educational attainment is a categorical variable but I reported it here as the median [quartile 1, quartile 3] for readability. HDL indicates high-density lipoprotein; SBP, systolic blood pressure; and SIMD, Scottish Index of Multiple Deprivation.

Risk factor	*n*	*Mean / Median / n*
Age, years	24 079	47.7 (15.4)
Male sex, n (%)	24 079	9,924 (41.2%)
SIMD, rank	21 136	4 331 [2 373, 5 451]
Diabetes, n (%)	23 590	804 (3.4%)
Rheumatoid arthritis, n (%)	23 585	431 (1.8%)
SBP, mm Hg	21 489	131.2 (17.8)
Total cholesterol, mmol/L	20 390	5.1 (1.1)
HDL cholesterol, mmol/L	20 350	1.4 [1.2, 1.7]
Educational attainment *	22 791	4 [3, 6]
Smoking, pack years	23 163	0 [0, 9]
Smoking status, n (%)	23 163	3 992 (17.2%) current smokers, 591 (2.6%) quit ≤12 months ago, 6 274 (27.1%) quit >12 months ago, 12 306 (53.1%) never smokers

4.1.6. CVD Diagnosis

CVD event dates were identified through linkage to routinely collected NHS health records. Primary care (general practice, GP) data were available only for a subset of participants, reflecting the limited number of GP practices that consented to publishing research conducted using their participants' details. In total, primary care data were available for 7,580 participants, while secondary care (hospitalisation) records were available for 21,725 individuals. For this reason, all analyses presented in this thesis are based solely on secondary care data. The composite CVD outcome was defined as per Welsh *et al*.[28] and included CHD, ischaemic stroke, MI, and CVD death (exact ICD-10 codes are given in **Chapters 6** and **7**). Censoring dates were at the end of the follow up period (**Chapter 6**: September 2021; **Chapter 7**: August 2023) or at time of death. Individuals with CVD events occurring prior to the baseline assessment (prevalent cases) or after the censoring date were classified as controls. As CVD risk varies with age (more details in **Chapter 6**), participants in **Chapters 6** and **7** were restricted to those within the SCORE2-recommended age range (40-69 years) [113]. Case and control sample sizes are reported in the relevant chapters, and time-to-event was calculated as age at event (CVD diagnosis or censoring) minus age at baseline.

4.2. The Lothian Birth Cohort 1936

LBC1936 is a longitudinal, population-based study established to investigate the determinants and trajectories of cognitive and brain ageing from childhood into later life. The cohort has been described previously [319,320]. It is unique in its ability to relate cognitive changes across the lifespan to a wide range of genetic, medical, psychosocial, and lifestyle factors, due to its linkage to childhood cognitive test scores.

LBC1936 comprises surviving participants of the Scottish Mental Survey 1947, in which nearly all children born in 1936 and attending school in Scotland were tested on June 4, 1947, using the Moray House Test No. 12, a validated measure of general intelligence. Decades later, residents of the Lothian area, born in 1936 who were likely to have taken part in the survey, were identified via Community Health Index and invited to participate. The recruitment took place between 2004 and 2007. Out of 70,805 children tested across Scotland, 1,091 individuals were recruited. At recruitment (wave 1 baseline), participants had a mean age of 70 years, and 50% were female. Follow-up assessments were conducted at approximate ages 73, 76, 79, 82, and 86, with a seventh wave currently underway.

Each wave collects extensive cognitive, medical, physiological, and psychosocial data, encompassing cognitive function (with childhood IQ as baseline), and detailed demographic, lifestyle, and mental health measures. Neuroimaging and omics data provide further insight into brain ageing. Structural MRI supports analyses of brain morphology and white matter integrity. Post-mortem brain tissue (n = 14) enables investigation of neuropathological processes. Genetic and epigenetic data are also available, including genome-wide DNAm profiles from whole blood (Illumina Infinium HumanMethylation450) and five post-mortem brain regions (Illumina EPIC850k).

In the next sections, I will provide an overview of the LBC1936 data used in this thesis, including blood and brain DNAm and smoking phenotypes.

4.2.1. Ethics and Funding

Ethical approval for the LBC1936 study was obtained from the Multi-Centre Research Ethics Committee for Scotland (MREC/01/0/56) and the Lothian Research Ethics committee (LREC/1998/4/183; LREC/2003/2/29). Use of human tissue for post-mortem studies has been reviewed and approved by the Edinburgh Brain Bank ethics committee and the ACCORD medical research ethics committee, AMREC (ACCORD is the Academic and Clinical Central Office for Research and Development, a joint office of the University of Edinburgh and NHS Lothian). All participants provided written informed consent. These studies were performed in accordance with the Helsinki declaration. LBC1936 jointly-funded by the Biotechnology and Biological Sciences Research Council and the Economic and Social Research Council (BB/W008793/1), and has also been supported by Age UK (Disconnected Mind project), and the Medical Research Council (G0701120, G1001245, MR/M013111/1, MR/R024065/1), the Milton Damerel Trust, and the University of Edinburgh. Methylation typing in LBC1921 and LBC1936 populations was supported by the Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), The University of Edinburgh, The University of Queensland, Age UK and The Wellcome Trust Institutional Strategic Support Fund.

4.2.2. Epigenetic Data

DNAm data were generated from whole-blood and *post-mortem* brain tissue samples. Blood samples were collected at baseline by trained research nurses at the Wellcome Trust Clinical Research Facility Genetics Core, Western General Hospital, Edinburgh [321]. Of the 1,901 baseline blood samples, 1,005 passed GWAS QC and were selected for DNAm profiling. Of these, 920 samples passed DNAm QC. Twenty-five samples were excluded due to missing phenotype data (n = 11) or unavailable white blood cell measurements (n = 14), resulting in a final dataset of 895 samples. DNAm was quantified at 485,512 CpG sites using the Illumina Human Methylation 450k BeadChips (Illumina Inc., San Diego, CA). Full details of DNAm QC were published previously [151,322]. Briefly, raw intensity data were background-corrected and normalised using internal control probes. Probes were excluded if they showed: (i) a low detection rate (<95% with detection P<0.01), or (ii) inadequate hybridisation, bisulfite conversion, nucleotide extension, or staining signal (based on manual inspection). Samples were removed if they showed (i) a low call rate (<450,000 CpGs at $p < 0.01$), (ii) sex mismatch, or (iii) genotype discordance based on SNP control probes. After quality control and normalisation using the danet method, 450,276 autosomal CpG sites remained for 895 individuals at Wave 1.

A brain tissue bank was established at wave 3 (from age ~76 years) [323]. At *post-mortem*, brains were divided into coronal slices. Samples from the cortical areas BA17, BA20–21, BA24, BA46, and the hippocampus were collected and immediately frozen. Approximately 25mg of tissue was obtained from these sections for DNA extraction (Dneasy kit (Qiagen)). DNAm levels were measured using Illumina MethylationEPIC BeadChips at the Edinburgh Clinical Research Facility. Background signal from type I and type II probes was equalised as part of danet normalisation. Probes were excluded using wateRmelon pfilter() function if (i) more than 1% of probes had a detection p-value > 0.05, (ii) beadcount was <3 in >5% of samples, or (iii) probes mapped to polymorphic or cross-hybridising targets or to sex chromosomes. Samples with >1% of probes failing detection (p > 0.05) were also removed. Following QC and normalisation, 807,163 CpG sites were retained (n = 14 individuals, $n_{hippocampus}$ = 13 individuals).

4.2.3. Phenotypic Data

Smoking information was self-reported at the baseline assessment (age 70) and included smoking status (never, former, or current smoker) and smoking behaviour (age at initiation, age at cessation, and an average number of cigarettes smoked per day). I calculated pack-years from these data using the formula in **Eq. 2**. Cotinine measurements were not available. **Table 10** shows descriptive statistics of the main LBC1936 variables used in this thesis.

Table 10. Overview of the main LBC1936 phenotypes used in this thesis. Continuous variables with a normal distribution are reported as mean (± SD), while those with a skewed distribution are reported as median [Q1, Q3]. Categorical variables are presented as n (%).

Phenotype	*n*	*Mean / Median / n*
Age, years	1 091	69.6 (0.8)
Sex, n (%)	1 091	548 (50.2%) males
Smoking, pack years	1 072	1.9 [0, 28]
Smoking status, n (%)	1 091	125 (11.5%) current smokers, 465 (42.6%) former smokers, 501 (45.9%) never smokers

4.3. The Avon Longitudinal Study of Parents and Children

ALSPAC is a population-based, multi-generational birth cohort study established to explore how genetic, environmental, and social factors shape health and development across the life course. The study recruited pregnant women living in the former Avon Health Authority area of South-West England, with expected delivery dates between 1st April 1991 and 31st December 1992. For this reason, ALSPAC is also known as the "Children of the 90's" study. Out of 20,248 eligible pregnancies, 14,541 women were enrolled, resulting in 14,062 live births, with 13,988 children surviving beyond their first year. During the antenatal period, mothers were invited to involve biological fathers; 12,113 fathers completed at least one questionnaire, and 3,807 are formally enrolled as participants.

This thesis uses the following four sub-cohorts of ALSPAC:

- Antenatal collection – mothers during pregnancy,
- Focus on Mothers (FOM) and Focus on Fathers (FOF) – mothers and fathers at midlife (~50 years),
- F17 – offspring assessed at ages 15–17 years,
- F24 – offspring assessed at age 24 years.

Data were gathered and administered utilising REDCap electronic data capture tools, which are hosted at the University of Bristol. REDCap (Research Electronic Data Capture) is a secure web application specifically designed to facilitate data capture for research [324]. The study website provides comprehensive details on all available data, accessible through a fully searchable data dictionary located at http://www.bristol.ac.uk/alspac/researchers/our-data/.

In the next sections, I will discuss ALSPAC data that I used in this thesis, including blood DNAm and smoking phenotypes.

4.3.1. Ethics and Funding

Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time. Children were invited to give assent where appropriate. Study participants have the right to withdraw their consent for elements of the study or from the study entirely at any time. Full details of the ALSPAC consent procedures are available on the study website (http://www.bristol.ac.uk/alspac/researchers/research-ethics/).

A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf).
The UK Medical Research Council and Wellcome (Grant ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. Funding for ALSPAC DNAm measurements were supported by the Wellcome (102215/2/13/2); the University of Bristol; the UK Economic and Social Research Council (ES/N000498/1); the UK Medical Research Council (MC_UU_12013/1, MC_UU_12013/2); the Biotechnology and Biological Sciences Research Council (BBI025751/1 and BB/I025263/1); and the John Templeton Foundation (60828).

4.3.2. Epigenetic Data

For a subset of ALSPAC participants DNAm was assayed as part of the Accessible Resource for Integrated Epigenomic Studies (ARIES) initiative, which has been described previously (http://www.ariesepigenomics.org.uk) [325]. Briefly, mothers, fathers and children had peripheral blood DNAm profiled using either the Illumina 450k array or the Illumina EPIC array at multiple time points. Blood samples from children were obtained at 7, 9, 15–17, and 24 years of age, and from mothers during pregnancy ("antenatal") and again approximately 18 years postpartum ("midlife").

In this thesis, I used the following DNAm data:

- F17 – a mixed 450k and EPIC array dataset (n_{450k} = 374, n_{EPIC} = 840)
- F24 – EPIC array dataset (n=496)
- FOM / FOF – 450k array dataset (n=1207)
- Antenatal – 450k array dataset (n=968)

DNAm wet-lab assays and data preprocessing were performed at the University of Bristol. QC and normalisation of the combined EPIC and 450K datasets followed the procedures implemented in the Meffil pipeline [302]. After QC, 450,838 CpG sites remained, representing those common to both array platforms.

4.3.3. Phenotypic Data

Smoking status was defined separately for offspring and parents, based on repeated questionnaire data collected across follow-up. **Table 11** summarises the criteria used to classify smoking status for each group and time point in ALSPAC.

Table 11. Criteria used to define smoking status for each studied sub-cohort of the Avon Longitudinal Study of Parents and Children. The antenatal collection includes data from mothers during pregnancy; the FOM/FOF collection represents mothers and fathers at midlife (~50 years); and the F17 and F24 collections correspond to offspring at ages 15–17 and 24, respectively.

Group	*Assessment period & questionnaires*	*Current smokers*	*Former smokers*	*Never smokers*
F17 (offspring age 14–17)	3 questionnaires (ages 14–17)	Smoked weekly in one or more questionnaires and had not quit	Reported having quit smoking at age 17 questionnaire	Reported never smoking at least once and never reported smoking
F24 (offspring age 14–24)	6 questionnaires (ages 14–24)	Smoked in last 30 days at age 24 and smoked ≥50 cigarettes lifetime	Smoked regularly prior to age 24 but not in previous 30 days at age 24	Reported never smoking at age 24
Antenatal (mothers)	2 questionnaires (18- and 32-weeks gestation)	Smoked regularly in first trimester	Smoked previously but stopped for pregnancy	Never smoked before or during pregnancy
FOM (mothers)	Multiple questionnaires (child ages ≤12 and final at age 18)	Current regular smoker on the 18y questionnaire	Smoked on ≥1 earlier questionnaire but reported not smoking on the 18y questionnaire	Consistently reported never smoking
FOF (fathers)	11 questionnaires (child ages ≤12 and final at age 20)	Current regular smoker on the 20y questionnaire.	Smoked on ≥1 earlier questionnaire but not smoking on the 20y questionnaire.	Consistently reported never smoking

Table *12* shows summary characteristics of ALSPAC sub-cohorts. Pack years and cotinine data were not available.

Table 12. Overview of the main phenotypes from the Avon Longitudinal Study of Parents and Children cohort used in this thesis. Continuous variables with a normal distribution are reported as mean (± SD), while those with a skewed distribution are reported as median [Q1, Q3]. Categorical variables are presented as n (%). The antenatal collection includes data from mothers during pregnancy; the FOM/FOF collection represents mothers and fathers at midlife (~50 years); and the F17 and F24 collections correspond to offspring at ages 15–17 and 24, respectively.

Variable	*F17*	*F24*	*FOM / FOF*	*Antenatal*
Sample size	1,214	496	1,207	968
Age, years	17.7 (0.4)	24.4 (0.8)	50.2 (5.4)	28.8 (4.4)
Male sex, n (%)	552 (45.5%)	232 (46.8%)	539 (44.7%)	0 (0%)
Current smokers, n (%)	110 (9.1%)	155 (31.3%)	90 (7.5%)	98 (10.1%)
Former smokers, n (%)	132 (10.9%)	96 (19.4%)	474 (39.3%)	268 (27.7%)
Never smokers, n (%)	972 (80.1%)	245 (49.4%)	643 (53.3%)	602 (62.2%)

4.4. Key Methods

This section provides an overview of the key analytical methods employed in this thesis. Full details on their implementation and application can be found in the methods sections of **Chapters 5** through **7**.

4.4.1. Statistical Analysis

In **Chapter 5**, I conducted EWASs of smoking using two statistical approaches. A large-scale EWAS of pack-years was conducted using Bayesian penalised regression in BayesR+. Other EWASs (high-dimensional, NGS-based EWASs and analyses conducted using DNAm measured in brain tissue) were conducted using frequentist marginal linear regression models. Both methods were described in detail and compared in **Section 2.3.2**.

Next, I used the results of these analyses to construct a smoking EpiScore (EpiScores were introduced in **Section 2.3.4**), called mCigarette. It was developed using penalised linear regression (elastic net) [326]. Elastic net regression is a regularised linear modelling approach that is particularly effective for high-dimensional datasets, where the number of predictors may exceed the number of observations or where predictors are highly correlated. The method performs variable selection and coefficient shrinkage simultaneously, identifying the subset of features most relevant for prediction while reducing overfitting. As in ordinary linear regression, the model aims to find the best-fitting line (or hyperplane) that relates predictors to the outcome. However, elastic net introduces a penalty term that constrains the size of the regression coefficients, preventing the model from fitting the data too closely and thus improving generalisability. Two common types of penalties are used in regularised regression. The LASSO penalty (L1) shrinks some coefficients exactly to zero, effectively removing uninformative predictors. In contrast, the Ridge penalty (L2) shrinks all coefficients towards zero but retains them in the model, which helps stabilise estimates when predictors are highly correlated. Elastic net combines these two penalties, balancing the advantages of each. In the context of mCigarette construction, elastic net regression was applied to identify and weight CpG sites that best predicted cumulative smoking exposure, measured as pack-years. The model was trained in the GS cohort, where both DNAm and pack-years data were available. The resulting regression coefficients were then applied to the LBC1936 to generate individual mCigarette EpiScores, representing predicted pack-years based solely on DNAm profiles.

The predictive performance of the mCigarette score was evaluated using AUC and through incremental R^2. AUC is a measure of a model's ability to discriminate between classes (binary classification), ranging from 0.5 to 1.0, with the former representing random guessing and higher values indicating better distinction between individuals with and without the outcome. Incremental R^2 was calculated to assess the additional variance in pack-years explained by the inclusion of the mCigarette score. Specifically, it was derived by comparing the R^2 values from models adjusted for age and sex with those that additionally included the mCigarette score.

Finally, I conducted an GWAS of epigenetic and phenotypic smoking. GWAS was described in detail in **Section 2.2.2**.

In **Chapter 6**, I used Cox PH regression [327] to investigate the associations between time to incident CVD and protein EpiScores trained by Gadd *et al*. [1]. The Cox model estimates the hazard function h(t) – the instantaneous risk of experiencing the event at time t, given survival up to that point – as a function of covariates X (**Eq. 3**):

$$h(t|X) = h_0(t)\exp(\beta X)$$

Eq. 3

where $h_0(t)$ is the baseline hazard and β represents log hazard ratios associated with the covariates. Cox regression assumes proportional hazards, meaning that the ratio of hazard functions between any two groups remains constant over time. This assumption was evaluated using Schoenfeld residuals (cox.zph function from the survival package [328] in R) and visual diagnostics.

Because the GS cohort includes related individuals, mixed-effects Cox PH regression (implemented as part of the coxme R package [329]) was used to account for familial relatedness. The kinship matrix was fitted in to represent genetic relationships among participants. Model coefficients were exponentiated to obtain HRs with corresponding 95% confidence intervals.

To construct a composite protein EpiScore, a penalised Cox PH regression model was trained using elastic net regression (via the glmnet package [330]), incorporating all 109 protein EpiScores as candidate predictors. The predictive performance of this model was compared with that of a Random Forest model, a non-parametric ensemble learning approach that aggregates predictions from multiple decision trees [331,332]. Predictive performance was assessed by calculating AUC and the concordance index (C-index). C-index represents the proportion of all pairs of individuals in which the person with the higher predicted risk experiences the event earlier. The C-index ranges from 0.5 (no better than chance) to 1.0 (perfect discrimination).

In **Chapter 7**, I used Cox PH regression to investigate associations between MS-derived protein concentrations and several outcomes: incident CVD, its subtypes, and mortality. Protein abundances were rank-based inverse normal transformed and standardised (mean=0, SD=1) prior to analysis to approximate a normal distribution and ensure comparability of effect sizes across proteins. As 439 individual models were fitted (one per protein), P-values were corrected for multiple testing. The primary correction method used was the Bonferroni adjustment [333]. However, given its conservative nature, results were also compared with those obtained using the Benjamini-Hochberg FDR method [334], which provides a balance between identifying true associations and limiting false positives. In a separate set of models, an interaction term (sex × protein) was used, to study the sex-specific effects of proteins on CVD risk.

Finally, a proteomic risk score was constructed using linear elastic net regression. In the training dataset, deviance residuals from an age- and sex-adjusted Cox PH model of composite CVD were used as the outcome, capturing variation in CVD risk not explained by these covariates. The model was fitted using glmnet [330] with 10-fold cross-validation. Cross-validation is a model validation technique that partitions the data into multiple subsets (folds) to iteratively train and test the model, thereby preventing overfitting and ensuring that the estimated model generalises well to unseen data. The cross-validation procedure was stratified by family ID to ensure that related individuals were not split across folds. Proteins with non-zero coefficients in the final model were retained, and these coefficients were used to compute the proteomic risk score, following the same approach described in **Chapter 6**. Model performance was assessed using AUC.

4.4.2. Data Visualisation

As part of **Chapter 6**, I developed an R Shiny web application designed to visualise associations between individual protein EpiScores incident CVD risk. The application is hosted by the University of Edinburgh and can be accessed at: https://shiny.igc.ed.ac.uk/3d2c8245001b4e67875ddf2ee3fcbad2/.

The interface comprises three tabs: (a) forest plot, (b) risk over time, and (c) survival probability. The forest plot tab (**Figure 12**) displays HRs with 95%CI for the 67 significant protein EpiScore – CVD associations (P < 0.05) as estimated from two models: a basic model adjusted for the ASSIGN score (red), and a full model additionally adjusted for cardiac troponin I (cTnI; blue). Users can interactively filter and select specific protein EpiScores for display.

Figure 12. Shiny application developed as part of Chapter 6 (Forest plot). The plot visualises the relationships between Protein EpiScores and risk of incident cardiovascular disease.

The second tab illustrates how associations between protein EpiScores and incident CVD change over time. HRs are estimated per SD increase in the EpiScore. For example, as shown in **Figure 13**, the risk of CVD associated with the CRP EpiScore peaks approximately three years before the event. The risk estimate gradually declines to around 1.2 per SD ten years before the event and remains stable up to 16 years prior. At all examined time points, the association between CRP EpiScore and CVD risk remains statistically significant (P < 0.05).

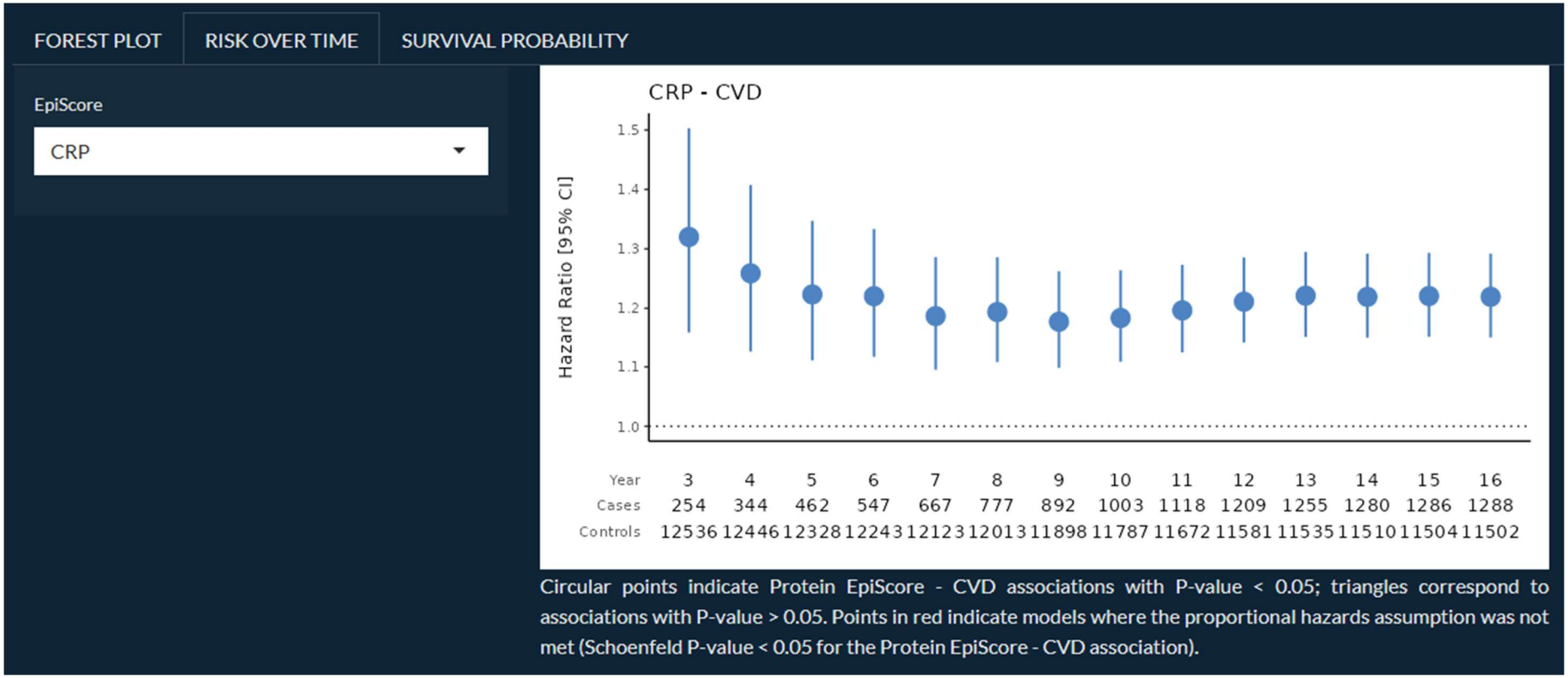


Figure 13. The relationship between the EpiScore for C-reactive protein (CRP) and incident cardiovascular disease plotted over time.

The third tab presents a Kaplan-Meier survival curve [335], which illustrates differences in CVD-free survival over time according to levels of the selected protein EpiScore. This non-parametric method estimates the probability of remaining free from CVD as a function of time since baseline. For example, individuals with higher metalloproteinase 12 (MMP12) EpiScores (above the 75th percentile) exhibited shorter CVD-free survival compared with those with lower EpiScores (below the 25th percentile), indicating an elevated risk of earlier disease onset (**Figure 14**).

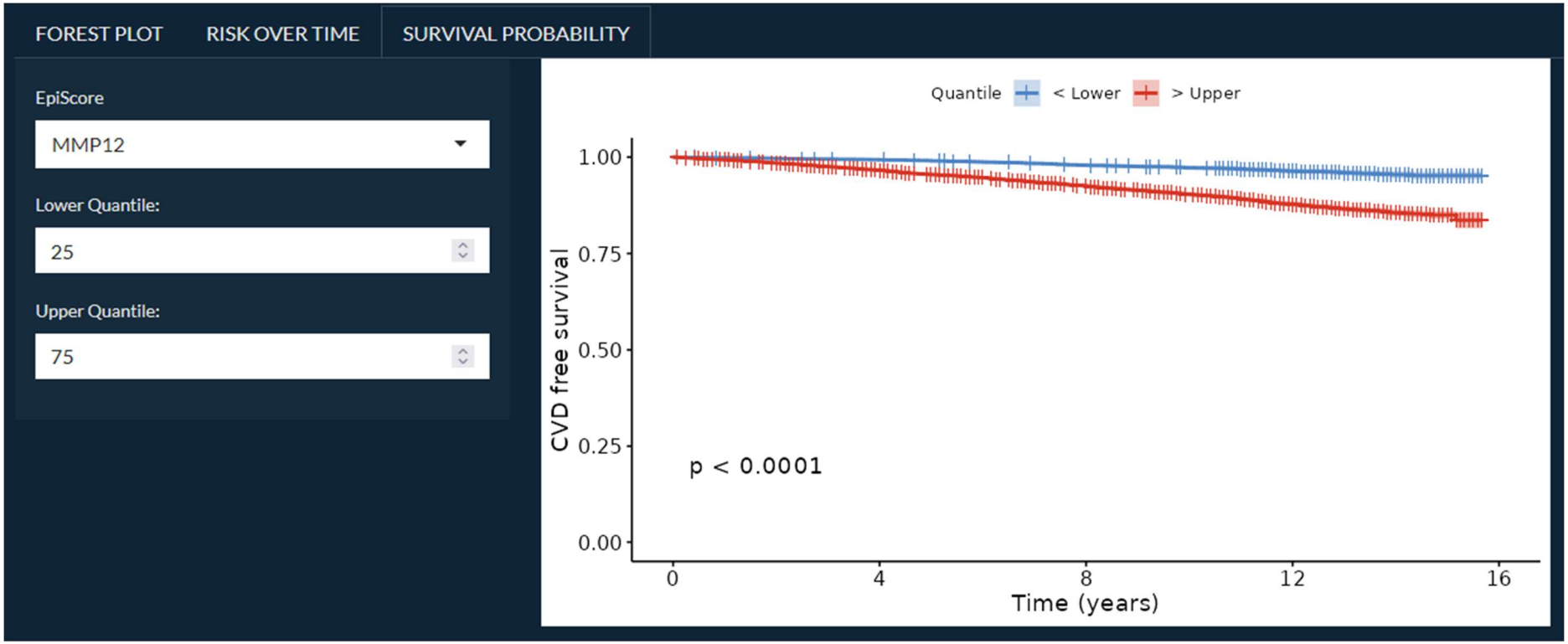


Figure 14. Differences in cardiovascular disease-free survival over time according to levels of the matrix metalloproteinase 12 (MMP12) EpiScore.

5. EWAS of Smoking

5.1. Introduction

Information on tobacco use is commonly incorporated into CVD risk prediction tools. However, self-reported smoking data are limited by recall bias and often fail to capture passive or occasional exposure. DNAm measured in blood can provide an objective molecular indicator of smoking exposure. Although previous research has identified robust smoking-related DNAm signatures, they were trained in DNAm from fewer than 10 000 individuals. Moreover, most studies have been restricted to array-based platforms and have not systematically investigated genome-wide methylation patterns or tissue specificity. In this chapter, I build upon existing evidence by performing a comprehensive analysis of the associations between DNA, DNAm and self-reported smoking behaviour.

To characterise the DNAm landscape associated with self-reported smoking, I conducted a large-scale Bayesian EWAS of pack-years in over 17,000 GS individuals, whose DNAm was measured using the Illumina EPIC array (~850,000 CpG sites). I then extended this analysis with a high-resolution EWAS in a subset of 46 GS individuals (23 age- and sex-matched smoker–non-smoker pairs), whose DNAm was sequenced using two complementary platforms: the TWIST Biosciences Human Methylome Panel (~4 million sites) and ONT sequencing (~21 million sites). Building on these results, I developed an EpiScore of smoking pack-years (mCigarette), which was trained in GS (n > 17,000), tested in LBC1936, and replicated across multiple age groups in ALSPAC. To assess tissue specificity, I compared smoking-associated DNAm patterns in blood and brain tissue (LBC1936). Finally, to explore the shared genetic architecture of smoking behaviour and its epigenetic surrogate, I conducted GWAS of both self-reported pack-years and an established DNAm-based smoking score (GrimAge DNAm pack-years). Collectively, these analyses provide a comprehensive, multi-layered view of smoking-associated methylation patterns, spanning array- and sequencing-based platforms, multiple tissues, and the intersection of genetic and epigenetic influences.

This study, published in *Nature Communications* in April 2025, is accompanied by supplementary material that can be accessed in the following repository https://github.com/aleksandra-chybowska/thesis/tree/main/A_blood-_and_brain-based_EWAS_of_smoking and through the electronic links provided by the publisher. Summary statistics from the EWAS and GWAS analyses are available on Zenodo (https://doi.org/10.5281/zenodo.14878399), and the code used to perform these analyses can be found in https://zenodo.org/records/14882849.

5.2. A Blood- and Brain-Based EWAS of Smoking

nature communications

Article

https://doi.org/10.1038/s41467-025-58357-6

A blood- and brain-based EWAS of smoking

Received: 12 June 2024

Accepted: 18 March 2025

Published online: 04 April 2025

Check for updates

Aleksandra D. Chybowska[1], Elena Bernabeu[1], Paul Yousefi[2,3,4], Matthew Suderman[2,3,4], Robert F. Hillary[1], Richard Clark[5], Louise MacGillivray[5], Lee Murphy[5], Sarah E. Harris[6], Janie Corley[6], Archie Campbell[1,7], Tara L. Spires-Jones[8,9], Daniel L. McCartney[1], Simon R. Cox[6,10], Jackie F. Price[7], Kathryn L. Evans[1] & Riccardo E. Marioni[1]

DNA methylation offers an objective method to assess the impact of smoking. In this work, we conduct a Bayesian EWAS of smoking pack years (n = 17,865, ~850k sites, Illumina EPIC array) and extend it by analysing whole genome data of smokers and non-smokers from Generation Scotland (n = 46, ~4–21 million sites via TWIST and Oxford Nanopore sequencing). We develop mCigarette, an epigenetic biomarker of smoking, and test it in two British cohorts. Results of brain- and blood-based EWAS (n_{brain}=14, n_{blood} = 882, >450k sites, Illumina arrays) reveal several loci with near-perfect discrimination of smoking status, but which do not overlap across tissues. Furthermore, we perform a GWAS of epigenetic smoking, identifying several smoking-related loci. Overall, we improve smoking-related biomarker accuracy and enhance the understanding of the effects of smoking by integrating DNA methylation data from multiple tissues and cohorts.

Cigarette smoking remains a leading cause of preventable death and disease, accounting for approximately 8 million global deaths annually[1]. It is a major risk factor of more than 50 diseases including cardiovascular disease, lung cancer and dementia[2]. As smoking history is often used in clinical risk stratification assessments, enhancing the accuracy of cumulative tobacco consumption measurements has the potential to improve prevention and treatment of smoking-related diseases.

Traditionally, tobacco use has often been quantified using self-report questionnaire-response data, such as pack years or indication of current smoking status (i.e. current, former, never smoker), which are prone to recall bias and do not account for passive smoke exposure[3]. A more objective approach to assess smoking is by measuring the concentration of tobacco-related chemicals. However, the commonly used nicotine biomarker, cotinine, has an average half-life of 15–20 hours[4]. Consequently, the concentration of serum cotinine does not inform about time since cessation in recent quitters and it cannot help to distinguish former smokers from never smokers. This limitation is especially pertinent when estimating the risk of diseases that take many years to develop, such as cardiovascular disease.

Blood-based DNA methylation patterns show great promise as a long-term biomarker of smoking[5]. DNA methylation (DNAm) is a cell- and tissue-specific epigenetic modification of DNA molecules that does not change the DNA sequence itself. It involves the addition of a methyl group to cytosine residues and occurs predominantly at cytosine-phosphate-guanine dinucleotides, also known as CpG sites. CpG methylation levels reflect not only smoking status but also cumulative tobacco exposure and can be informative of time since quitting in former smokers[6–8]. In the majority of CpG sites, smoking-related DNAm changes are dose-dependent and reversible after cessation[9].

[1]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh EH4 2XU, UK. [2]Medical Research Council Integrative Epidemiology Unit at the University of Bristol, University of Bristol, Bristol, UK. [3]NIHR Bristol Biomedical Research Centre, University Hospitals Bristol and Weston NHS Foundation Trust and University of Bristol, Bristol BS8 2BN, UK. [4]Population Health Science, Bristol Medical School, University of Bristol, Bristol, UK. [5]Edinburgh Clinical Research Facility, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK. [6]Lothian Birth Cohorts, Department of Psychology, The University of Edinburgh, Edinburgh, UK. [7]Usher Institute, University of Edinburgh, 5-7 Little France Road, Edinburgh EH16 4UX, UK. [8]Centre for Discovery Brain Sciences, University of Edinburgh, Edinburgh, UK. [9]UK Dementia Research Institute, University of Edinburgh, Edinburgh, UK. [10]Scottish Imaging Network, A Platform for Scientific Excellence (SINAPSE) Collaboration, Edinburgh, UK.
e-mail: riccardo.marioni@ed.ac.uk

In previous studies, blood-based DNAm biomarkers of smoking almost perfectly discriminated smokers from never smokers[7,10]. However, their ability to differentiate former smokers from never smokers was relatively modest. Additionally, most of these studies relied on whole-blood DNAm assessments using arrays, which only measure a preselected subset of CpG sites present in the epigenome. For example, the current largest epigenome wide association study (EWAS) of smoking in adults (n = 15,907) was a meta-analysis conducted using the Illumina 450k BeadChip array (n ~ 450,000 CpG sites)[6]. The largest EWAS of smoking, in which DNAm was measured with Illumina EPIC array (n ~ 850,000 CpG sites, approximately 5% of all sites on the epigenome), considered blood samples of 15,014 individuals[11].

In this work, we update the existing biomarkers of smoking by analysing whole-blood DNAm levels measured with Illumina EPIC array in 17,865 individuals. For a subset of 46 individuals, we implement a high-resolution approach (~4 million CpG sites, TWIST human methylome panel and ~21 million sites, Oxford Nanopore Sequencing) aimed at measuring methylation levels at CpG sites which are currently absent from arrays. Furthermore, we develop a smoking biomarker (mCigarette) using the EPIC dataset and investigate its associations with self-reported smoking in two external cohorts (Lothian Birth Cohort – LBC1936 and The Avon Longitudinal Study of Parents and Children - ALSPAC). We also investigate variations in methylation patterns in both blood and brain, with DNA methylation measured across five *post-mortem* brain regions in 14 individuals. Finally, we compare the epigenetic proxy of smoking to self-reported smoking by considering the genetic loci associated with these phenotypes in genome-wide association studies (GWASs).

Results

EWAS of smoking

First, we ran a Bayesian EWAS of smoking (Fig. 1). There were 17,865 Generation Scotland (GS) volunteers with a measure of pack years of smoking and Illumina EPIC DNAm data. The average age of the sample was 47.6 years (standard deviation [SD] = 14.9), and 59.1% of the participants were female (Supplementary Data 1). In the BayesR EWAS, DNAm explained 50.0% (95% Credible Interval 46.0 – 53.9) of the variance in the pack years phenotype.

Forty-two independent CpGs were associated with smoking at posterior inclusion probability (PIP) > 80%, with 26 of these associations reaching a PIP > 95% (Supplementary Data 2). Among the associations with PIP > 80%, 33 had previously been reported in the EWAS catalogue[12] at $P < 1\times10^{-4}$, with 30 of these reaching $P < 1\times10^{-7}$. This catalogue is a resource that curates findings from published EWAS studies.

Associations unreported in the EWAS catalogue as being associated with smoking included nine sites at PIP > 80% and two CpGs with PIP > 95%. The former group included intergenic CpGs linked to neurodevelopment and addiction, such as cg22454588 (annotated to *SCAMP5*), cg27110277 (*FGF20*), and cg19404444 (*SKI*). The latter, high confidence associations, included cg02517189 (*GRIK5*) and cg00562553 (*HOXA4*).

High resolution EWASs of smoking

Next, we extended this analysis by running high resolution EWASs of smoking on a subset of 23 pairs (n = 46) of current vs never smokers with Illumina EPIC array (~850k CpG sites), TWIST human methylation panel (~4 million CpG sites), and Oxford Nanopore Technologies (ONT) sequencing data (~21 million CpG sites). At $P < 3.6\times10^{-8}$ (significance threshold set as per Saffari et al.[13]), the EPIC-based analysis revealed 15 CpG sites associated with smoking status (EWAS inflation factor, $\lambda = 0.94$), while the TWIST-based and ONT-based analyses identified 33 ($\lambda = 1.60$) and 9 ($\lambda = 1.11$) associations, respectively. At a less stringent threshold ($P < 1\times10^{-5}$), these counts increased to 42, 102, and 63 for the EPIC-, TWIST-, and ONT-based analyses, respectively. The overlap between the sites identified by these technologies is detailed in Supplementary Data 3. Figure 2 and Supplementary Fig. 1 (comparison of beta estimates) display the results obtained from the TWIST, ONT and EPIC EWAS.

Among the 33 associations identified as significant in the TWIST EWAS at $P < 3.6\text{x}10^{-8}$, two had been previously reported in the EWAS catalog (based on DNAm profiled with array technologies). These included *AHRR* (chr5-373263-373264, beta = −0.35, $P = 1.2\text{x}10^{-10}$) and an intergenic locus found on chromosome 2 (chr2-232419951-232419952, beta = −0.24, $P = 3.5\times10^{-8}$). The remaining 31 loci significant at this threshold included sites annotated to *F2RL3* (chr19-16889741-16889742, beta = -0.31, $P = 1.3\text{x}10^{-11}$) and *USP42* (chr7-6126706-6126707, beta=0.05, $P = 1.8\times10^{-8}$). At a less stringent threshold of $P < 1\times10^{-5}$, 98 uncatalogued sites were identified. They included *SST* (chr3-187670342-187670343, beta = −0.09, $P = 2.2\text{x}10^{-7}$) and *TSPAN5* (chr4-98472405-98472406, beta = -0.06, $P = 6.8\times10^{-6}$). Further details are provided in Supplementary Data 4.

In the ONT EWAS, 9 sites were significant at $P < 3.6\times10^{-8}$, of which only one had been previously listed in the EWAS catalog: a site mapping to *AHRR* (chr5-373263-373264, beta = -0.47, $P = 2.4\text{x}10^{-12}$). The remaining eight included additional loci within the *AHRR* region, loci from an intergenic region on chromosome 2 (e.g., chr2-232420079-232420080, beta = -0.37, $P = 3.4\text{x}10^{-10}$) and a site annotated to *CNTNAP2* (chr7-147245588-147245589, beta=0.37, $P = 1.1\times10^{-8}$). At $P < 1\times10^{-5}$, the ONT analysis revealed 62 uncatalogued loci, such as *SEPTIN9* (chr17-77351321-77351322, beta = -0.27, $P = 8.4\text{x}10^{-6}$) and *TERF2* (chr16-69398923-69398924, beta=0.12, $P = 8.7\text{x}10^{-6}$). Complete results are available in Supplementary Data 5.

A gene set enrichment analysis of genes mapped to the 102 CpGs with $P < 1\times10^{-5}$ identified in the TWIST EWAS revealed 13 enriched gene sets (FDR $p < 0.05$; see Supplementary Data 6). These included tissue degradation through altered extracellular matrix dynamics and chronic inflammation stemming from dysregulated ATP release and immune cell recruitment. Many of the enriched pathways were driven by the presence of collagen genes, such as *COL4A4* and *COL4A3*. In contrast, enrichment analysis based on the significant CpGs from the ONT EWAS did not identify any significantly enriched gene sets.

DNAm biomarker of cigarette consumption - mCigarette

A DNAm biomarker of cigarette consumption, mCigarette, was then developed. The biomarker was trained in GS (n = 17,865) using elastic net regression with 10-fold cross-validation. Prior to training, CpG sites were prefiltered to those associated with tobacco use at FDR < 0.05 (n_{CpG}=18,760) in the previous largest EWAS of smoking[6]. This meta-analysis did not include GS and was conducted using Illumina 450k BeadChip array. Following 10-fold cross validation, an optimal lambda value that minimised the mean prediction error was selected (lambda=0.012577) and fed into an elastic net model of smoking pack years. As a result, non-zero coefficients were assigned to 1,255 CpG sites (Supplementary Data 7).

The biomarker was tested in the external LBC1936 study (n = 882, mean age 69.6 years, SD = 0.8). Assessing the predictive performance using Area Under the Curve (AUC) (Fig. 3) revealed very good (AUC = 0.85) to near perfect (AUC = 0.98) ability to distinguish between current (n = 101), former (n = 368) and never smokers (n = 413). Similar results were obtained when Area Under the Precision-Recall Curve was used as a performance metric ($PRAUC_{\text{current vs never}} = 0.96$, $PRAUC_{\text{former vs never}} = 0.85$, PRAUC current vs former = 0.71).

The predictive performance of mCigarette was benchmarked against four previously developed epigenetic scores for smoking, as well as a single-site biomarker based on *AHRR* methylation at cg05575921, in wave 1 of LBC1936 (n = 882). The results of this comparison can be found in Table 1. mCigarette yielded improved incremental R^2 estimates compared to EpiSmokEr, the score developed by McCartney et al.[10] and GrimAge DNAm pack years.

https://doi.org/10.1038/s41467-025-58357-6

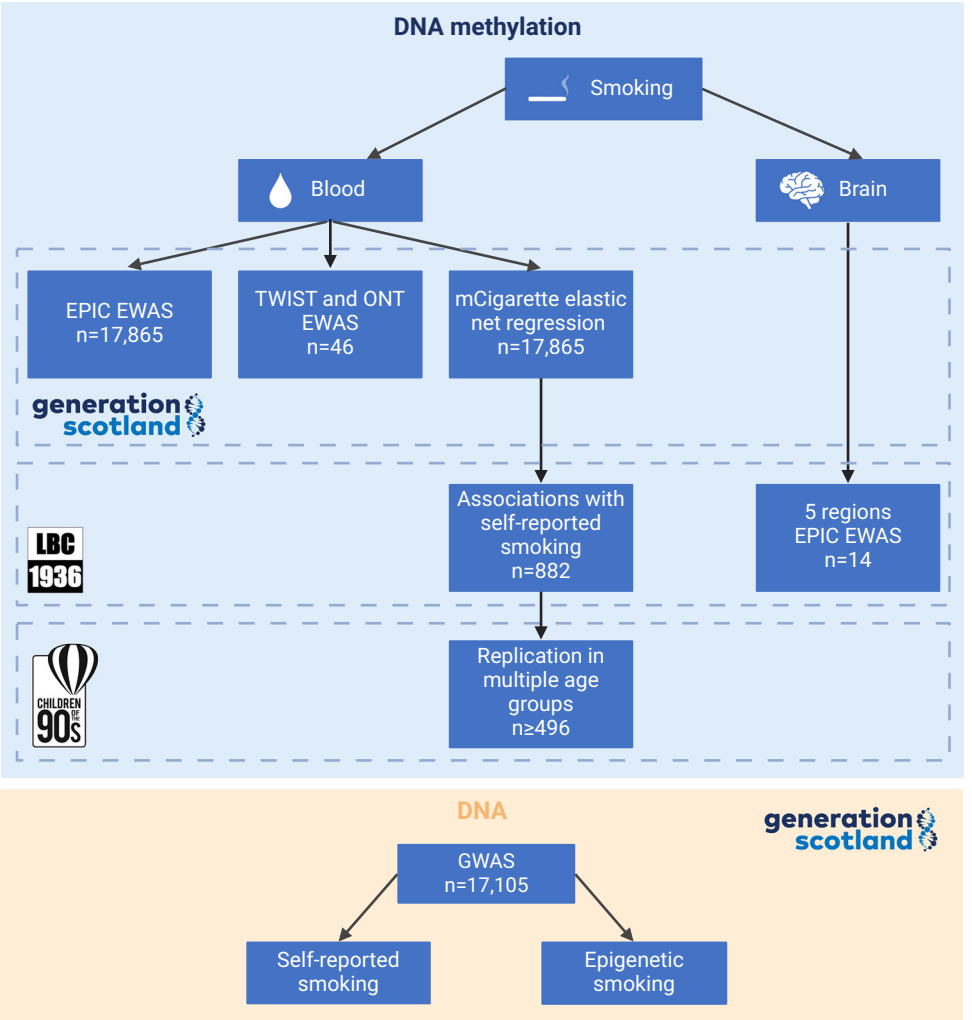


Fig. 1 | Project overview. The analysis was conducted using data from three British cohorts: Generation Scotland (GS), the Lothian Birth Cohort of 1936 (LBC1936) and the Avon Longitudinal Study of Parents and Children (ALSPAC). While blood-based DNA methylation (DNAm) data were available in all three cohorts, LBC1936 also contained information about the DNAm levels in post-mortem brain tissue, covering five brain regions from 14 individuals. A Bayesian Epigenome-Wide Association Study (EWAS) of smoking was performed in GS (~ 850k sites, Illumina EPIC array). For 23 pairs of age- and sex-matched smokers and non-smokers from GS, a high-resolution methylation measurement approach was implemented (~ 4 million sites, TWIST human methylome panel and ~21 million sites, Oxford Nanopore Technologies sequencing), followed by an EWAS analysis. An epigenetic biomarker of smoking, mCigarette, was developed in GS and tested as a predictor of self-reported smoking in LBC1936. The association between mCigarette and self-reported smoking was replicated in multiple age groups present in ALSPAC. Next, EWASs of smoking were run across five brain regions for 14 individuals using EPIC DNAm from LBC1936. Finally, Genome-Wide Association Studies (GWAS) were run in GS to compare genetic signal of self-reported and epigenetic smoking (GrimAge DNAm pack years score). EPIC – Illumina EPIC array, TWIST – TWIST Biosciences Human Methylome Panel, ONT – Oxford Nanopore Technologies Sequencing. Created in BioRender. Marioni, R. (2024) https://BioRender.com/h44z126.

Subsequently, the weights used to construct the studied scores (apart from the GrimAge DNAm pack years, as weights are not publicly available) were applied to methylation data in the ALSPAC cohort. Within this dataset, no single methylation score consistently outperformed others in distinguishing between current, former, and never smokers across studied age groups. While in young adults (mean age 18 and 24 years old) EpiSmokEr and the score constructed by McCartney et al.[10] achieved the highest AUCs (median AUC = 0.720 [IQR:0.642-0.763]), mCigarette and EpiSmokEr showed excellent performance in older adults (mean age 29 and 50 years old, median AUC = 0.890 [IQR:0.771-0.931]). Full results of the replication study are available in Supplementary Data 8 and are visualised in Supplementary Fig. 2.

Tissue specificity

To assess whether the findings translated across different tissue types, the association between tobacco use and DNAm levels across five brain regions were explored using *post-mortem* samples from LBC1936 ($n = 14$, $n_{hippocampus} = 13$). At significance threshold $P < 1x10^{-5}$, five loci in the hippocampus (BA35), one locus in dorsolateral prefrontal cortex (BA46), four loci in primary visual cortex (BA17), nine loci in anterior cingulate cortex (BA24) and three loci in ventral/lateral inferior temporal cortex (BA20/21) were associated with smoking status (Supplementary Data 9). There was no overlap between the significant loci across the studied brain regions.

Some loci demonstrated nearly perfect discrimination of smoking status in blood and brain; however, these loci did not overlap (Fig. 4 and Supplementary Fig. 3). For instance, the methylation status at cg05575921, annotated to the *AHRR* gene, is a well-established marker of smoking status in whole-blood DNAm. However, this marker did not discriminate smoking category in hippocampal DNAm. On the other hand, cg26381592, annotated to the *PMS1* gene, did not effectively distinguish smokers in blood samples, but it exhibited a strong correlation with smoking status in hippocampus samples.

https://doi.org/10.1038/s41467-025-58357-6

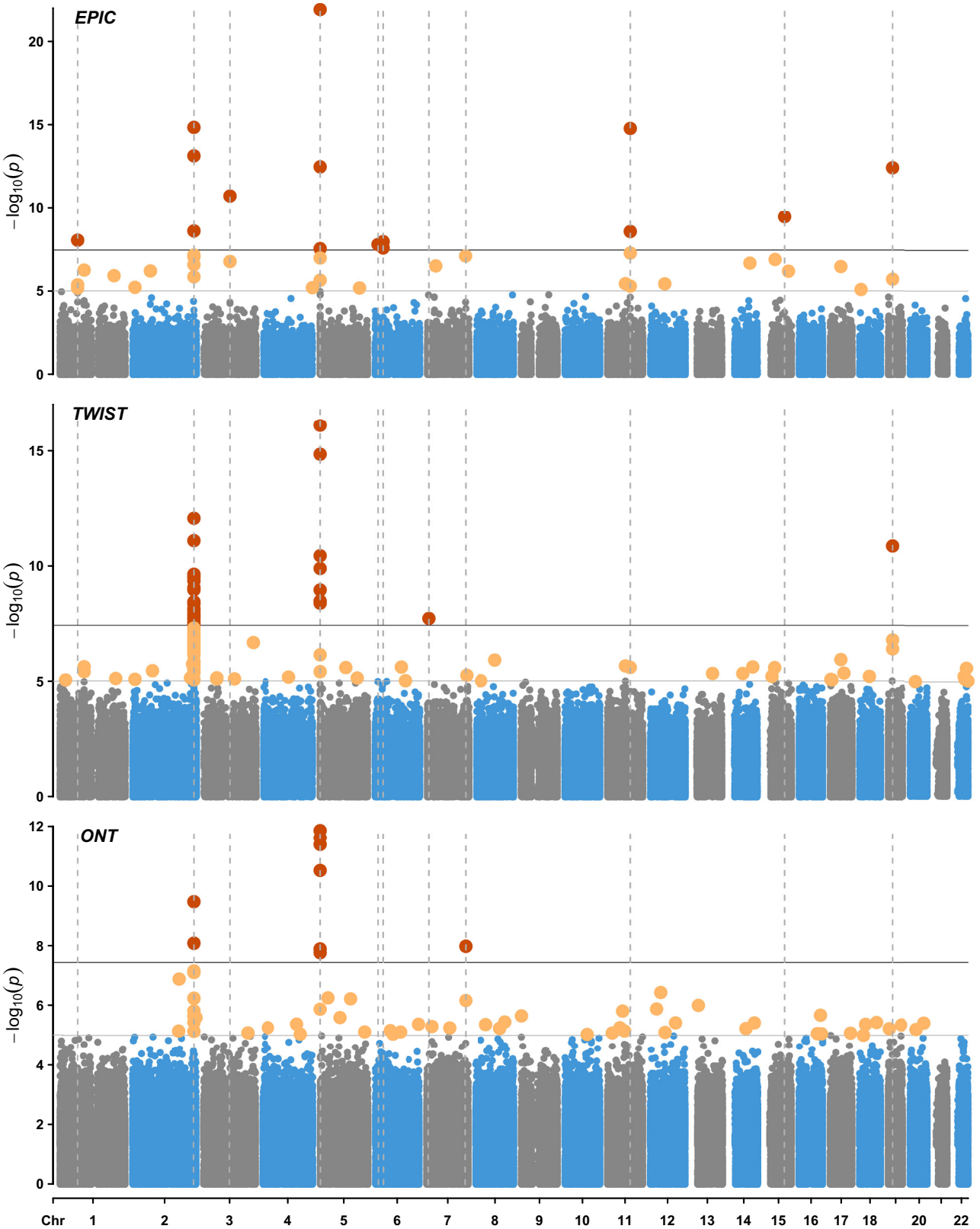


Fig. 2 | Epigenome-Wide Association Study (EWAS) of current versus never smokers in Generation Scotland ($n = 23$ pairs). Analyses were performed using DNA methylation data obtained using the Illumina EPIC array (~850k CpG sites), the TWIST human methylation panel (4 million CpG sites, targeted short read sequencing) and Oxford Nanopore Technologies sequencing (21 million CpG sites, long read sequencing). The X-axis represents chromosomes 1–22, while the Y-axis shows −log10(P-values). The top horizontal line marks genome-wide significant associations ($P < 3.6 \times 10^{-8}$, red dots), based on the multiple testing threshold estimated by Saffari *et al.*[13]. The bottom horizontal line denotes the suggestive significance threshold ($P < 1 \times 10^{-5}$, yellow dots). Dotted vertical lines highlight loci associated with smoking status at $P < 3.6 \times 10^{-8}$. All statistical tests were two-sided. EPIC – Illumina EPIC array, TWIST – TWIST Biosciences Human Methylome Panel, ONT – Oxford Nanopore Technologies Sequencing.

https://doi.org/10.1038/s41467-025-58357-6

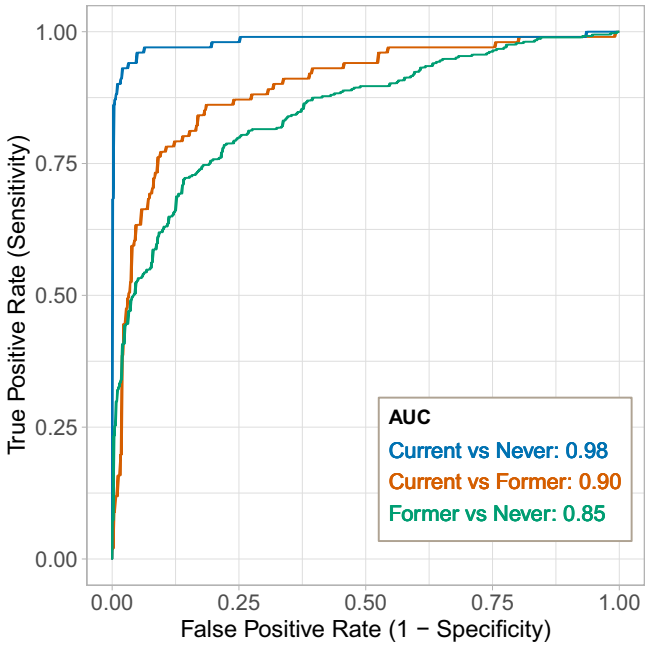


Fig. 3 | Predictive performance of the epigenetic biomarker of smoking (mCigarette). The Areas Under the Curves (AUCs) represent the ability of the model to distinguish between the following smoking categories: blue for current smokers vs. never smokers, orange for current smokers vs. former smokers, and green for former smokers vs. never smokers. Source data are provided as a Source Data file.

Subsequently, we tested the performance of a blood-derived DNAm biomarker of smoking (mCigarette) in the brain tissue. When applied to brain DNAm data, mCigarette did not distinguish between smoking categories in any of the studied brain regions (Supplementary Fig. 4).

GWAS of self-reported and epigenetic smoking

Finally, GWASs of self-reported and epigenetic smoking were conducted. To avoid overfitting, we used the GrimAge DNAm pack years estimator (trained externally in 1731 individuals from the Framingham Heart Cohort study[14]) instead of mCigarette, which was trained in GS.

In 17,105 GS individuals, there was a single nucleotide polymorphism (SNP)-based heritability of 27.3% ($P = 7.0\text{x}10^{-46}$) for self-reported pack years of smoking compared to 41.0% ($P = 8.2\times10^{-98}$) for GrimAge DNAm pack years (Fig. 5). The genomic inflation factors (λ) were 1.06 and 1.10 for pack years and GrimAge DNAm pack years, respectively. At the genome-wide significance level of $P < 5.0\text{x}10^{-8}$, only one locus (rs117836409 annotated to *GDPD1*) was associated with self-reported pack years, in contrast to 39 SNPs (three lead SNPs at two genomic risk loci) associated with GrimAge DNAm pack years (detailed in Supplementary Data 10 and 11). Of the three lead SNPs associated with GrimAge DNAm pack years, two (rs1800440 and rs6495309, annotated to *CYP1B1* and *CHRNA3 - CHRNB4*, respectively) have been previously documented in the GWAS catalogue, which aggregates data from published GWAS studies[15] (Supplementary Data 12). They have been associated with carcinogenesis and nicotine dependence, respectively[16,17]. The SNP not previously annotated in the GWAS Catalog (rs114342890) maps to *RMDN2:RMDN2-AS1*, a long non-coding RNA previously studied in relation to eosinophil counts and melanoma[18,19]. According to GeneHancer, an online database of enhancers, promoters and their inferred targets, all 27 regulatory elements which target *RMDN2:RMDN2-AS1* also regulate the expression of *CYP1B1*[20]. The beta coefficients of the lead loci identified in the GrimAge DNAm pack years GWAS and the smoking pack years GWAS are compared in Supplementary Fig. 5.

The three lead SNPs have been characterised as methylation quantitative trait loci (mQTLs). They all act in *cis* on 43 CpGs (Supplementary Data 13). Only one of these CpGs (cg06264984) is featured in mCigarette, while none are included in EpiSmokEr – the list of CpGs included for the GrimAge DNAm pack years is not publicly available.

Next, the GrimAge DNAm pack years GWAS results were compared to previously published GWAS studies of tobacco use (Supplementary Data 14). At a significance level of $P < 5\times10^{-8}$, seven SNPs annotated to *CHRNA3* and *CHRNA5* aligned with the findings of the largest pack years GWAS to date ($n = 131{,}892$)[21]. Thirty-seven SNPs, mapping to *CHRNA3, CHRNA5*, and *CHRNB4*, overlapped at $P < 5\text{x}10^{-8}$ with the results of a related phenotype, cigarettes per day ($n = 618{,}489$)[22].

GrimAge DNAm pack years and self-reported pack years were moderately correlated (Spearman's r = 0.65). The genetic correlation (r_g) between GrimAge DNAm pack years and pack years from Erzurumluoglu et al.[21] was 0.62 (SE = 0.12, $P = 4.4\text{x}10^{-7}$), with an LD score regression intercept of 0.99 (SE = 0.01). The r_g between self-reported pack years in the 17,105 GS and pack years from Erzurumluoglu et al.[21] ($n = 131{,}892$) was 0.67 (SE = 0.19, $P = 5.0\text{x}10^{-4}$), with an LD score regression intercept of 0.99 (SE = 0.01). Additional information on the genetic correlation between epigenetic smoking and previously studied self-reported smoking behaviours (ranging from −0.62 to 0.72)[22] can be found in Supplementary Fig. 6.

Discussion

This multi-tissue, multi-cohort analysis of the relationship between smoking and DNAm (assessed via arrays and sequencing) has improved both our understanding of the biological consequences of smoking and our ability to measure it objectively. The array-based study, which identified two loci not listed in the EWAS catalog as being associated with smoking, represents the largest single cohort EWAS of smoking and the largest EPIC array EWAS of smoking, to date. The updated epigenetic biomarker of tobacco-use, mCigarette, reliably predicted smoking status and was strongly correlated with self-reported pack years of tobacco use. The analysis of sites differentially methylated in the brains of smokers and non-smokers revealed evidence of tissue-specific signals. There was a partial overlap between the results of the GrimAge DNAm pack years GWAS conducted in GS ($n = 17{,}105$) and the most extensive GWAS of self-reported smoking to date ($n = 131{,}892$).

Among the loci not present in the EWAS catalog but identified in the EPIC array EWAS of smoking at PIP > 80%, five array-based CpGs are annotated to *FGF20*, *SCAMP5*, *GRIK5*, *SKI*, and *HOXA4*. *FGF20* plays a key role in the survival and function of dopamine-producing neurons[23], which are crucial in the brain's reward system[24] that nicotine stimulates. *SCAMP5* is essential for dopamine release[25] and the pleasurable sensations associated with smoking. It reinforces smoking behaviour and potentially make individuals more susceptible to nicotine addiction. *GRIK5* encodes a glutamate receptor[26] which modulates dopaminergic neurons within the brain's reward pathways, influencing how strongly these pathways respond to nicotine[27]. A proto-oncogene called *SKI* regulates cell growth and apoptosis[28]. It could potentially modify neural circuitry related to addiction, which affect the risk of developing nicotine dependence or influence the severity of addiction. Lastly, *HOXA4* is a homeobox domain gene that is normally involved in embryonic development[29]. Homeobox genes are abnormally expressed in cancer cells and changes in the expression of *HOXA4* has been specifically associated with colorectal, ovarian and lung cancer[29].

The findings from the next generation sequencing EWASs, which were less well-powered, also underscored the role of smoking in carcinogenesis and disrupted neurodevelopment. Significant loci not listed in the EWAS catalog ($P < 1\times10^{-5}$) identified in the TWIST EWAS included sites mapping to *TSPAN5*, which regulates tumour suppressor gene expression[30]; *USP42*, involved in head and neck cancer pathogenesis[31]; and *SST*, encoding somatostatin, a hormone

https://doi.org/10.1038/s41467-025-58357-6

Table 1 | Benchmarking of mCigarette against six biomarkers of smoking: EpiSmokEr[7], GrimAge DNA methylation (DNAm) pack years[14], and three scores developed using Generation Scotland (GS) data: BayesR score[59], a score developed by McCartney et al.[10] and two scores developed in this study – single-site biomarker based on AHRR (cg05575921) blood DNAm level and mCigarette

Metric	AHRR	EpiSmokEr	BayesR	McCartney et al.[10]	GrimAge	mCigarette
N training	17,865	1793	9448	5087	1731	17,865
a) Variance explained in measured pack years (R^2) and correlation (r) metrics						
Incremental R^2	0.329	0.351	0.514	0.330	0.419	0.534
r	0.589	0.610	0.735	0.581	0.664	0.750
b) Binary classification performance (AUC)						
Current / Never	0.972	0.985	0.977	0.982	0.983	0.984
Current / Former	0.930	0.916	0.853	0.921	0.939	0.897
Former / Never	0.742	0.755	0.846	0.725	0.815	0.852

The BayesR score and the score developed by McCartney et al.[10] were trained on smaller subsets of the GS DNAm dataset. Table a) compares the variance in self-reported pack years explained by null and full models. While the null model was adjusted for age and sex, the full model also included the studied score. The difference between variance explained by null and full models is denoted as the incremental R^2. Pearson's correlation coefficient is referred to as r. b) Performance of the studied scores in distinguishing between different smoking categories. Binary classification performance of the scores was measured using Areas Under the Curve (AUC).

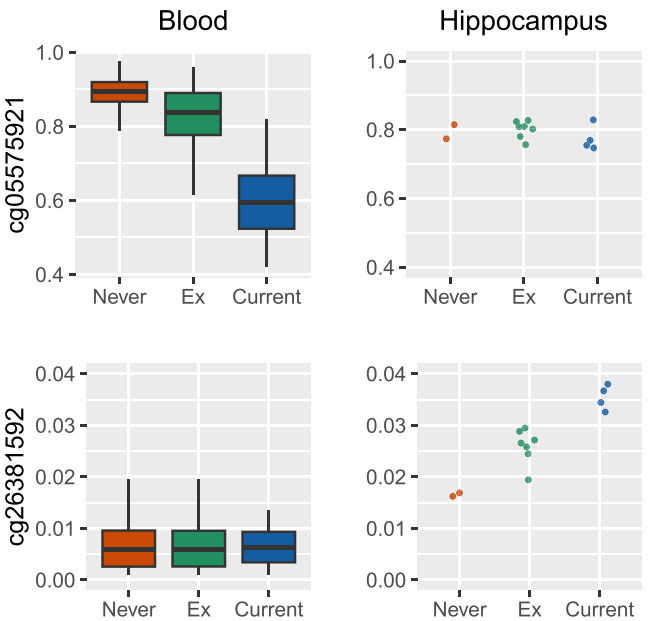


Fig. 4 | CpG methylation levels in blood (Lothian Birth Cohort 1936 baseline, $n_{blood\_DNAm} = 882$) and brain across different self-reported smoking categories (Lothian Birth Cohort 1936, $n_{hippocampus\_DNAm} = 13$). Box plots are defined as follows: the centre line represents the median (50th percentile). The box bounds indicate the interquartile range (IQR; 25th to 75th percentile). Whiskers extend to the smallest and largest values within 1.5×IQR. Source data are provided as a Source Data file.

implicated in the development of pancreatic cancer[32]. Significant loci revealed by ONT EWAS included CpGs associated with *SEPTIN9*, a tumour suppressor gene[33]; *TERF2*, a telomeric protein linked to tumour formation and progression[34]; and *CNTNAP2*, a potential marker of tumour aggressiveness in oligodendrogliomas[35], which is also implicated in neurodevelopmental disorders. An enrichment analysis of the TWIST results indicated that smoking influences methylation patterns across genes involved in extracellular matrix interactions, chronic inflammation, platelet activation, and ATP regulation, among other pathways. Again, the genes present in enriched pathways typically included *COL4A3* and *COL4A4* which play crucial roles in cancer invasion and metastasis[36]; and inflammation and remodelling of the lung extracellular matrix[37].

When compared to previously published epigenetic biomarkers of tobacco use, mCigarette was a better predictor of smoking pack years. It also showed an excellent performance in discriminating smoking status (current, former, never) in two external cohorts. The robust performance of mCigarette among pregnant women in ALSPAC may reflect the unique biological context of pregnancy, where hormonal changes and epigenetic plasticity make smoking-related epigenetic changes more pronounced. Loci which are most responsive to smoking during pregnancy could provide insights into changes transmitted to the foetus, and affecting the child's health later in life. In the future, mCigarette could be used to monitor smoking cessation efforts during pregnancy, ensuring compliance with cessation programs and potentially improving their effectiveness. The *AHRR* single-site biomarker provided a highly practical option for smoking classification, particularly in settings where limited data points or resources are available. However, multi-site models offered greater precision, accounting for a broader smoking-related methylation signature.

While individual CpG sites offered excellent discrimination of cases and controls, these loci varied by tissue. This is consistent with the low correlations in methylation patterns between blood and brain tissue reported previously[38]. Future work should explore if tissue-specific signals identify pathways and mechanisms by which smoking influences brain health.

In GS, the GrimAge DNAm pack years GWAS results did not align with self-reported smoking GWAS findings but did show partial overlap of lead loci with the most extensive GWAS of self-reported smoking to date. This may suggest an increased power to detect significant loci when the epigenetic score is analysed as a phenotype. However, the genetic correlation between GrimAge DNAm pack years and the meta-analysis smoking pack years was lower than the latter and self-reported pack years in GS. The moderate correlation could reflect differences in the biological pathways that phenotypic and epigenetic measures of smoking are capturing. The GrimAge-based DNAm estimator is designed to capture the cumulative biological impact of smoking, which might include broader aging-related processes beyond direct tobacco exposure. In contrast, the smoking pack years represent a more straightforward measure of cumulative smoking exposure. The shared loci between the GrimAge DNAm pack years GWAS and previous self-reported smoking GWAS included *CHRNA3, CHRNA5*, and *CHRNB4*. These genes encode subunits of the nicotinic acetylcholine receptor, responsible for neurotransmission and binding of nicotine in the brain. Variations in these genes can affect nicotine dependence and may be associated with neurological conditions as well as lung cancer[39,40].

https://doi.org/10.1038/s41467-025-58357-6

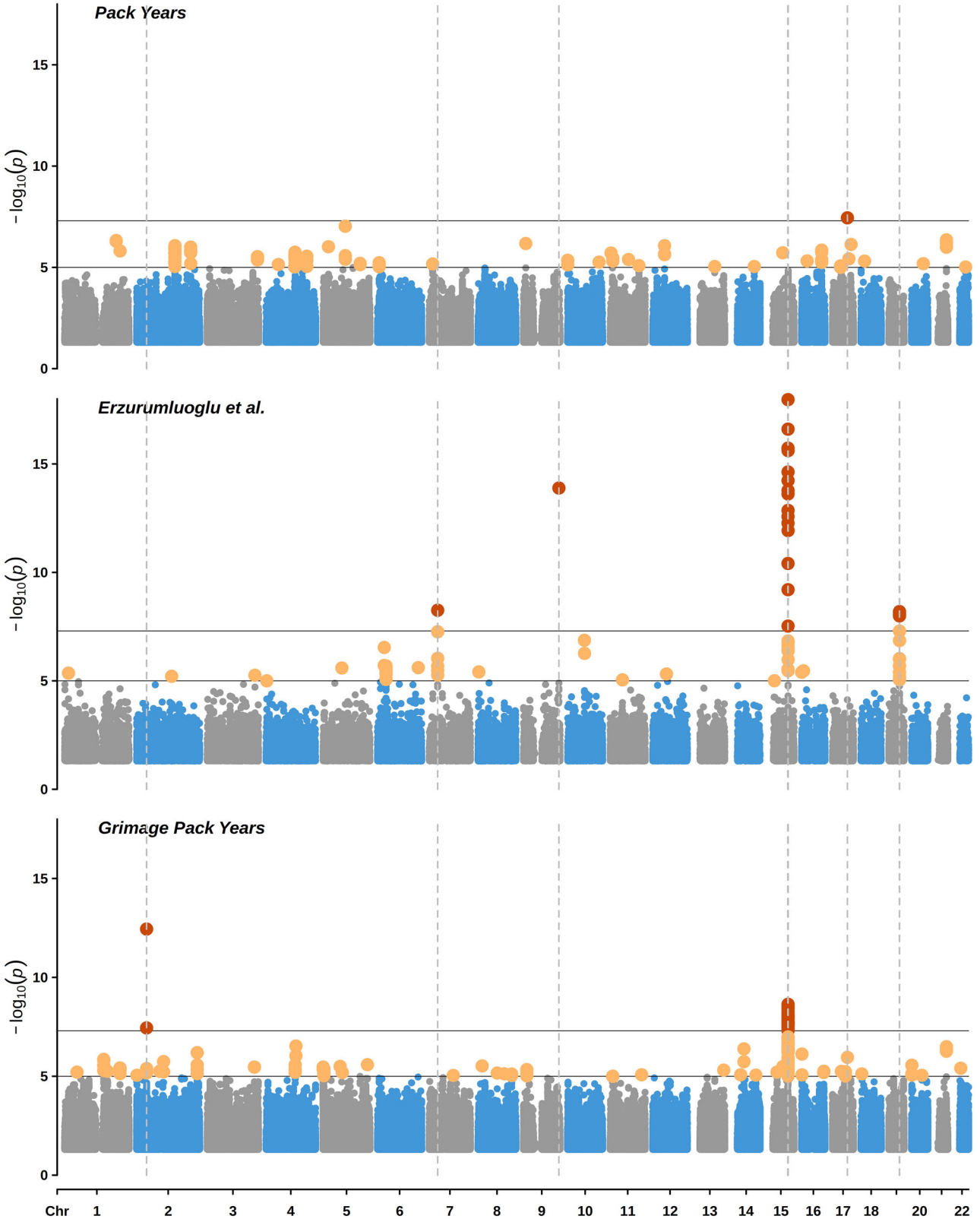


Fig. 5 | Manhattan plots visualising overlap between various Genome-Wide Association Studies (GWASs) of smoking. Studied phenotypes included: pack years (Generation Scotland, $n = 17{,}105$), pack years (largest meta-analysis to date[21], $n = 131{,}892$), GrimAge-based DNAm estimator of smoking pack years ($n = 17{,}105$). The X-axis represents genomic positions, while the Y-axis shows −log10(P-values). The top horizontal line marks genome-wide significant associations ($P < 5{\times}10^{-8}$, red dots), based on the Bonferroni threshold for multiple testing. The bottom horizontal line denotes the suggestive significance threshold ($P < 1{\times}10^{-5}$, yellow dots). Dotted vertical lines mark genomic positions of significant associations at $P < 5{\times}10^{-8}$. All tests were two sided.

The key strengths of this study include the high resolution of epigenetic data used in the main EWAS analysis (both sample size and the number of measured CpG sites), diverse DNAm profiling techniques and the availability of multi-tissue DNAm. Using targeted sequencing allowed for identifying associations between smoking and loci not included on the EPIC chip. Given the mean age of GS volunteers, mCigarette represents many years of exposure to cigarette smoke, and is likely to capture long term DNAm changes associated with smoking.

Limitations of this study include its potential lack of generalizability to non-Europeans and the small number of brain- and whole methylome samples. Nevertheless, given that smoking is associated with major epigenetic alternations, the effects of tobacco use are detectable, even in small datasets. The absence of serum cotinine concentrations prevented us from comparing mCigarette against a clinically established smoking biomarker. We were also unable to verify self-reported smoking status. To mitigate potential bias, we cross-referenced the reported smoking status with other variables (such as self-reported smoking initiation and cessation) and eliminated records with conflicting responses. Additionally, we acknowledge a limitation in our study regarding second-hand smoke exposure. In family-based cohorts such as GS, where related participants may live in the same household, passive smoking is an important factor that could confound the associations between direct smoking and health outcomes. Unfortunately, data on second-hand smoke exposure were not collected in this cohort.

In conclusion, this study explored methylation patterns associated with smoking in blood and brain. The blood-based analyses, using both sequencing- and array-based approaches, identified additional loci associated with tobacco use and led to the development of a highly accurate blood-based DNAm biomarker for smoking. Furthermore, the study provided insights into the differential effects of smoking across tissues. Together, these enhance our understanding of the epigenetic architecture of smoking and shed light on the molecular mechanisms by which tobacco use influences health.

Methods

Generation Scotland

Volunteer recruitment to GS has been detailed in a previous publication[41]. Between 2006 and 2011, patients of collaborating general medical practices in Scotland between 35 and 65 years of age were selected at random and invited to take part in the study. These individuals were then encouraged to recruit family members to volunteer to join the cohort. A total of 24,088 individuals between 18-99 years of age completed a health questionnaire. All individuals were asked to report their smoking status (classified as non-smoker, ex-smoker who stopped more than 12 months ago, ex-smoker who stopped within the past 12 months, and current smoker). In addition, current and former smokers provided information about the age they started smoking, the age they quit smoking (former smokers only), and the number of cigarettes they smoked per day. Pack years were computed by multiplying years of smoking by the number of cigarettes smoked per day divided by 20 (number of cigarettes in a pack), and assigning a value of zero to those who never smoked. Further information about the distribution of phenotypic data is available in Supplementary Data 1. DNA extracted from whole blood collected at the baseline visit was genotyped using the Illumina HumanOmniExpressExome array (8v1-2 and 8v1) for 19,992 individuals. Following quality control (QC), imputation to the Haplotype Reference Consortium (HRC) panel[42] and post imputation QC, 7,626,922 SNPs remained for downstream analyses. Methylation levels were measured with Illumina EPIC850k array (18,413 individuals, 752,722 sites after QC). GWAS and EWAS quality control steps have been described before[10] and are also documented in Supplementary Data 15 and 16. Phenotype pre-processing steps are detailed in Supplementary Fig. 7.

Lothian Birth Cohort 1936

The Lothian Birth Cohort of 1936 ($n = 1091$) comprises community-dwelling older adults in Scotland, most of whom completed an intelligence test aged around 11 years in 1947. Later in life, those living in the Edinburgh and Lothians region were recruited to the cohort at a mean age of ~70 years and then followed up at 3-yearly intervals. Data collected at each wave comprised cognitive test scores as well as biological measures obtained from blood samples. During a baseline interview at age 70, the participants' self-reported smoking status (never smoker, past smoker, current smoker) and smoking behaviour (age at starting, age at stopping, average number of cigarettes smoked per day) were determined. Pack years were calculated as in GS. A brain tissue bank was established at wave 3 (from age ~76 years). Detailed information about the cohort, brain imaging and post-mortem brain samples can be found in a cohort update and brain protocol papers[43,44]. DNAm from whole blood has been measured using Illumina Infinium HumanMethylation450 BeadChip array, while DNAm in five post-mortem brain tissues was profiled using the Illumina EPIC850k array[45]. Quality control and processing details are provided in Supplementary Data 15. Phenotype pre-processing steps are detailed in Supplementary Fig. 8.

ALSPAC

ALSPAC is a cohort study conducted among pregnant women residing in Avon, UK, with expected delivery dates falling between 1st April 1991 and 31st December 1992[46,47]. Out of the 20,248 eligible pregnancies, 14,541 were enrolled to the study, resulting in 14,062 live births, of which 13,988 children survived to age one. During pregnancy, mothers invited fathers to take part in the study. A total of 12,113 fathers completed questionnaires, with 3807 currently formally enrolled. For a subset of ALSPAC participants (mothers, fathers and children) DNAm was assayed as part of the Accessible Resource for Integrated Epigenomic Studies (ARIES) initiative[48,49]. DNA was extracted from blood samples collected at various time intervals between birth and death, and methylation levels were measured using the Illumina Infinium HumanMethylation450 or MethylationEPIC BeadChip arrays. 450,838 CpG sites passed quality control and were common to these methylation arrays. This study used four collections of DNAm data. Antenatal collection includes data from the ALSPAC mothers only, the Focus on Mothers (FOM)/ Focus on Fathers (FOF) collection corresponds to the mothers/fathers at midlife (~ 50 years), and the F17 and F24 collections contain ALSPAC children at ages 15–17 (time-point '15up') and 24 (time-point 'F24'), respectively[50]. Smoking status for F17 was based on three questionnaires administered ages 14–17. Former smokers reported having quit at the age 17 questionnaire. Current smokers reported that they had smoked weekly in one or more questionnaires but had not quit. Never smokers reported having never smoked at least one time and never reported having smoked. Smoking status for F24 was assessed using six questionnaires administered ages 14-24. Former smokers reported having smoked regularly at some point prior to age 24, but at age 24 reported not having smoked in the previous 30 days. Current smokers reported at age 24 having smoked in the last 30 days and having smoked at least 50 cigarettes in their lifetime. Never smokers reported never having smoked at age 24. Maternal antenatal smoking was assessed by questionnaires administered at 18- and 32-weeks gestation. Former smokers reported having smoked previously but have stopped smoking for the pregnancy. Current smokers reported smoking regularly in the first trimester. Never smokers reported having never smoked before or during the pregnancy. Smoking status of FOM mothers was assessed using 15 questionnaires administered at study child ages up to age 12 and a final questionnaire at age 18. Former smokers reported having smoked on at least one questionnaire but reported not smoking on the 18 y questionnaire. Current smokers reported being current regular smokers on the 18 y questionnaire. Never smokers consistently reported never having smoked on questionnaires. Smoking status of FOF fathers was

assessed using 11 questionnaires administered at study child ages up to age 12 and a final questionnaire at age 20. Former smokers reported having smoked on at least one questionnaire but reported not smoking on the 20 y questionnaire. Current smokers reported being current regular smokers on the 20 y questionnaire. Never smokers consistently reported never having smoked on questionnaires.

Data for the study were gathered and administered utilizing REDCap electronic data capture tools, which are hosted at the University of Bristol. REDCap (Research Electronic Data Capture) is a secure web application specifically designed to facilitate data capture for research[51]. The study website provides comprehensive details on all available data, accessible through a fully searchable data dictionary located at http://www.bristol.ac.uk/alspac/researchers/our-data/.

Inclusion & ethics

This study is based on self-reported biological sex (cross-referenced with genetic data). Detailed sex distributions are provided in Supplementary Data 1. Sex/gender was not a primary focus of this study, and analyses were conducted on the full cohort to maximize statistical power. Participants in GS, LBC1936 and ALSPAC did not receive major financial compensation for participation.

All components of GS received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89). All participants provided broad and enduring written informed consent for biomedical research. This study was performed in accordance with the Helsinki declaration.

Ethical approval for the LBC1936 study was obtained from the Multi-Centre Research Ethics Committee for Scotland (MREC/01/0/56) and the Lothian Research Ethics committee (LREC/1998/4/183; LREC/2003/2/29). Use of human tissue for post-mortem studies has been reviewed and approved by the Edinburgh Brain Bank ethics committee and the ACCORD medical research ethics committee, AMREC (ACCORD is the Academic and Clinical Central Office for Research and Development, a joint office of the University of Edinburgh and NHS Lothian). All participants provided written informed consent. These studies were performed in accordance with the Helsinki declaration.

Ethical approval for the ALSPAC study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Consent for biological samples has been collected in accordance with the Human Tissue Act (2004). Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Sequencing-based approach

Whole methylome sequencing data were generated for 48 unrelated smokers and non-smokers from GS as part of this study. Two approaches were used: the TWIST methylome panel (~ 4 million CpG sites) and ONT sequencing (~ 21 million CpG sites). To ensure a robust methylation signal would be present when comparing cases against controls, only heavy smokers (12 males and 12 females chosen from a pool of 40 potential cases with tobacco use ranging from 53.9 to 87.7 pack years; see Supplementary Fig. 9) were selected as cases. Additional details on the sample selection process are provided in the Supplementary Methods. Controls were matched by age and sex using the Matchit package in R[52], with the maximum age difference of less than 12 months (0 years) within a matched pair. Cases and controls showed a clear separation in terms of their methylation at the *AHRR* CpG probe cg05575921 (Supplementary Fig. 10).

Sequencing using the TWIST Human Methylome Panel was performed by the Genetics Core, Edinburgh Clinical Research Facility according to TWIST Targeted Methylation Sequencing Protocol[53]. Sequencing using the ONT kit was performed by Edinburgh Genomics (the first 24 libraries, without basecalling) and the Genetics Core, Edinburgh Clinical Research Facility on the Oxford Nanopore PromethION 24, with R10.4.1 flow cells, running for 72 hours. Further details available in Supplementary Methods. For TWIST pre-processing of raw FASTQ files, read aligning to human reference genome (GRCh38, $n = 29{,}401{,}795$ total reference CpGs) with bwa-meth and quality-control of the results was performed using MethylSeq bioinformatics analysis pipeline, version 2.2.0[54,55]. This analysis yielded information about the methylation level and depth of coverage (DoC) at 18,248,472 covered CpG sites.

Dorado, optimized for NVIDIA GPUs, was used for high-accuracy basecalling and modified base detection in raw ONT data. Reads were aligned to GRCh38 human reference genome. Variants were called with epi2me-labs/wf-human-variation nextflow pipeline, version 23.10.1. Methylation level and depth of coverage was measured at 28,989,402 covered CpG sites.

The bedGraph (TWIST) and bedMethyl (ONT) files were subsequently processed in R version 4.3.1[56] using Methrix package[57]. As part of post-processing, loci a) of extremely low (minimal DoC = 2) and high coverage (beyond 0.99 quantile), and b) overlapping with known cytosine to thymine polymorphisms were removed from the methylation dataset. Finally, a coverage filter was applied, retaining only the loci that were covered in at least 40 samples by: a) 10 or more reads (TWIST), b) 5 or more reads (ONT). This left an analysis sample of 3,391,718 and 21,167,712 CpGs for TWIST and ONT data, respectively. CpG sites were annotated with Annotatr R package[58].

Blood-based DNAm EWAS

A blood-based EWAS of smoking was carried out in 17,865 GS individuals using BayesR[59]. Before running the EWAS, each CpG was corrected for the effects of age, sex, and batch using linear regression (saving the residuals from each model as the new variable). As BayesR implicitly corrects for confounding effects without requiring a full characterization of all covariates, our models were only adjusted for measured variables i.e., estimated white blood cell proportions were not included. However, we conducted a sensitivity analysis by running an EWAS that included adjustments for estimated cell proportions. The results of this sensitivity analysis showed a strong overlap with the primary findings and are presented in Supplementary Data 17. The smoking phenotype (measured in pack years, with a pack defined as 20 cigarettes) was natural log+1 transformed and adjusted for age and sex using linear regression (again, the residuals were saved and used for the downstream analyses). Both the smoking and the CpG variables were scaled to have mean of 0 and variance of 1. The data served as inputs of a Bayesian penalised linear regression model. Four Gaussian priors were specified to model CpGs with varying effect sizes (mixture variances of 0.1%, 1%, 10% and 100%) along with a discrete spike at the origin to model CpGs with no effect. A Gibbs sampler was used to sample over the posterior distribution, conditioning on the input data. A burn-in of 5000 samples was used, after which every fifth sample was retained across 10,000 iterations. A CpG with a posterior inclusion probability of greater than 0.95 was considered as epigenome-wide significant. Previously unpublished associations were identified by searching the literature and the EWAS catalog[12].

Comparison between TWIST, ONT and EPIC850k

DNAm of 24 pairs of smokers and non-smokers from GS was profiled using the EPIC array, ONT kit, and TWIST platform (see Methods - Sequencing-based approach). During quality control, one pair was identified as mismatched due an age gap exceeding the pre-defined threshold of 12 months. To preserve the integrity of the study design and prevent potential biases in the analysis, this pair was excluded from subsequent analyses. This left an analysis sample of 46 individuals. For each DNAm profiling method, an EWAS of smoking (pack

years) was conducted. The association between DNAm level at each CpG site (outcome) and binary smoking status, age and sex was modelled using linear regression. The results were displayed on a Manhattan plot generated with the CMplot R package[60]. Gene names associated with CpG sites reaching a suggestive significance threshold ($P < 1x10^{-5}$) were extracted and subjected gene set enrichment analysis in Functional Mapping and Annotation (FUMA) GENE2FUNC tool[61], which implements a hypergeometric test. An FDR-adjusted p-value threshold of 0.05 was applied, and a minimum of 2 overlapping genes within each gene set was required.

Biomarkers of cumulative smoking

An elastic net biomarker of pack years (mCigarette) was trained in 17,865 GS individuals using the glmnet library in R[62]. As part of data pre-processing, CpG sites were filtered to 18,760 loci associated with smoking at False Discovery Rate (FDR) < 0.05 in a previous meta-analysis EWAS ($n = 18,760$) of tobacco use[6]. Alpha was fixed at 0.5 and the lambda value that minimised the mean prediction error was selected via 10-fold cross validation. The selected model assigned non-zero coefficients to 1255 CpGs. A single-site biomarker for smoking, based on methylation at *AHRR* (cg05575921), was also trained in the same subset of GS individuals ($n = 17,865$) using linear regression. Both mCigarette and the single-site biomarker were tested in wave 1 of LBC1936 ($n = 882$, mean age 70 years), while mCigarette alone was tested in ALSPAC ($n = 496–1207$ across four time points). These biomarkers were benchmarked against three epigenetic scores for smoking: EpiSmokEr score[7,63,64], and two scores derived from previous GS analyses on smaller subsets of the dataset - one based on BayesR weights[59], the other via lasso penalised regression by McCartney et al.[10]. In LBC1936, the predictive performance of mCigarette was additionally compared to that of GrimAge DNAm pack years[14]. Pearson's r was calculated to estimate the degree of correlation between self-reported pack years of smoking and the studied scores. The amount of variance in pack years explained by the studied scores was assessed by comparing R^2 estimates of null and full models. While the null model was adjusted for age and sex, the full model also included the studied score. Incremental R^2 was calculated as the difference between variance explained by null and full models.

The ability of the scores to distinguish between current, former, and never smokers was assessed by AUC. Receiver operating characteristic (ROC) curves were produced using pROC R package[65]. Additional prediction performance metrics such as PRAUC were obtained using MLmetrics R package[66].

Tissue specificity analyses in LBC1936

DNAm was measured in 5 brain regions (hippocampus - BA35, dorsolateral prefrontal cortex – BA46, primary visual cortex - BA17, anterior cingulate cortex - BA24, ventral/lateral inferior temporal cortex - BA20/21) from post-mortem brain samples of 14 LBC1936 individuals, with one sample missing from hippocampus. Tissue acquisition and processing details are detailed in Stevenson et al.[45]. Using blood-DNAm measured at wave 1 and brain-DNAm data, exploratory EWAS analyses were performed (CpG ~ smoking category). In these analyses, smoking was treated as a continuous variable encoded as 0 = never smoker, 1 = former smoker, 2 = current smoker. Given the small sample size, nominally significant CpG-smoking associations were defined as having $P < 1x10^{-5}$ and were displayed using a ggplot2 boxplot. Due to the same constraints, an enrichment analysis for significant associations was not carried out.

GWAS of smoking

Associations between genetic variants and smoking were examined using GWASs. Two phenotypes were considered: natural log(-transformed pack years of smoking + 1) and an epigenetic score for smoking pack years generated by an online calculator that uses the algorithm derived for the GrimAge epigenetic clock[14]. After initial filtering (see Supplementary Data 3), there were 17,105 GS individuals with data available at 7,626,922 SNPs. Both GWASs were conducted using the GCTA software[67], with a Genetic Relationship Matrix fitted into a fastGWA-lmm model to account for relatedness. Each trait was adjusted for age, sex, and 20 genetic principal components. The results of these analyses, along with the findings of previous GWASs of tobacco use, were visualised using CMplot[60]. Lead and methylation quantitative loci among the results of GrimAge DNAm pack years GWAS were identified using the default settings in FUMA[61] and goDMC[68], respectively. GWAS catalog was accessed via FUMA website. Genetic correlations were calculated with LDSC[69].

Data availability

Cohort data are available under restricted access. According to the terms of consent for Generation Scotland participants, access to data must be reviewed by the Generation Scotland Access Committee. Applications should be made to genscot@ed.ac.uk and normally take up to six weeks for approval. Further details can be found at https://genscot.ed.ac.uk/for-researchers/access/. Lothian Birth Cohort data are available on request from the Lothian Birth Cohort Study, University of Edinburgh. Information on data access can be found at https://lothian-birth-cohorts.ed.ac.uk/data-access-collaboration, including data dictionaries and a Data Request Form (DRF). Data access requests, including a completed DRF, should be sent to the study director (simon.cox@ed.ac.uk); requests are normally reviewed by the team within four weeks and data are shared subject to completion of a Data or Material Transfer Agreement. ALSPAC is run as a resource for the research community. Instructions for accessing ALSPAC data can be found here: https://www.bristol.ac.uk/alspac/researchers/access/. A research proposal must be submitted via the research proposal system for consideration by the ALSPAC Executive Committee. For any questions regarding accessing data or samples please email alspac-data@bristol.ac.uk (data) or bbl-info@bristol.ac.uk (samples). Approval may take up to two weeks. The raw data underlying figures are provided in the Supplementary Information/Source Data file. The GWAS and EWAS summary statistic output is available in the Zenodo database [https://doi.org/10.5281/zenodo.14878399]. For any further correspondence and material requests please contact Dr Riccardo Marioni at riccardo.marioni@ed.ac.uk. Source data are provided with this paper.

Code availability

All custom R (version 4.3.1), Python (version 3.9.7), and bash code is available with open access at the following Zenodo repository: https://doi.org/10.5281/zenodo.14882848[70].

References

1. GBD 2019 Tobacco Collaborators. Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990–2019: a systematic analysis from the Global Burden of Disease Study 2019. *Lancet* **397**, 2337–2360 (2021).
2. West, R. Tobacco smoking: Health impact, prevalence, correlates and interventions. *Psychol. Health* **32**, 1018–1036 (2017).
3. Connor Gorber, S., Schofield-Hurwitz, S., Hardt, J., Levasseur, G. & Tremblay, M. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. *Nicotine Tob. Res* **11**, 12–24 (2009).
4. Murphy, S. E., Wickham, K. M., Lindgren, B. R., Spector, L. G. & Joseph, A. Cotinine and trans 3'-hydroxycotinine in dried blood spots as biomarkers of tobacco exposure and nicotine metabolism. *J. Expo. Sci. Environ. Epidemiol.* **23**, 513–518 (2013).
5. Breitling, L. P. Current genetics and epigenetics of smoking/tobacco-related cardiovascular disease. *Arteriosclerosis, Thrombosis, Vasc. Biol.* **33**, 1468–1472 (2013).

6. Joehanes, R. et al. Epigenetic signatures of cigarette smoking. *Circulation: Cardiovascular Genet.* **9**, 436–447 (2016).
7. Bollepalli, S., Korhonen, T., Kaprio, J., Anders, S. & Ollikainen, M. EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data. *Epigenomics* **11**, 1469–1486 (2019).
8. Christiansen, C. et al. Novel DNA methylation signatures of tobacco smoking with trans-ethnic effects. *Clin. Epigenetics* **13**, 36 (2021).
9. McCartney, D. L. et al. Epigenetic signatures of starting and stopping smoking. *EBioMedicine* **37**, 214–220 (2018).
10. McCartney, D. L. et al. Epigenetic prediction of complex traits and death. *Genome Biol.* **19**, 136 (2018).
11. Hoang, T. T. et al. Comprehensive evaluation of smoking exposures and their interactions on DNA methylation. *eBioMedicine* **100**, 104956 (2024).
12. Battram, T. et al. The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Res* **7**, 41 (2022).
13. Saffari, A. et al. Estimation of a significance threshold for epigenome-wide association studies. *Genet Epidemiol.* **42**, 20–33 (2018).
14. Lu, A. T. et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY)* **11**, 303–327 (2019).
15. Sollis, E. et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977–D985 (2022).
16. Flora, A. V. et al. Functional characterization of SNPs in *CHRNA3/B4* intergenic region associated with drug behaviors. *Brain Res.* **1529**, 1–15 (2013).
17. Choquet, H. et al. Multi-ancestry genome-wide meta-analysis identifies novel basal cell carcinoma loci and shared genetic effects with squamous cell carcinoma. *Commun. Biol.* **7**, 33 (2024).
18. Ransohoff, K. J. et al. Two-stage genome-wide association study identifies a novel susceptibility locus associated with melanoma. *Oncotarget* **8**, 17586–17592 (2017).
19. Kichaev, G. et al. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am. J. Hum. Genet* **104**, 65–75 (2019).
20. Fishilevich, S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxf.)* **2017**, bax028 (2017).
21. Erzurumluoglu, A. M. et al. Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci. *Mol. Psychiatry* **25**, 2392–2409 (2020).
22. Saunders, G. R. B. et al. Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature* **612**, 720–724 (2022).
23. Boshoff, E., Fletcher, E. & Duty, S. Fibroblast growth factor 20 is protective towards dopaminergic neurons in vivo in a paracrine manner. *Neuropharmacology* **137**, 156 (2018).
24. Baik, J.-H. Stress and the dopaminergic reward system. *Exp. Mol. Med* **52**, 1879–1890 (2020).
25. Chen, Y., Fan, J., Xiao, D. & Li, X. The role of SCAMP5 in central nervous system diseases. *Neurological Res.* **44**, 1024–1037 (2022).
26. Unlu, G. et al. GRIK5 genetically regulated expression associated with eye and vascular phenomes: discovery through iteration among biobanks, electronic health records, and zebrafish. *Am. J. Hum. Genet* **104**, 503–519 (2019).
27. D'Souza, M. S. & Markou, A. The "Stop" and "Go" of Nicotine Dependence: Role of GABA and Glutamate. *Cold Spring Harb. Perspect. Med* **3**, a012146 (2013).
28. Peng, Y. et al. Ski promotes proliferation and inhibits apoptosis in fibroblasts under high-glucose conditions via the FoxO1 pathway. *Cell Prolif.* **54**, e12971 (2020).
29. Cheng, S. et al. HOXA4, down-regulated in lung cancer, inhibits the growth, motility and invasion of lung cancer cells. *Cell Death Dis.* **9**, 1–13 (2018).
30. Schreyer, L. et al. Tetraspanin 5 (TSPAN5), a Novel Gatekeeper of the Tumor Suppressor DLC1 and Myocardin-Related Transcription Factors (MRTFs), Controls HCC Growth and Senescence. *Cancers (Basel)* **13**, 5373 (2021).
31. Rong, C. et al. Ubiquitin carboxyl-terminal hydrolases and human malignancies: the novel prognostic and therapeutic implications for head and neck cancer. *Front. Oncol.* **10**, 592501 (2021).
32. Manoochehri, M. et al. SST gene hypermethylation acts as a pan-cancer marker for pancreatic ductal adenocarcinoma and multiple other tumors: toward its use for blood-based diagnosis. *Mol. Oncol.* **14**, 1252–1267 (2020).
33. Sun, J., Zheng, M.-Y., Li, Y.-W. & Zhang, S.-W. Structure and function of Septin 9 and its role in human malignant tumors. *World J. Gastrointest. Oncol.* **12**, 619–631 (2020).
34. Iachettini, S. et al. The telomeric protein TERF2/TRF2 impairs HMGB1-driven autophagy. *Autophagy* **19**, 1479 (2022).
35. Rautajoki, K. J. et al. PTPRD and CNTNAP2 as markers of tumor aggressiveness in oligodendrogliomas. *Sci. Rep.* **12**, 14083 (2022).
36. Nie, X. et al. COL4A3 expression correlates with pathogenesis, pathologic behaviors, and prognosis of gastric carcinomas. *Hum. Pathol.* **44**, 77–86 (2013).
37. Zheng, T., Zheng, Z., Zhou, H., Guo, Y. & Li, S. The multifaceted roles of COL4A4 in lung adenocarcinoma: An integrated bioinformatics and experimental study. *Computers Biol. Med.* **170**, 107896 (2024).
38. Walton, E. et al. Correspondence of DNA methylation between blood and brain tissue and its application to schizophrenia research. *Schizophr. Bull.* **42**, 406–414 (2016).
39. Wassenaar, C. A. et al. Relationship Between CYP2A6 and CHRNA5-CHRNA3-CHRNB4 Variation and Smoking Behaviors and Lung Cancer Risk. *JNCI J. Natl Cancer Inst.* **103**, 1342 (2011).
40. Kawamata, J. & Shimohama, S. Association of novel and established polymorphisms in neuronal nicotinic acetylcholine receptors with sporadic Alzheimer's disease. *J. Alzheimers Dis.* **4**, 71–76 (2002).
41. Smith, B. H. et al. Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med Genet* **7**, 74 (2006).
42. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet* **48**, 1279–1283 (2016).
43. Henstridge, C. M. et al. Post-mortem brain analyses of the Lothian Birth Cohort 1936: extending lifetime cognitive and brain phenotyping to the level of the synapse. *Acta Neuropathologica Commun.* **3**, 53 (2015).
44. Taylor, A. M., Pattie, A. & Deary, I. J. Cohort Profile Update: The Lothian Birth Cohorts of 1921 and 1936. *Int. J. Epidemiol.* **47**, 1042–1042r (2018).
45. Stevenson, A. J. et al. A comparison of blood and brain-derived ageing and inflammation-related DNA methylation signatures and their association with microglial burdens. *Eur. J. Neurosci.* **56**, 5637–5649 (2022).
46. Boyd, A. et al. Cohort Profile: the'children of the 90s'–the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J. Epidemiol.* **42**, 111–127 (2013).
47. Fraser, A. et al. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J. Epidemiol.* **42**, 97–110 (2013).
48. Relton, C. L. et al. Data resource profile: accessible resource for integrated epigenomic studies (ARIES). *Int. J. Epidemiol.* **44**, 1181–1190 (2015).
49. Min, J. L., Hemani, G., Davey Smith, G., Relton, C. & Suderman, M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics* **34**, 3983–3989 (2018).
50. Northstone, K. et al. The Avon Longitudinal Study of Parents and Children (ALSPAC): an update on the enrolled sample of index children in 2019. *Wellcome Open Res* **4**, 51 (2019).

51. Harris, P. A. et al. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inf.* **42**, 377–381 (2009).
52. Ho, D., Imai, K., King, G. & Stuart, E. A. MatchIt: nonparametric preprocessing for parametric causal inference. *J. Stat. Softw.* **42**, 1–28 (2011).
53. Twist Targeted Methylation Sequencing Protocol. https://www.twistbioscience.com/resources/protocol/twist-targeted-methylation-sequencing-protocol (2023).
54. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
55. Ewels, P. et al. nf-core/methylseq: nf-core/methylseq version 1.6.1 [Nauseous Serpent]. *Zenodo* https://doi.org/10.5281/zenodo.4744708 (2021).
56. R: The R Project for Statistical Computing. https://www.r-project.org/ (2023).
57. Mayakonda, A. et al. Methrix: an R/Bioconductor package for systematic aggregation and analysis of bisulfite sequencing data. *Bioinformatics* **36**, 5524–5525 (2020).
58. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).
59. Trejo Banos, D. et al. Bayesian reassessment of the epigenetic architecture of complex traits. *Nat. Commun.* **11**, 2865 (2020).
60. Yin, L. et al. rMVP: A memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics, Proteom. Bioinforma.* **19**, 619–628 (2021).
61. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
62. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
63. Zeilinger, S. et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLOS ONE* **8**, e63812 (2013).
64. Elliott, H. R. et al. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. *Clin. Epigenetics* **6**, 4 (2014).
65. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinforma.* **12**, 77 (2011).
66. Yan, Y. CRAN - Package MLmetrics. https://cran.r-project.org/web/packages/MLmetrics/index.html (2024).
67. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet* **88**, 76–82 (2011).
68. Min, J. L. et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat. Genet* **53**, 1311–1321 (2021).
69. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet* **47**, 291–295 (2015).
70. Chybowska, A. aleksandra-chybowska/Smoking_EpiScore. Zenodo https://doi.org/10.5281/zenodo.14882848 (2025).

Acknowledgements

We sincerely appreciate all Generation Scotland study participants, staff, and research team members for their past and ongoing contributions to these studies. 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 Generation Scotland samples was carried out by the Genetics Core Laboratory at the Edinburgh Clinical Research Facility, Edinburgh, Scotland and was funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award STratifying Resilience and Depression Longitudinally (STRADL; Reference 104036/Z/14/Z). The DNA methylation data assayed for Generation Scotland was partially funded by a 2018 NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation (Ref: 27404; awardee: Dr David M Howard) and by a JMAS SIM fellowship from the Royal College of Physicians of Edinburgh (Awardee: Dr Heather C Whalley). The authors thank all LBC1936 study participants and research team members who have contributed, and continue to contribute, to ongoing studies. The LBC1936 is supported by the BBSRC, and the Economic and Social Research Council [BB/W008793/1] (which supports S.E.H.), Age UK (Disconnected Mind project), the Milton Damerel Trust, the Medical Research Council (MR/M01311/1), and the University of Edinburgh. Methylation typing of LBC1936 was supported by the Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), Age UK, The Wellcome Trust Institutional Strategic Support Fund, The University of Edinburgh, and The University of Queensland. Genotyping was funded by the BBSRC (BB/F019394/1). S.R.C. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 221890/Z/20/Z). We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team. The UK Medical Research Council and Wellcome (Grant ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors and they will serve as guarantors for the contents of this paper. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf). Funding for ALSPAC DNAm measurements were supported by the Wellcome (102215/2/13/2); the University of Bristol; the UK Economic and Social Research Council (ES/N000498/1); the UK Medical Research Council (MC_UU_12013/1, MC_UU_12013/2); the Biotechnology and Biological Sciences Research Council (BBI025751/1 and BB/I025263/1); and the John Templeton Foundation (60828). P.Y. and M.S. work is supported by the National Institute for Health and Care Research Bristol Biomedical Research Centre, the Medical Research Council Integrative Epidemiology Unit at the University of Bristol (MC_UU_00032/3, MC_UU_00032/4, MC_UU_00032/6), and Cancer Research UK [C18281/A29019, EDDISA-Jan22\100003]. A.D.C. is supported by a Medical Research Council PhD Studentship in Precision Medicine with funding from the Medical Research Council Doctoral Training Program and the University of Edinburgh College of Medicine and Veterinary Medicine. R.F.H is supported by an MRC IEU Fellowship. E.B. and R.E.M. are supported by Alzheimer's Society major project grant AS-PG-19b-010. 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 221890/Z/20/Z). For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Author contributions

A.D.C. analysed the data. E.B. and R.F.H. developed the Bayesian EWAS pipeline. P.Y. and M.S. replicated results in the ALSPAC cohort. D.L.M., R.F.H., R.C., L.McG., L.M., S.E.H., J.C., A.C., T.L.S., S.R.C., and K.L.E. were involved in the data generation. A.D.C. and R.E.M. drafted the initial manuscript. A.D.C., J.F.P., K.L.E., and R.E.M. designed the study. All authors read and approved the final manuscript.

Competing interests

R.E.M. is an advisor to the Epigenetic Clock Development Foundation. R.F.H. has received consultant fees from Illumina. R.E.M. and R.F.H. have received consultant fees from Optima partners. L.M. received speaker fees from Illumina and Oxford Nanopore Technologies. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-58357-6.

Correspondence and requests for materials should be addressed to Riccardo E. Marioni.

Peer review information *Nature Communications* thanks Suzanne Martos and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

5.3. Conclusion

The EPIC array EWAS of smoking pack years (n > 17,000) identified 42 CpG sites with a PIP > 80%, including nine loci not listed in the EWAS catalog. These loci included genes involved in dopaminergic signalling, addiction biology, and carcinogenesis. Complementary NGS EWASs revealed additional sites implicated in cancer development and neurodevelopmental disruption. Although individual CpGs almost perfectly discriminated smokers from non-smokers, these sites did not overlap across blood and brain tissue.

The updated smoking EpiScore, mCigarette, demonstrated strong performance in predicting smoking status in British cohorts and showed a high correlation with self-reported pack years. Finally, I observed a partial overlap between genetic signals of self-reported and epigenetic smoking (GrimAge DNAm pack-years).

This multi-cohort, multi-tissue study investigated the relationship between smoking and DNAm using both array-based and NGS approaches. The findings enhance our understanding of smoking's biological effects and demonstrate the utility of the mCigarette smoking EpiScore as a robust, objective biomarker of tobacco exposure. In the following chapter, I turn to other potential epigenetic predictors of CVD: protein EpiScores.

6. Protein EpiScores as Biomarkers of CVD

6.1. Introduction

Protein EpiScores have emerged as promising biomarkers for CVD and its sub-conditions, including IHD and stroke, with some demonstrating stronger associations with clinical outcomes than the corresponding measured protein levels. One of the first studies to generate protein EpiScores for multiple proteins and link them to health outcomes was conducted by Gadd *et al*. [1]. Building on this foundation, the work presented in this chapter focuses on protein EpiScores that could improve the predictive performance of established clinical CVD risk tools, including ASSIGN and SCORE2, and potentially provide insights into the molecular mechanisms underlying disease risk.

The analysis employed data of 12 657 adults from GS study (n_{events} = 1274) who were followed-up for 16 years. I started by examining associations between 109 Gadd's EpiScores and time-to-CVD using Cox PH models, running one model per EpiScore. Basic models were adjusted for the ASSIGN, reflecting established clinical risk factors, while fully adjusted models additionally incorporated cTnI levels to account for subclinical cardiac injury. Next, I conducted a prediction analysis by training a composite CVD EpiScore using elastic net regression ($n_{training}$ = 6880 GS individuals), with all 109 protein EpiScores as potential input features. This composite score was then evaluated in an independent test sample (n_{test} = 3659 GS individuals) to determine whether it could enhance risk prediction beyond ASSIGN and SCORE2. This two-step approach – first evaluating individual protein contributions, then combining them into a single predictive score – enabled a comprehensive assessment of both the biological relevance and practical utility of protein EpiScores for CVD risk stratification.

This study was published in Circulation: Genomic and Precision Medicine in February 2024. The supplementary material for this study can be accessed in the following repository https://github.com/aleksandra-chybowska/thesis/tree/main/Epigenetic_contributions_to_clinical_risk_prediction_of_cardiovascular_disease and through the electronic links provided by the publisher. The code used to perform this analysis can be found in https://github.com/aleksandra-chybowska/troponin_episcores.

6.2. Epigenetic Contributions to CVD Risk Prediction

Circulation: Genomic and Precision Medicine

ORIGINAL ARTICLE

Epigenetic Contributions to Clinical Risk Prediction of Cardiovascular Disease

Aleksandra D. Chybowska, MSci; Danni A. Gadd, MSc; Yipeng Cheng, MSc; Elena Bernabeu, PhD; Archie Campbell, MA; Rosie M. Walker, PhD; Andrew M. McIntosh, PhD; Nicola Wrobel, BSc; Lee Murphy, MSc; Paul Welsh, PhD; Naveed Sattar, PhD; Jackie F. Price, MB, ChB, MD; Daniel L. McCartney, PhD; Kathryn L. Evans, PhD; Riccardo E. Marioni, PhD

BACKGROUND: Cardiovascular disease (CVD) is among the leading causes of death worldwide. The discovery of new omics biomarkers could help to improve risk stratification algorithms and expand our understanding of molecular pathways contributing to the disease. Here, ASSIGN—a cardiovascular risk prediction tool recommended for use in Scotland—was examined in tandem with epigenetic and proteomic features in risk prediction models in ≥12 657 participants from 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 cTnI (cardiac troponin I). The associations between individual protein EpiScores and the CVD risk were examined using Cox regression (n_{cases}≥1274; $n_{controls}$≥11 383) and visualized in a tailored R application. Splitting the cohort into independent training (n=6880) and test (n=3659) subsets, a composite CVD EpiScore was then developed.

RESULTS: Sixty-five protein EpiScores were associated with incident CVD independently of ASSIGN and the measured concentration of cTnI (P<0.05), over a follow-up of up to 16 years of electronic health record linkage. The most significant EpiScores were for proteins involved in metabolic, immune response, and tissue development/regeneration pathways. A composite CVD EpiScore (based on 45 protein EpiScores) was a significant predictor of CVD risk independent of ASSIGN and the concentration of cTnI (hazard ratio, 1.32; P=3.7×10^{-3}; 0.3% increase in C-statistic).

CONCLUSIONS: EpiScores for circulating protein levels are associated with CVD risk independent of traditional risk factors and may increase our understanding of the etiology of the disease.

Key Words: biomarkers ■ cardiovascular diseases ■ epigenomics ■ multiomics ■ troponin

See Editorial by Bozack et al

For the past 20 years, cardiovascular disease (CVD) has been among the leading causes of mortality and morbidity worldwide. Given that many CVD cases are preventable, it is important to identify at-risk individuals early, when an intervention is most likely to be effective, and translate this knowledge into preventative strategies.[1,2] Although there are many CVD risk prediction algorithms, currently, they have limited predictive performance.[3] It may be possible to improve on that by discovering novel factors strongly associated with the disease, for example, the type and the concentrations of proteins expressed as a response to the damage to the cardiovascular system.

Correspondence to: Riccardo E. Marioni, PhD, Centre for Genomic and Experimental Medicine, The University of Edinburgh, Western General Hospital, Crewe Rd, Edinburgh EH4 2XU, United Kingdom. Email riccardo.marioni@ed.ac.uk

Supplemental Material is available at https://www.ahajournals.org/doi/suppl/10.1161/CIRCGEN.123.004265.
For Sources of Funding and Disclosures, see page 45.

Circulation: Genomic and Precision Medicine is available at www.ahajournals.org/journal/circgen

Nonstandard Abbreviations and Acronyms

ADM	adrenomedullin
cTnC	cardiac troponin C
cTnI	cardiac troponin I
cTnT	cardiac troponin T
CVD	cardiovascular disease
DNAm	DNA methylation
EpiScores	epigenetic scores
GDF15	growth differentiation factor 15
GS	Generation Scotland
HR	hazard ratio
NT-proBNP	N-terminal pro-B-type natriuretic peptide
PH	proportional hazard

Several proteins have been highlighted as possible biomarkers for CVD. These include GDF15 (growth differentiation factor 15), NT-proBNP (N-terminal pro-B-type natriuretic peptide), and ADM (adrenomedullin).[4–7] An established and highly sensitive marker of myocardial damage is cardiac troponin.[8] It is a complex of 3 proteins, namely, cTnI (cardiac troponin I), cTnT (cardiac troponin T), and cTnC (cardiac troponin C) regulating the contraction of the cardiac muscle. Cardiac forms of troponin T[9,10] and troponin I are expressed almost exclusively in the heart.[11] Following myocyte damage, cardiac troponin enters the circulation and can be detected in blood samples. A high-sensitivity cardiac troponin test plays a role in the rapid diagnosis of myocardial infarction.[8] Low-grade elevations in cardiac troponin are associated with an increased risk of CVD.[8]

Individual differences in protein concentration can be well captured by DNA methylation (DNAm). DNAm is a type of epigenetic modification characterized by the addition of methyl groups to DNA. Typically, the methyl group is added to cytosine-phosphate-guanine dinucleotides that are found mostly (but not exclusively) in gene promoters.[12] Blocking promoters, to which activating transcription factors should bind to initiate transcription, is one of the mechanisms by which DNAm can precisely regulate gene expression.[13] Conversely, changes in DNAm patterns can also be a result of changes in gene expression and chromatin state.[14,15]

DNAm-based proxies for protein levels are referred to as protein epigenetic scores (EpiScores) and are broadly analogous to polygenic risk scores. These methylation scores can be derived from penalized linear regression models of protein concentrations. Due to their temporal stability, protein EpiScores may exhibit stronger associations with disease outcomes than singular protein measurements, which are known to fluctuate between measurements.[16–19] We have shown that EpiScores for 109 circulating protein levels are associated with the time to diagnosis for a host of leading causes of morbidity and mortality, including cardiovascular outcomes.[20] Protein EpiScores are, therefore, useful biomarker tools for disease risk stratification.

Here, we examine whether protein EpiScores, calculated for ≥12 657 participants of the Generation Scotland (GS), study can augment predictions made by a CVD risk calculator developed for use in Scotland (ASSIGN[21]). We first run individual Cox proportional hazard (PH) models to discover relationships between individual protein EpiScores and incident CVD. We then create a CVD EpiScore (based on the protein EpiScores) and test the additional predictive performance offered by it for CVD risk stratification. A graphical overview of the analyses is presented in Figure 1.

METHODS

All methods are described in the Supplemental Material. A key resource in this study, GS, is a family-based research initiative focusing on genetic and environmental factors influencing health. Briefly, from 2006 to 2011, eligible individuals were selected from participating general medical practices in Scotland and invited at random to take part in the study.[22] All participants provided written informed consent for research. The study received ethical approval from the National Health Service Tayside Committee on Medical Research Ethics (REC reference number: 05/S1401/89). The GS data set is not publicly available as it contains information that could compromise participant consent and confidentiality. However, the data, research materials, and analytical methods will be made accessible to other researchers for the purpose of replicating the findings. Access will be granted upon successful project application to the GS Access Committee and obtaining ethical approval for accessing linked health data from NHS Scotland. Instructions for accessing GS data can be found at https://www.ed.ac.uk/generation-scotland/for-researchers/access; the GS Access Request Form can be downloaded from this site.

RESULTS

Clinical Risk Prediction Tools

ASSIGN scores were calculated for 16 366 individuals with nonmissing risk factor data. To meet the PH assumption of the Cox model, the data set was filtered to individuals aged between 30 and 70 years (results split by decade are presented in Table SI) and trimmed of outliers (points beyond 3 SDs of the mean; n=181). This left a cohort of 12 790 individuals, which was further filtered to records with nonmissing concentrations of cTnI (n=12 657). Table 1 summarizes the training, test, and full data sets.

Incremental Model Using Cardiac Troponin and Cardiac Troponin EpiScores

We tested whether concentrations of cardiac troponin were associated with CVD risk above ASSIGN over 16 years of follow-up. While the measured concentration of cTnI was associated with a hazard ratio (HR) of 1.20 per SD increase in the full (n=12 657) cohort (95% CI,

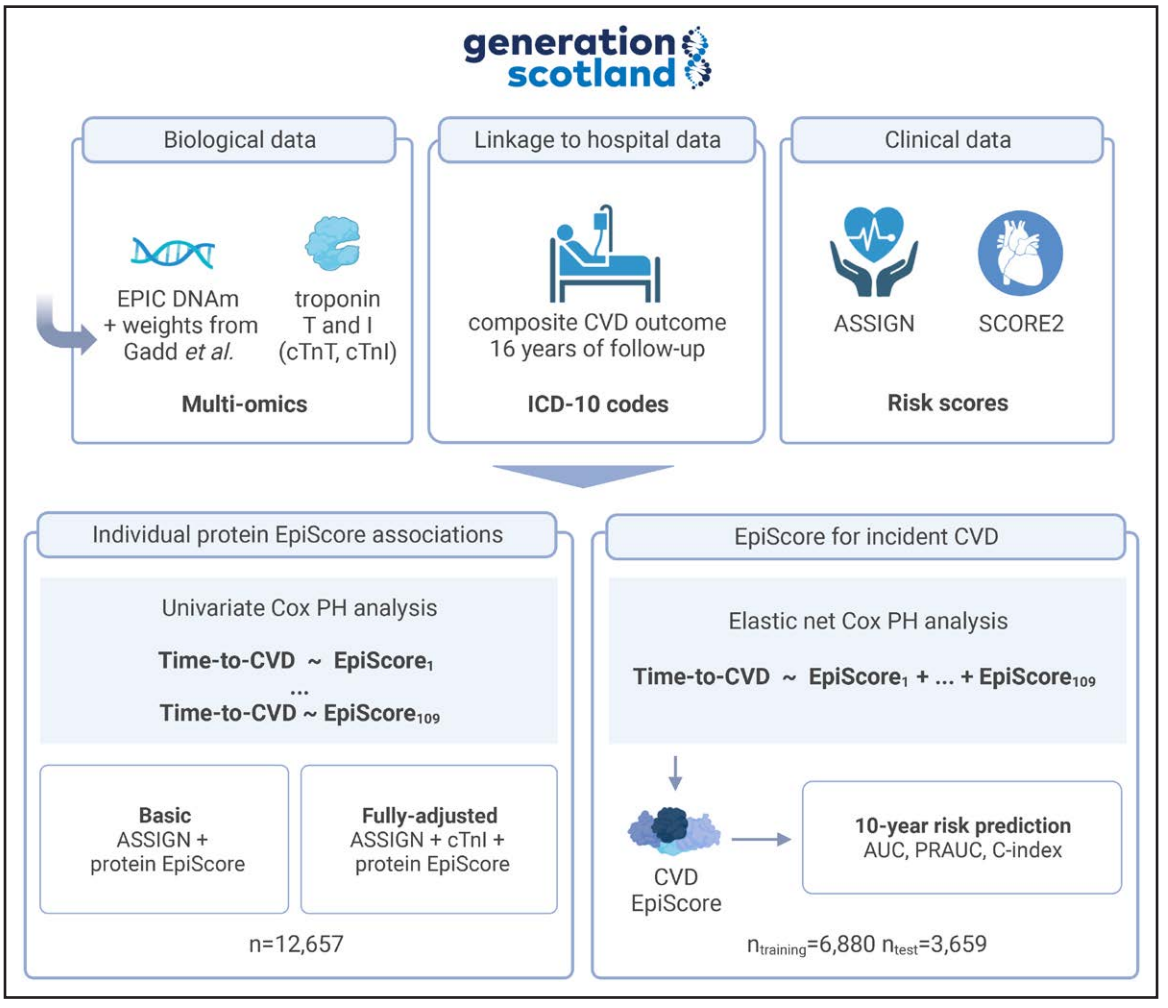


Figure 1. Project overview.
A series of Cox proportional hazard (PH) models were run to model the relationship between time-to-cardiovascular disease (CVD) and 109 protein epigenetic scores (EpiScores). Basic models were adjusted for the ASSIGN score, whereas fully adjusted models also included the concentration of cTnI (cardiac troponin I). This was followed by a prediction analysis where a composite protein EpiScore was trained. The CVD EpiScore was derived using elastic net and 109 protein EpiScores as possible input features. The score was assessed in the test sample to quantify the additional predictive performance offered by it over and above ASSIGN and SCORE2. The test Cox PH models were adjusted for age, sex, cTnI, and the CVD EpiScore, with time-to-CVD as the outcome. ASSIGN indicates the cardiovascular risk score chosen for use by SIGN (Scottish Intercollegiate Guidelines Network) and Scottish Government Health Directorates; AUC, area under the receiver operating characteristic curve; cTnT, cardiac troponin T; PRAUC, area under the precision recall curve; and SCORE2, an algorithm derived, calibrated, and validated to predict 10-year risk of first-onset CVD in European populations. Created with BioRender.com

1.13–1.29; P=1.9×10^{-8}), an EpiScore generated for cTnI (see Methods for details) was not associated with the measured concentrations in the n=3659 test set (incremental R^2, 0.027%; P=0.31) and did not predict CVD risk in Cox models adjusted for ASSIGN in the same test set (P=0.59). For that reason, it was not considered a feature 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[20] (protein description available in Table SII) were associated with CVD risk over 16 years of follow-up (n=12 657; n_{events}=1274).

First, we generated 109 Cox PH CVD risk models adjusted for ASSIGN. Each model was additionally adjusted for a different protein EpiScore. Two EpiScores failed to satisfy the PH assumption (Schoenfeld residual test P>0.05), and 6 EpiScores were not unique (proxied the concentration of the same protein). Of the remaining 101 protein EpiScores, 67 were significantly associated with CVD risk (P<0.05). After applying a conservative Bonferroni threshold for multiple testing (P<0.05/101=5.0×10^{-4}), 36 associations remained statistically significant.

Secondly, to understand whether protein EpiScores were associated with CVD risk beyond established biomarkers such as cardiac troponin, we included the concentration of cTnI as a covariate in the model along with ASSIGN, and we repeated the analysis. Of the 101

Table 1. Summary of Training, Test, and Full Data Sets. The Full Data Set Contains Related Individuals

	Training		Test		Full	
	Cases	Controls	Cases	Controls	Cases	Controls
n	658	6222	337	3322	1274	11 383
Time-to-event (years to onset or censoring)	7.0 (4.1–9.9)	11.8 (11.1–13.0)	4.8 (2.6–7.6)	11.8 (11.0–13.6)	6.8 (3.7–9.8)	11.8 (11.1–13.2)
Age, y	58.3 (50.8–62.6)	50.0 (40.8–58.8)	56.6 (51.6–60.0)	51.5 (43.9–57.6)	57.4 (51.0–62.2)	50.4 (41.5–58.2)
Sex, male	345 (52.4%)	2452 (39.4%)	165 (49.0%)	1219 (36.7%)	655 (51.4%)	4399 (38.6%)
SIMD, score/10	41.6 (22.2–53.3)	45.3 (26.3–55.1)	45.1 (20.2–54.9)	44.5 (22.6–54.8)	41.8 (21.6–54.0)	44.9 (25.5–54.9)
Family history of CHD/stroke, yes	443 (67.3%)	3171 (51.0%)	224 (66.5%)	1781 (53.6%)	862 (67.7%)	5881 (51.7%)
Diabetes, yes	19 (2.9%)	63 (1.0%)	16 (4.7%)	72 (2.2%)	43 (3.4%)	165 (1.4%)
Rheumatoid arthritis, yes	29 (4.4%)	140 (2.3%)	23 (6.8%)	110 (3.3%)	71 (5.6%)	281 (2.5%)
Nonsmoker, yes	534 (81.2%)	5306 (85.3%)	280 (83.1%)	2752 (82.8%)	1044 (81.9%)	9623 (84.5%)
Systolic blood pressure, mm Hg	142.1 (16.7)	130.8 (16.9)	140.3 (17.6)	130.3 (16.5)	141.5 (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 (11–28)	9 (4–17)

To make sure that members of the same family are not present across training and test data sets, any individuals in the training set who shared family ID with individuals from the test set were excluded from subsequent analyses (n=2118). For continuous variables with normal distributions, summary values are reported as mean (SD). Median (Q1-Q3) are given for continuous variables that do not follow a normal distribution. A number and a percentage of samples are reported for categorical variables.

ASSIGN indicates the cardiovascular risk score chosen for use by SIGN (Scottish Intercollegiate Guidelines Network) and Scottish Government Health Directorates; CHD, coronary heart disease; HDL, high-density lipoprotein; ID, identification number; and SIMD, Scottish Index of Multiple Deprivation.

aforementioned protein EpiScores, 65 were associated with CVD over and above the ASSIGN score and the concentration of cTnI ($P<0.05$; Figure 2). Thirty-three associations remained significant after correcting for multiple tests. Of the 65 protein EpiScores, higher levels of 41 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 per SD of 1.23 (95% CI, 1.16–1.30; $P=9.2\times10^{-12}$) and 1.13 (95% CI, 1.06–1.22; $P=5.4\times10^{-4}$; Figure 3A), respectively. In contrast, higher levels of 24 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, 0.84 [95% CI, 0.79–0.89]; $P=1.6\times10^{-9}$) and OMD (HR per SD, 0.87 [95% CI, 0.82–0.92]; $P=1.0\times10^{-6}$). The relationships between individual EpiScores and CVD risk have been visualized in the form of risk-over-time (Figure 3B), forest, and Kaplan Meier plots in an online R application (https://shiny.igc.ed.ac.uk/3d2c8245001b4e67875ddf2ee3fcbad2/).

As DNAm levels vary between different types of white blood cells, there is a concern that the associations that we observe may be influenced by cellular heterogeneity. To mitigate this potential effect, we incorporated estimated white blood cell proportions as covariates in the model adjusted for the concentration of cTnI and the ASSIGN score. In this model, 50 protein EpiScores were significantly associated with CVD risk ($P<0.05$). The comparison of HRs associated with protein EpiScores in each of the studied models can be found in Table SIII.

Finally, to learn whether individual protein EpiScore can augment CVD prediction beyond established biomarkers and clinical risk prediction tools, we calculated C-statistics for null and full models. While the null model was adjusted for ASSIGN and the concentration of cTnI (C-stat, 0.728), the full model also contained the studied protein EpiScore. Table 2 lists the top 10 associations that result in the greatest improvement in CVD risk prediction.

Composite Episcore for CVD Risk Prediction

To understand whether the abovementioned protein EpiScores can be used as biomarkers that add additional predictive value over and above typically used clinical risk scores (ASSIGN and SCORE2) and the concentration of cTnI, we generated a composite CVD EpiScore—a weighted linear combination of individual protein EpiScores. The score was trained using 2 modeling techniques: Cox PH Elastic Net and Random Survival Forest. There were 6880 records in the training set and 3659 records in the test set. The Elastic Net assigned nonzero coefficients to 45 of 109 protein EpiScores (Table SIV).

In a 10-year Elastic Net prediction analysis, the null model (containing age, sex, and ASSIGN) had an area under the receiver operating characteristic curve (AUC) of 0.719. The model with the CVD EpiScore increased the AUC to 0.723. The addition of cTnI to the null model resulted in an AUC of 0.721. The full model (null model+cTnI+CVD EpiScore) AUC was 0.724. Full output for the CVD models including C-statistics and a comparison with SCORE2 can be found in Tables V through VII. These analyses were a carbon copy of the aforementioned ASSIGN models—a null model (containing age, sex, and

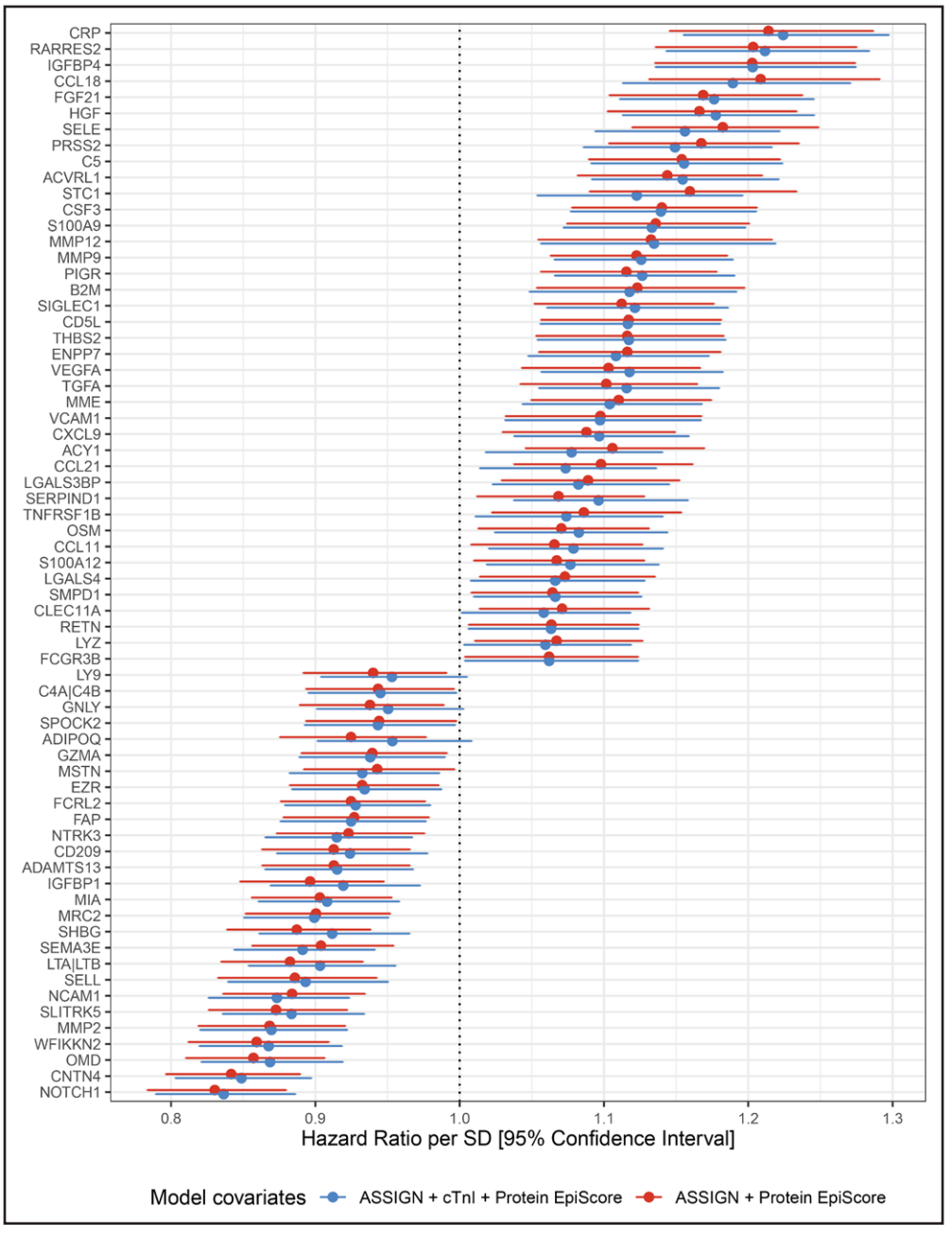


Figure 2. Associations between protein epigenetic scores (EpiScores) and incident cardiovascular disease.
Hazard ratios are plotted for the 67 significant associations (P<0.05) with 95% CI limits. Basic models were adjusted for ASSIGN (red), whereas full models included the ASSIGN score and concentration of cTnI (cardiac troponin I) as covariates (blue).

SCORE2) was compared with models with cTnI and the CVD EpiScore. The CVD EpiScore remained statistically significant after adjusting for the concentration of cTnI in models incorporating ASSIGN and SCORE2 (HR, 1.32; $P=3.7\times10^{-3}$ and HR, 1.36; $P=1.4\times10^{-3}$, respectively).

Random Survival Forest–based analysis (see Methods) yielded similar results. The null model (as above) had an AUC of 0.719. Adding the CVD EpiScore to the null model increased the AUC to 0.721. The full model adjusted for CVD EpiScore and the concentration of cardiac troponin had an AUC of 0.723.

DISCUSSION

In this study, we describe 65 novel epigenetic biomarkers that are associated with long-term risk of CVD independently of a clinical risk prediction tool (ASSIGN) and the concentration of an established protein biomarker (cTnI). The most statistically significant EpiScores reflected concentrations of proteins involved in metabolic, immune, and developmental pathways. A weighted linear combination of protein EpiScores (the composite protein-CVD EpiScore) was significantly associated with CVD risk in

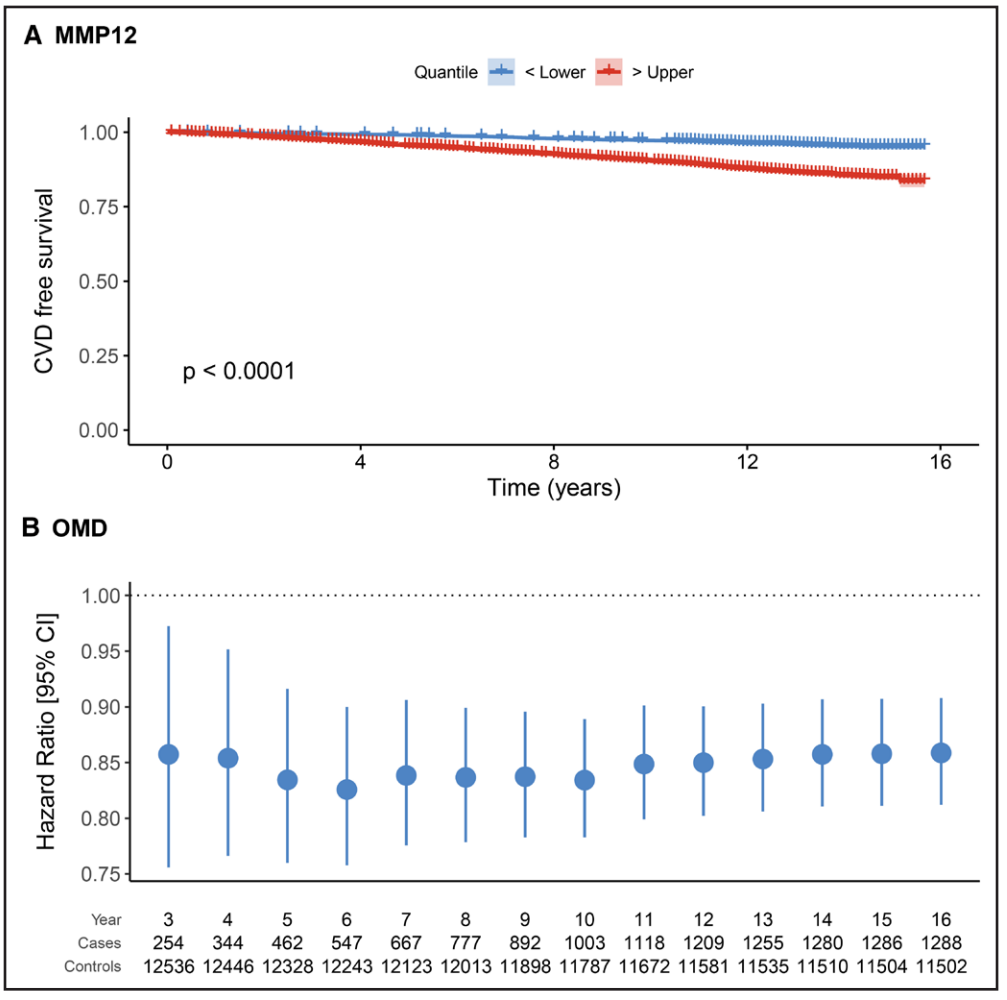


Figure 3. Changes in cardiovascular disease (CVD) free survival and CVD risk plotted for two selected protein EpiScores. **A**, Individuals with higher levels of MMP12 (>75th percentile) had shorter CVD-free survival when compared with those with lower levels of this EpiScore (<25th percentile). **B**, Hazard ratios (per SD of the EpiScore) and 95% CIs associated with the levels of OMD EpiScore plotted over time. At all examined time points, the association with CVD risk was significant (P<0.05).

models adjusted for ASSIGN. Although the score may be a useful addition to other omic features in future CVD risk prediction tools, at present, it is unlikely to be measured in a clinical setting.[23]

One previous study focused on how DNAm biomarkers improve CVD risk prediction.[24] 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[24] 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 are associated with CVD. Furthermore, we developed 10-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.[25] MMP12 and OMD, in turn, are involved in maintaining the stability of atherosclerotic plaques. While MMP12 contributes to the growth and destabilization of plaques,[26] increased levels of OMD have been observed in macrocalcified plaques from asymptomatic patients.[27] Finally, multiple studies have demonstrated that NOTCH1 signaling protects the heart from CVD-induced myocardial damage. The Notch1 pathway is involved in neoangiogenesis and revascularization of a failing heart.[28] It limits the extent of ischemic injury,[28] reduces fibrosis,[29] and improves cardiac function.[30] Several protein EpiScores associated with CVD in our study, such as SELE and C5, have also been shown to be associated with stroke and ischemic heart disease in our previous work.[20]

Table 2. C-Statistics Calculated for Null and Full Protein EpiScore Models. Risk Was Ascertained Over 16 Years of Follow-Up. The Null Model Was Adjusted for ASSIGN and the Concentration of Cardiac Troponin I, While the Full Model Also Included a Studied EpiScore

EpiScore	C_{full}–C_{null}	Function
IGFBP4	0.0050	Metabolic/growth promoter
CRP	0.0050	Immune response
NTRK3	0.0046	Neural development/cell signaling
FGF21	0.0042	Metabolic
CSF3	0.0039	Immune response
HGF	0.0035	Growth factor/tissue regeneration
ACVRL1	0.0035	Vascular
CNTN4	0.0035	Cell adhesion/maintenance
PIGR	0.0034	Immune response
RARRES2	0.0032	Metabolic

EpiScore indicates epigenetic score.

Whereas some of the EpiScores reflect known protein-CVD associations, others reflect novel pathways. This includes, but is not limited to, PRSS2 and CNTN4. PRSS2, which encodes the digestive enzyme trypsin 2, has been mainly studied in the context of pancreatitis. However, recent studies provide evidence that trypsin can leak from the small intestine into the bloodstream and digest myocardial tissue during heart failure.[31] Trypsin-mediated degradation of heart tissue was also observed in cases of dilated cardiomyopathy following influenza A infection.[32] CNTN4, in turn, encodes a cell adhesion molecule implicated in the development of autism spectrum disorders.[33] Recent studies have shown that mutations in CNTN4 were associated with an elevated production of a prothrombotic agent called thromboxane A2 and an increased risk of cardiovascular events.[34]

The protein EpiScore that we trained for cTnI was not associated with the incidence of CVD. Therefore, we excluded it from composite CVD score generation. This highlights an important consideration in the development of multiomics 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[20] that we make use of in our study were extracted as the best-performing EpiScores from a total set of 953 proteins tested as potential outcomes. It is, therefore, not always possible to generate a meaningful protein EpiScore that reflects the protein biology. In the case of cardiac troponins, the elevations in circulating cTnI and cTnT are a result of a leakage of these proteins from the damaged heart muscle into the bloodstream.[35] As opposed to transcription, this process is not regulated by DNAm. Therefore, the methylation signal underlying an increased concentration of cardiac troponin in the bloodstream may be too weak to enable the generation of a meaningful EpiScore. This limitation may also extend to other proteins derived in the heart or other tissues involved in CVD onset. Nonetheless, the ability of a DNAm array to capture surrogate markers for hundreds of proteins—many of which are not routinely measured in the clinic—offers promise in the development of CVD biomarkers.

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 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 generalizability beyond a Scottish population. In this study, we trained and tested predictors in a Scottish cohort to augment the ASSIGN score. However, many of the protein EpiScores were trained in a German cohort (KORA [Cooperative Health Research in the Region Augsburg]) and projected to GS.[20] This suggests that the EpiScore biomarkers part-translate across European ancestry populations. 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. To generalize the findings further, replication of the EpiScore associations with CVD (while adjusting for SCORE2) across other European ancestry populations is required.

CONCLUSIONS

In conclusion, we identified novel epigenetic signals that were associated with the incidence of CVD independently of ASSIGN and the concentration of cardiac troponin. The exploration of associations between protein EpiScores and CVD shed light on the etiology and molecular biology of the disease. 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.

ARTICLE INFORMATION

Received April 22, 2023; accepted November 30, 2023.

Affiliations

Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer (A.D.C., D.A.G., Y.C., E.B., A.C., D.L.M., K.L.E., R.E.M.), Division of Psychiatry, Royal Edinburgh Hospital (A.M.M.), Edinburgh Clinical Research Facility, Western General Hospital (N.W., L.M.), and Usher Institute, Old Medical School (J.F.P.), The University of Edinburgh, United Kingdom. School of Psychology, University of Exeter, United Kingdom (R.M.W.). Institute of Cardiovascular and Medical Sciences, British Heart Foundation Glasgow Cardiovascular Research Centre, University of Glasgow, United Kingdom (P.W., N.S.).

Acknowledgments

The authors 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, health care assistants, and nurses. They would also like to thank Dr Robert Hillary for helpful feedback on the article and analyses.

Sources of 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 license to any author-accepted manuscript version arising from this submission. A.D. Chybowska was supported by a Medical Research Council PhD Studentship in Precision Medicine with funding from the Medical Research Council Doctoral Training Program and the University of Edinburgh College of Medicine and Veterinary Medicine. D.A. Gadd was 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). Drs Bernabeu and Marioni were supported by the Alzheimer's Society major project grant AS-PG-19b-010. Generation Scotland (GS) 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 were performed by the Genetics Core Laboratory at the Clinical Research Facility, The University of Edinburgh, United Kingdom, and was funded by the Medical Research Council UK and the Wellcome Trust (references 104036/Z/14/Z, 220857/Z/20/Z, and 216767/Z/19/Z). The DNA methylation profiling and analysis were supported by Wellcome Investigator Award 220857/Z/20/Z and grant 104036/Z/14/Z (principal investigator: AM McIntosh) and through funding from National Alliance for Research on Schizophrenia & Depression (reference 27404; awardee: Dr D.M. Howard) and the Royal College of Physicians of Edinburgh (Sim Fellowship; awardee: Dr H.C. Whalley). All components of GS:SFHS (Generation Scotland: the Scottish Family Health Study), including the protocol and written study materials, have received formal and national ethical approval from the NHS Tayside Research Ethics Committee (Research Ethics Committee reference number 05/S1401/89). In addition, local approval has been obtained from the NHS Glasgow Research Ethics Committee and the NHS Glasgow and NHS Tayside Research and Development Offices, as required.

Disclosures

L. Murphy and Dr Marioni received a speaker fee from Illumina. Dr Marioni is an advisor to the Epigenetic Clock Development Foundation. D.A. Gadd and Dr Marioni received consultancy fees from Optima Partners. The other authors report no conflicts.

Supplemental Material

Supplemental Methods
Tables S1–S7
Figures S1–S3
References 36–46

REFERENCES

1. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*. 2017;357:j2099. doi: 10.1136/bmj.j2099
2. van Staa TP, Gulliford M, Ng ESW, Goldacre B, Smeeth L. Prediction of cardiovascular risk using Framingham, ASSIGN and QRISK2: how well do they predict individual rather than population risk? *PLoS One*. 2014;9:e106455. doi: 10.1371/journal.pone.0106455
3. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, Lassale CM, Siontis GCM, Chiocchia V, Roberts C, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;353:i2416. doi: 10.1136/bmj.i2416
4. Ho JE, Lyass A, Courchesne P, Chen G, Liu C, Yin X, Hwang S, Massaro JM, Larson MG, Levy D. Protein biomarkers of cardiovascular disease and mortality in the community. *J Am Heart Assoc*. 2018;7:e008108. doi: 10.1161/JAHA.117.008108
5. Hijazi Z, Lindbäck J, Alexander JH, Hanna M, Held C, Hylek EM, Lopes RD, Oldgren J, Siegbahn A, Stewart RAH, et al; ARISTOTLE and STABILITY Investigators. The ABC (age, biomarkers, clinical history) stroke risk score: a biomarker-based risk score for predicting stroke in atrial fibrillation. *Eur Heart J*. 2016;37:1582–1590. doi: 10.1093/eurheartj/ehw054
6. Wallentin L, Eriksson N, Olszowka M, Grammer TB, Hagström E, Held C, Kleber ME, Koenig W, März W, Stewart RAH, et al. Plasma proteins associated with cardiovascular death in patients with chronic coronary heart disease: a retrospective study. *PLoS Med*. 2021;18:e1003513. doi: 10.1371/journal.pmed.1003513
7. Gadd DA, Hillary RF, Kuncheva Z, Mangelis T, Admanit R, Gagnon J, Lin T, Ferber K, Runz H, Team BB, et al. Blood protein levels predict leading incident diseases and mortality in UK Biobank. 10.1101/2023.05.01.23288879.
8. Welsh P, Preiss D, Hayward C, Shah ASV, McAllister D, Briggs A, Boachie C, McConnachie A, Padmanabhan S, Welsh C, et al. Cardiac troponin T and troponin I in the general population. *Circulation*. 2019;139:2754–2764. doi: 10.1161/CIRCULATIONAHA.118.038529
9. Fridén V, Starnberg K, Muslimovic A, Ricksten S-E, Bjurman C, Forsgard N, Wickman A, Hammarsten O. Clearance of cardiac troponin T with and without kidney function. *Clin Biochem*. 2017;50:468–474. doi: 10.1016/j.clinbiochem.2017.02.007
10. Michielsen ECHJ, Wodzig WKWH, Van Dieijen-Visser MP. Cardiac troponin T release after prolonged strenuous exercise. *Sports Med*. 2008;38:425–435. doi: 10.2165/00007256-200838050-00005
11. Wu AHB. Release of cardiac troponin from healthy and damaged myocardium. *Front Lab Med*. 2017;1:144–150. doi: 10.1016/j.flm.2017.09.003
12. Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes Dev*. 2011;25:1010–1022. doi: 10.1101/gad.2037511
13. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet*. 2008;9:465–476. doi: 10.1038/nrg2341
14. Cedar H, Bergman Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet*. 2009;10:295–304. doi: 10.1038/nrg2540
15. Hop PJ, Luijk R, Daxinger L, van Iterson M, Dekkers KF, Jansen R, Heijmans BT, 't Hoen PAC, van Meurs J, Jansen R, et al. Genome-wide identification of genes regulating DNA methylation using genetic anchors for causal inference. *Genome Biol*. 2020;21:220.
16. 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. *J Appl Physiol*. 2000;89:1499–1504. doi: 10.1152/jappl.2000.89.4.1499
17. Koenig W, Sund M, Fröhlich M, Löwel H, Hutchinson WL, Pepys MB. 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. *Am J Epidemiol*. 2003;158:357–364. doi: 10.1093/aje/kwg135
18. Liu Y, Buil A, Collins BC, Gillet LCJ, Blum LC, Cheng LY, Vitek O, Mouritsen J, Lachance G, Spector TD, et al. Quantitative variability of 342 plasma proteins in a human twin population. *Mol Syst Biol*. 2015;11:786. doi: 10.15252/msb.20145728
19. Stevenson AJ, McCartney DL, Hillary RF, Campbell A, Morris SW, Bermingham ML, Walker RM, Evans KL, Boutin TS, Hayward C, et al. Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. *Clin Epigenetics*. 2020;12:113. doi: 10.1186/s13148-020-00903-8
20. Gadd DA, Hillary RF, McCartney DL, Zaghlool SB, Stevenson AJ, Cheng Y, Fawns-Ritchie C, Nangle C, Campbell A, Flaig R, et al. Epigenetic scores for the circulating proteome as tools for disease prediction. Lo YD, Ferrucci L, eds. *eLife*. 2022;11:e71802. doi: 10.7554/eLife.71802
21. Estimate the risk-ASSIGN Score. Accessed April 14, 2022. https://www.assign-score.com/estimate-the-risk/visitors/
22. Smith BH, Campbell H, Blackwood D, Connell J, Connor M, Deary IJ, Dominiczak AF, Fitzpatrick B, Ford I, Jackson C, et al. Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med Genet*. 2006;7:74. doi: 10.1186/1471-2350-7-74
23. Wagner W. How to translate DNA methylation biomarkers into clinical practice. *Front Cell Dev Biol*. 2022;10:854797. doi: 10.3389/fcell.2022.854797
24. Cappozzo A, McCrory C, Robinson O, Freni Sterrantino A, Sacerdote C, Krogh V, Panico S, Tumino R, Iacoviello L, Ricceri F, et al. A blood DNA methylation biomarker for predicting short-term risk of cardiovascular events. *Clin Epigenetics*. 2022;14:121. doi: 10.1186/s13148-022-01341-4
25. Sharif S, Van der Graaf Y, Cramer MJ, Kapelle LJ, de Borst GJ, Visseren FLJ, Westerink J, van Petersen R, Dinther BGF, Algra A, et al. Low-grade inflammation as a risk factor for cardiovascular events and all-cause mortality in patients with type 2 diabetes. *Cardiovasc Diabetol*. 2021;20:220.
26. Goncalves I, Bengtsson E, Colhoun HM, Shore AC, Palombo C, Natali A, Edsfeldt A, Dunér P, Fredrikson GN, Björkbacka H, et al; SUMMIT Consortium. Elevated plasma levels of MMP-12 are associated with atherosclerotic burden and symptomatic cardiovascular disease in subjects with type 2 diabetes. *Arterioscler Thromb Vasc Biol*. 2015;35:1723–1731. doi: 10.1161/ATVBAHA.115.305631
27. Gonçalves I, Oduor L, Matthes F, Rakem N, Meryn J, Skenteris NT, Aspberg A, Orho-Melander M, Nilsson J, Matic L, et al. Osteomodulin gene expression is associated with plaque calcification, stability, and fewer cardiovascular events in the CPIP cohort. *Stroke*. 2022;53:e79–e84. doi: 10.1161/STROKEAHA.121.037223
28. Li Y, Hiroi Y, Ngoy S, Okamoto R, Noma K, Wang CY, Wang H-W, Zhou Q, Radtke F, Liao R, et al. Notch1 in bone marrow–derived cells mediates

cardiac repair after myocardial infarction. *Circulation*. 2011;123:866–876. doi: 10.1161/CIRCULATIONAHA.110.947531
29. Ferrari R, Rizzo P. The Notch pathway: a novel target for myocardial remodelling therapy? *Eur Heart J*. 2014;35:2140–2145. doi: 10.1093/eurheartj/ehu244
30. Gude NA, Emmanuel G, Wu W, Cottage CT, Fischer K, Quijada P, Muraski JA, Alvarez R, Rubio M, Schaefer E, et al. Activation of notch-mediated protective signaling in the myocardium. *Circ Res*. 2008;102:1025–1035. doi: 10.1161/CIRCRESAHA.107.164749
31. Courelli V, Ahmad A, Ghassemian M, Pruitt C, Mills PJ, Schmid-Schönbein GW. Digestive enzyme activity and protein degradation in plasma of heart failure patients. *Cell Mol Bioeng*. 2021;14:583–596. doi: 10.1007/s12195-021-00693-w
32. Pan HY, Sun HM, Xue LJ, Pan M, Wang YP, Kido H, Zhu JH. Ectopic trypsin in the myocardium promotes dilated cardiomyopathy after influenza a virus infection. *Am J Physiol Heart Circ Physiol*. 2014;307:H922–H932. doi: 10.1152/ajpheart.00076.2014
33. Zhang SQ, Fleischer J, Al-Kateb H, Mito Y, Amarillo I, Shinawi M. Intragenic CNTN4 copy number variants associated with a spectrum of neurobehavioral phenotypes. *Eur J Med Genet*. 2020;63:103736. doi: 10.1016/j.ejmg.2019.103736
34. McCarthy N, Vangjeli C, Surendran P, Treumann A, Rooney C, Ho E, Sever P, Thom S, Hughes A, Munroe P, et al. Genetic variants in PPARGC1B and CNTN4 are associated with thromboxane A2 formation and with cardiovascular event free survival in the Anglo-Scandinavian Cardiac Outcomes Trial (ASCOT). *Atherosclerosis*. 2018;269:42–49. doi: 10.1016/j.atherosclerosis.2017.12.013
35. 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. doi: 10.2147/vhrm.s5306
36. Smith BH, Campbell A, Linksted P, Fitzpatrick B, Jackson C, Kerr SM, Deary IJ, MacIntyre DJ, Campbell H, McGilchrist M, 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. *Int J Epidemiol*. 2013;42:689–700. doi: 10.1093/ije/dys084
37. Welsh P, Preiss D, Shah ASV, McAllister D, Briggs A, Boachie C, McConnachie A, Hayward C, Padmanabhan S, Welsh C, et al. Comparison between high-sensitivity cardiac troponin T and cardiac troponin I in a large general population cohort. *Clin Chem*. 2018;64:1607–1616. doi: 10.1373/clinchem.2018.292086
38. Cheng Y, Gadd DA, Gieger C, Monterrubio-Gómez K, Zhang Y, Berta I, Stam MJ, Szlachetka N, Lobzaev E, Wrobel N, et al. Development and validation of DNA methylation scores in two European cohorts augment 10-year risk prediction of type 2 diabetes. *Nat Aging* 2023;3:450–458. doi: 10.1038/s43587-023-00391-4
39. Faqs-ASSIGN Score. Accessed April 14, 2022. https://www.assign-score.com/faqs/
40. Woodward M, Brindle P, Tunstall-Pedoe H., for the SIGN group on risk estimation*. 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. doi: 10.1136/hrt.2006.108167
41. SCORE2 Working Group and ESC Cardiovascular Risk Collaboration. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *Eur Heart J*. 2021;42:2439–2454. doi: 10.1093/eurheartj/ehab309
42. SCORE2 risk calculator for low-risk regions. Accessed October 6, 2022. https://heartscore.escardio.org/Calculate/quickcalculator.aspx?model=low
43. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. 1st ed. Springer New York; 2013. doi: 10.1007/978-1-4757-3294-8
44. Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Statist Soft*. 2010;33:1–22.
45. Lin DY. On the Breslow estimator. *Lifetime Data Anal*. 2007;13:471–480. doi: 10.1007/s10985-007-9048-y
46. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinf*. 2011;12:77. doi: 10.1186/1471-2105-12-77

6.3. Conclusion

The analysis of associations between individual protein EpiScores and CVD risk identified 67 potential biomarkers that were significantly associated with incident CVD beyond the ASSIGN score ($P<0.05$). After applying a conservative Bonferroni correction for multiple testing ($P<0.05/101 = 5.0\times10^{-4}$), 36 associations remained statistically significant. When circulating cTnI concentration was additionally included as a covariate, 65 protein EpiScores were associated with CVD risk at nominal significance ($P<0.05$), of which 33 retained significance after Bonferroni correction. The most significant associations were observed for EpiScores corresponding to proteins involved in metabolic regulation, immune response, and tissue development or regeneration pathways, reflecting biological processes central to CVD pathophysiology.

Building on these findings, a composite protein-based CVD EpiScore was derived using 45 individual EpiScores to capture the combined predictive signal. This composite score modestly improved 10-year CVD risk prediction beyond both ASSIGN and cTnI, corresponding to a 0.3% increase in the C-statistic. Comparable improvements were observed when SCORE2 was used as the primary clinical covariate. In both models, the composite CVD EpiScore remained a statistically significant predictor of incident CVD, indicating that epigenetic signatures of circulating proteins provide complementary prognostic information to established clinical and biochemical risk factors.

In summary, this work identified novel epigenetic signatures associated with incident CVD independently of clinical risk prediction tools (ASSIGN, SCORE2) and cardiac troponin. These findings underscore the potential of methylation-derived protein proxies to capture biological processes relevant to CVD. The subsequent chapter examines associations between directly quantified circulating proteins, measured using mass spectrometry, and CVD outcomes.

7. Proteomic Biomarkers of CVD and Death

7.1. Introduction

The levels of circulating proteins provide valuable insights into future risk of CVD. Nevertheless, the associations of many proteins with CVD remain poorly characterised. One key limitation is that commercially available proteomic panels capture only a subset of the circulating proteome, restricting the discovery of novel biomarkers. MS-based proteomics offers a powerful and unbiased means of addressing this gap by enabling comprehensive profiling of proteins without reliance on predefined targets. However, the high cost, technical demands, and relatively low throughput of untargeted approaches have limited their use in large population cohorts, meaning that the full potential of proteomic discovery for elucidating disease mechanisms and improving risk prediction has yet to be realised.

In this chapter, I applied a novel, cost-effective mass spectrometry approach to study the relationships between circulating protein abundances and CVD in 8,343 individuals from the GS cohort. Using data on 439 proteins, I first explored potential biomarkers of early disease, highlighting proteins whose circulating levels were associated with incident CVD and its sub-types. Next, I explored sex-specific effects of proteins on CVD risk and the associations of circulating proteins with individual CVD risk factors. Finally, I developed a proteomic risk score that integrated information from multiple protein concentrations into a single composite measure, with the aim of improving CVD risk prediction beyond established clinical risk factors.

This study has been submitted to medRxiv and can be accessed using the following link: https://www.medrxiv.org/content/10.1101/2025.11.20.25340673v2. Supplementary materials are available at medRxiv (https://www.medrxiv.org/content/10.1101/2025.11.20.25340673v2.supplementary-material) and at https://github.com/aleksandra-chybowska/thesis/tree/main/Untargeted_proteomic_profiling_identifies_candidate_biomarkers_for_early_detection_of_cardiovascular_disease_and_mortality. The source code can be accessed at: https://github.com/aleksandra-chybowska/MS_proteins_and_CVD.

7.2. Proteomic Biomarkers of CVD and Death

Untargeted Proteomic Profiling Identifies Candidate Biomarkers for Early Detection of Cardiovascular Disease and Mortality

Aleksandra D. Chybowska[1], Spyros Vernardis[2,3], Daniel L. McCartney[1], Jure Mur[1], Josephine Robertson[1], Hannah M. Smith[1], Archie Campbell[1], Camilla Drake[4], Hannah Grant[1], Poppy Adkin[2,5], Matthew White[2], Christoph B. Messner[2,6], Arturas Grauslys[3], Sergej Andrejev[3], Charles Brigden[3], David J. Porteous[1], Caroline Hayward[4], Jackie F. Price[7], Kathryn L. Evans[1], Aleksej Zelezniak[2,3,8,9,10], Markus Ralser[2,3,11], Riccardo E. Marioni[1]

[1.] Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, UK
[2.] The Francis Crick Institute, Molecular Biology of Metabolism Laboratory, London, UK
[3.] Eliptica Limited, The London Cancer Hub, Cotswold Road, Sutton, London, UK
[4.] MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK
[5.] MRC Clinical Trials Unit, Institute of Clinical Trials and Methodology, University College London, UK
[6.] Precision Proteomics Center, Swiss Institute of Allergy and Asthma Research, University of Zurich, Switzerland
[7.] Usher Institute, University of Edinburgh, Edinburgh, UK
[8.] Department of Life Sciences, Chalmers University of Technology, Gothenburg, Sweden
[9.] Randall Centre for Cell & Molecular Biophysics, King's College London, London, UK
[10.] Institute of Biotechnology, Life Sciences Centre, Vilnius University, Vilnius, Lithuania
[11.] Department of Biochemistry, Charité Universitätsmedizin Berlin, Berlin, Germany

*Corresponding author:
Name and Address: Riccardo Marioni, Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, UK
Contact Details: riccardo.marioni@ed.ac.uk

Abstract:

Background:
The serum proteome can provide valuable insights into the development and progression of diseases. This is particularly important for cardiovascular disease (CVD), a leading cause of death worldwide. In this large-scale cohort study, we employ an untargeted mass-spectrometry-based approach to explore associations between highly expressed proteins, incident CVD (analysed as six individual outcomes and one composite outcome) and all-cause mortality.

Methods:
The abundances of 439 proteins and protein groups quantified by mass spectrometry in serum were related to incident outcomes in 8,343 Generation Scotland participants (age 40–69 years), who were free of CVD at baseline ($n_{all\_cause\_death}$=618, $n_{composite\_CVD}$=666, follow-up ≤17 years). Cox proportional hazards (PH) models were run before and after adjustment for pre-selected known CVD risk factors. Sex-specific effects were explored. A protein-based risk score for composite CVD outcome was developed using penalised regression.

Results:
Forty-eight high abundance serum proteins and protein groups were significantly associated with incident CVD and death outcomes ($P_{Bonferroni}<1.14x10^{-4}$), including 24 associations not reported in the Open Targets database. Proteins involved in immune and oxidative stress responses were associated with composite CVD (Immunoglobulin heavy variable 3/OR16-9, Hazard Ratio per SD (HR)=0.85 [95%CI 0.79,0.92]) and death (Alpha-1-antitrypsin, HR=1.27 [1.17, 1.38]), while heart failure was linked to proteins playing a role in lipid metabolism (Apolipoprotein A-II, HR=0.70 [0.59, 0.84]) and complement cascade (Complement C1q subcomponent subunit B, HR=1.40 [1.18, 1.66]). Applied to the test set, the proteomic risk score improved 17-year incident CVD prediction over models including age, sex, and nine lifestyle and clinical risk factors (ΔAUC = 0.010, ROC P = 0.013).

Conclusion:
The highly abundant serum proteome, readily assessed by mass spectrometry, reveals candidate biomarkers for incident CVD and provides predictive value for early risk stratification.

1. Non-standard Abbreviations and Acronyms

ABC - Ammonium Bicarbonate
ACN - Acetonitrile
BMI - Body Mass Index
BNP - B-type natriuretic peptide
CRP - C-reactive Protein
CVD - Cardiovascular Disease
DTT - Dithiothreitol
FA - Formic Acid
FDR - False Discovery Rate
GS - Generation Scotland
HCU - Hormonal Contraception Use
HR - Hazard Ratio
IAA - Iodoacetamide
LC - Liquid Chromatography
MS - Mass Spectrometry
NT - N-terminal
PG - Protein Group
PH - Proportional Hazards
SBP - Systolic Blood Pressure
SD - Standard Deviation
SIMD - Scottish Index of Multiple Deprivation
VIF - Variance Inflation Factor

2. Introduction

CVD has been among the leading causes of morbidity and mortality worldwide for over 20 years [1]. Despite progress in diagnosis and treatment, there is an ongoing need to gain further insights into the underlying molecular mechanisms of CVD. This knowledge is essential for improving early diagnosis and developing more effective, targeted therapies.

Currently, there are a number of protein biomarkers known to be associated with acute or chronic CVD, including: 1) cardiac troponin T (cTnT) and I (cTnI) [2,3], 2) B-type natriuretic peptide (BNP) and its N-terminal form (NT-pro BNP) [4], 3) D-dimer and C-reactive protein (CRP) [5–7], and 4) Apolipoprotein A-I [8]. Elevated levels of cTnT and cTnI are strongly associated with acute coronary syndrome and myocardial infarction [2,3]. High-sensitivity troponin assays enhance the diagnostic sensitivity for acute coronary syndrome and can also be used for risk stratification. BNP and NT-proBNP are employed to diagnose congestive heart failure [4]. D-dimer [7] and C-reactive protein [5,6] are non-specific inflammatory markers that facilitate the diagnosis of thromboembolic conditions and assess cardiovascular risk, particularly in patients with no known cardiovascular disease. Lastly, apolipoprotein A-I is an excellent predictor of HDL metabolism-associated CVD [8]. While numerous studies have demonstrated the diagnostic and prognostic value of established biomarkers, a key limitation is that many of them reflect downstream or acute manifestations of CVD [9]. As such, they may be less effective at identifying individuals in preclinical stages of disease or capturing early molecular changes before clinical symptoms emerge. This highlights the need for novel biomarkers capable of detecting early, subclinical pathophysiological processes.

One method of identifying novel biomarkers is by relating circulating protein measures to incident CVD in large cohort studies [10,11]. Nonetheless, despite this promise, challenges remain. Firstly, CVD is an umbrella term for several inter-related clinical conditions, with atherosclerosis being the common underlying pathology. These conditions include ischaemic heart disease (e.g., angina, myocardial infarction) and cerebrovascular disease (e.g., ischaemic stroke, haemorrhagic stroke, transient ischemic attack). Each sub-condition has its own distinctive risk factor profile and therapeutic strategies, for instance, hypertension may be more critical for stroke [12], while cholesterol levels may be useful for coronary heart disease [13] though these show considerable overlap. Modelling CVD as a single, composite outcome increases the number of testable cases and can reveal shared biomarkers. However, the heterogeneity of diseases included in the composite outcome can obscure disease-specific associations, and may result in hypothetical biomarkers that are of low specificity.

There are multiple ways of measuring protein concentrations in accessible body fluids. Thus far, most proteomic studies that focused on identifying novel biomarkers of CVD have made use of affinity-reagent based technologies [14–17]. These targeted technologies can quantify proteins across a wide dynamic range, including many low-abundance components of the circulating proteome. However, because they are limited to a predefined set of proteins, they may miss novel or unanticipated biomarkers. In contrast, mass spectrometry (MS)–based proteomics can be performed in an untargeted fashion, enabling a more comprehensive survey of the proteome. Mass spectrometry is particularly well suited in the quantification of the high abundance serum protein fraction which is enriched in proteins that execute their function in metabolism and immune system, and has thus far proven as the main reservoir of protein biomarkers that eventually made it into clinical use. Experience towards the robust mass spectrometry-based profiling of big cohorts has been gained throughout the recent years [18].

Lastly, men and women differ in their CVD profiles. While young men are typically at a higher risk of developing CVD compared to women of the same age, the risk of CVD in women increases and often surpasses that of men after the menopause, largely due to the loss of the protective effects of oestrogen [19]. Oestrogen modulates the immune response, balancing pro-inflammatory and anti-inflammatory pathways [20,21]. By doing so, it helps prevent chronic low-grade inflammation - a key factor in atherosclerosis and other cardiovascular diseases [22]. Moreover, hormonal contraception use (HCU) has been shown to affect the high abundant proteome [23], which may influence the associations with CVD observed in females.

In this work, we use a cost-effective high-throughput MS platform [24] to discover candidate biomarkers for the early stages of CVD and gain insights into the molecular aetiology of the studied diseases. The analysis focuses on individuals of European ancestry from the Generation Scotland (GS) cohort. To capture both shared and condition-specific signals, we analysed CVD as a composite outcome - including ischaemic stroke, myocardial infarction, coronary heart disease, and CVD-related death – and as individual events (fatal or non-fatal coronary heart disease, myocardial infarction, heart failure, ischaemic stroke, transient ischaemic attack and CVD death). All-cause mortality was assessed as a separate outcome. Serum proteomic profiling was conducted in 15,818 GS participants, yielding quantitative data on 439 serum proteins and protein groups. Then, we employed Cox regression to link concentration changes in these proteins to CVD and death outcomes. Next, we analysed sex-specific effects of proteins on CVD risk and the influence of established lifestyle and health risk factors on these associations. Finally, we derived a proteomic risk score for the composite CVD outcome and evaluated its utility for 17-year risk prediction.

Methods

2.1. Generation Scotland

Generation Scotland (GS) is a family-based cohort of individuals aged 18 to 98 from 7,000 family groups spread throughout Scotland [25]. Recruitment of study participants occurred between 2006 and 2011. Eligible individuals were invited to attend a clinic, where their clinical and physical characteristics were measured following a standardized protocol. In total, 24,088 participants completed a health questionnaire. Fasting blood samples were collected from 21,521 individuals using a standard operating procedure. All participants provided written informed consent for research. The study received ethical approval from the National Health Service Tayside Committee on Medical Research Ethics (REC reference number: 05/S1401/89). The GS dataset is not publicly available as it contains information that could compromise participant consent and confidentiality. However, the data, research materials, and analytical methods will be made accessible to other researchers for the purpose of replicating the findings. Access will be granted upon successful project application to the GS Access Committee and obtaining ethical approval for accessing linked health data from NHS Scotland. Instructions for accessing GS data can be found at https://www.ed.ac.uk/generation-scotland/for-researchers/access; the GS Access Request Form can be downloaded from this site. All code used in the analyses is available with open access at the following Github repository: https://github.com/aleksandra-chybowska/MS_proteins_and_CVD.

2.2. Materials

For the mass spectrometric proteome analysis study, the following chemicals and items were used: acetonitrile (ACN, Thermo Fisher, 10001334), ammonium bicarbonate (ABC, Thermo Fisher, 15645440), C18 96-well plates (BioPureSPN Macro 96-well, 100 mg PROTO 300 C18, HNS S18V-L), dithiothreitol (DTT, Sigma-Aldrich, 43815), formic acid (FA, Thermo Fisher Scientific, 85178), iodoacetamide (IAA, Sigma-Aldrich, I1149), methanol (MeOH, Thermo Fisher, 10767665), trypsin (Promega, Sequence Grade, V5117), urea (Sigma-Aldrich, 1084870500) and water (Thermo Fisher, 10505904).

2.3. Sample Preparation for Serum Proteomics

The sample preparation process has been described previously [24,26]. Briefly, 5 µL of serum samples were added to a solution of 50 µL of 8 M urea and 0.1 M ammonium bicarbonate at pH 8.0 to denature the proteins. Subsequently, the proteins were reduced using 5 µL of 50 mM dithiothreitol for 1 hour at 30 °C and alkylated with 5 µL of 100 mM iodoacetamide for 30 minutes in the dark. The sample was then diluted with 340 µL of 0.1 M ammonium bicarbonate to a concentration of 1.5 M urea. The next step was the trypsinisation of the proteins and 200 µL of the solution was used, and the proteins were digested overnight with trypsin (12.5 µL, 0.1 µg/µL) at 37 °C at a 1/40 trypsin/total protein ratio. The digestion was quenched with the addition of 25 µL of 0.1% v/v FA. The peptides were cleaned up with C18 96-well plates, eluted with 50% v/v ACN, dried by a vacuum concentrator (Eppendorf Concentrator Plus), and redissolved in 50 µL of 10% v/v FA for processing by liquid chromatography (LC)-MS.

2.4. Liquid Chromatography and Mass Spectrometry Acquisition

An Agilent 1290 Infinity II system (Agilent Technologies) was used for liquid chromatography, and it was connected to a TripleTOF 6600 mass spectrometer (SCIEX). 2 µg of total peptides were injected and separated using a Luna Omega LC Column (1.6u PS C18 100A, 30 x 2.1 mm (Phenomenex)) over a 3-minute gradient. A linear gradient was utilised, starting from 1% B and reaching 40% B over 3 minutes (Buffer A: 0.1% v/v FA; Buffer B: ACN/0.1% v/v FA) with an 800 µL/min flow rate. When the gradient reached 40%, the washing and re-equilibration steps were carried out in the following manner: B was increased from 40% to 80% over 0.5 min, followed by 80% B for 0.2 min, and then decreased from 80% to 3% B over 0.1 min. Re-equilibration at 3% B was maintained for 1 min until the next injection.

For the sample acquisition, we used a Scanning SWATH method [27]. The mass spectrometer was operated at high-sensitivity mode. The source conditions were as follows: source gas one at 15 psi, source gas two at 20 psi, curtain gas at 25 psi, temperature at 0 °C, IonSpray floating voltage at 5,500 V, and declustering potential at 80 V. Rolling collision energies were determined using the following equation: CE=0.034×m/z+2, where m/z represents the centre of the scanning quadrupole bin. The Scanning SWATH parameters were as follows: total cycle time 0.69 s, transmission window width 10 Da, Sample duration 3.103 min, precursor range 450-850 Da, fragment range 100-1500 Da and effective accumulation time 16.90 ms.

2.5. Mass Spectrometry Data Processing, Batch Correction and Quality Control

Raw data was analysed by DIA-NN [28] as described previously [24]. DIA-NN (version 1.8.12) was run in Robust LC (high precision) quantification mode, using the default parameter set. Identification was performed using a previously generated spectral library [29], which was refined using DIA-NN, as described before [24]. During all steps, precursor false discovery rate (FDR) filtering was set to 1%. **Supplemental Figures 1** and **2** show the distribution of identified precursors.

Postprocessing was carried out in R v4.3.1 [30]. The data was filtered at 1% gene group q-value and only samples with minimum precursor identifications of 2000 were included in the analysis. The precursors were filtered to have at least 80% prevalence in QC samples. Within-batch signal drift correction was performed by fitting a linear model to the repeat injections of pooled QC samples and using the model to correct the remaining samples (adapted from [31]). The between-batch correction was done using the “limma” v3.54.2 [32] linear batch-correction algorithm. Prior to running subsequent analyses, all proteins were annotated (**Supplemental Table 1, Supplemental Methods**) and their abundances were rank-based inverse normalised (n=15,818).

2.6. Selection of Covariates for Survival Analysis

Covariates of our Cox PH models included known CVD risk factors, such as age, sex, average systolic blood pressure, total cholesterol, HDL cholesterol, smoking status, presence of rheumatoid arthritis, diabetes status, years of education, and socioeconomic deprivation (measured by the Scottish Index of Multiple Deprivation, SIMD). The assessment of these variables is explained in detail in the **Supplemental Methods** section. In line with the approach implemented by the authors of SCORE2 clinical risk predictor [33], all individuals aged less than 40 years and more than 69 years were excluded from the analysis. Average systolic blood pressure, log transformed pack years of smoking, HDL cholesterol and total cholesterol levels were trimmed of outliers (points beyond 4 standard deviations (SDs) of the mean). Body mass index (BMI) was log transformed and filtered to values between 18 and 50 kg/m^2. Participants with missing covariate data were excluded from the dataset, leaving a sample size of n=8,343 (**Supplemental Figure 3**).

Given that recent studies suggest a link between the circulating proteome and HCU in females[23,34], we examined this relationship using linear regression (**Supplemental Methods**). As significant associations between the studied proteins and HCU were detected (**Supplemental Table 2**), we decided to further adjust our models for contraception use.

To assess multicollinearity among the covariates, we generated a Spearman correlation matrix (**Supplemental Figure 4**) and calculated the variance inflation factor (VIF) (**Supplemental Table 3**). As all variable pairs had Spearman's $r < 0.6$ and the VIF values for all variables were less than 2, we did not exclude any of the covariates from the analysis due to multicollinearity.

2.7. Outcome Definitions for Survival Analysis

CVD (hospitalised cases only) and all-cause mortality data were identified through linkage to NHS data and death records. Incident cases were ascertained over a follow up period of up to 17 years. The outcomes of interest were divided into three categories: a) individual CVD events, which included fatal or non-fatal coronary heart disease, myocardial infarction, heart failure, ischaemic stroke, transient ischaemic attack, and CVD death, b) a composite CVD outcome, which included diseases considered in a previous GS publication by Welsh *et al*. [35] (coronary heart disease, ischaemic stroke, myocardial infarction and CVD death), and c) all-cause death. All disease outcomes were defined as per CALIBER/HDR UK [36] consensus definitions. **Supplemental Table 4** details International Classification of Diseases, 10th Revision (ICD10) codes included in each sub-category of outcomes. Only the first occurrences of events were considered per sub-category. More information about intersections of events within GS can be found in **Supplemental Figure 5**. CVD death was defined in line with Welsh *et al*. [35] (**Supplemental Table 4**).

2.8. Survival Analysis

The study employed Cox PH regression implemented in Python [37] version 3.10.13 using lifelines library [38] version 0.27.8 to investigate the relationship between individual protein abundances and the studied outcomes. CVD cases included individuals diagnosed after baseline who subsequently died, as well as those who remained alive after diagnosis. Prevalent cases were excluded from the analysis. Controls were censored at the end of the follow-up period (August 2023, up to 17 years of follow up) or at the time of death.

For each protein, two types of models were considered: a) basic models adjusted for age and sex, and b) fully-adjusted models including age, sex, average systolic blood pressure, total cholesterol, HDL cholesterol, smoking (pack years), rheumatoid arthritis, diabetes, years of education, SIMD score, and contraceptive use. Protein abundance was included in all models as the primary independent variable.

Sex-specific effects of proteins on the outcome risk were studied by incorporating a sex by protein abundance interaction term (protein*sex) to fully adjusted models (b) and plotted using "plot_partial_effects_on_outcome" function from lifelines library [38]. This function uses previously fitted Cox model to visualise the effect of varying one or more covariates on the predicted survival probability.

Proteins significantly associated with the studied outcomes in fully-adjusted models (b) were further analysed in R version 4.3.1 [30]. The linearity assumption was assessed by examining deviance residuals using ggcoxdiagnostics function from the survminer package [39] version 0.4.9. Full models were re-run using coxme library [40] version 2.2-20 with a kinship matrix fitted as a random effect to adjust for relatedness.

All models were adjusted for multiple testing using a Bonferroni correction applied in a disease-specific manner ($P_{Bonferroni}$=0.05/439). Given its conservative nature, we compared the findings after applying an FDR-correction (P_{FDR}<0.05).

2.9. Statistical Attenuation After Covariate Adjustment

An additional series of Cox PH regressions were run to model time to a composite CVD outcome for proteins that were significant in the basic model but not in the fully adjusted model. The analysis aimed to identify proteins that initially associated with CVD outcomes but failed to meet the threshold for statistical significance after adjusting for additional variables. The basic model was adjusted for age, sex, and individual protein abundance. Subsequently, additional covariates - average blood pressure, total cholesterol, HDL cholesterol, smoking, rheumatoid arthritis, diabetes, years of education, SIMD score, and contraceptive use - were independently incorporated into the basic model. The impact of the covariates was assessed by comparing the protein's significance to that from the basic model.

Using a similar approach, associations from the basic model were compared with those from a model additionally adjusted for baseline use of cardiovascular drugs, to assess potential confounding by treatment. Details of medication use assessment are provided in the **Supplemental Methods.**

2.10. Proteomic CVD Risk Score

A proteomic risk score for a composite CVD outcome was developed using elastic net regression. The proteomic dataset with incident CVD data collected over 17-year follow-up period (n=8343) was split into training and test sets. Individuals unrelated to each other and to all individuals in the training set ("singletons") were selected based on family ID variable and assigned to the test set (n=2660 individuals, 229 events). The dataset containing remaining participants was used for training (n=5683 individuals, 437 events). Protein abundances were rank-based inverse normal transformed and scaled (mean=0, SD=1) separately within training and test datasets.

In the training set, deviance residuals from an age- and sex-adjusted Cox model of composite CVD were used as the outcome. Linear elastic net regression (glmnet v4.1) [41] was fitted with 10-fold cross-validation stratified by family ID to prevent related individuals from splitting across folds. Proteins with non-zero coefficients were retained, and their coefficients were used to compute a weighted linear combination of protein abundances (the proteomic score). In the test set, scores were calculated using the selected coefficients. Four nested Cox PH models were evaluated for each outcome: (a) minimally adjusted (age, sex), (b) model (a) + proteomic score, (c) extended model including age, sex, and nine lifestyle and clinical risk factors (average systolic blood pressure, total cholesterol, HDL cholesterol, smoking pack years, rheumatoid arthritis status, diabetes status, educational attainment, SIMD, contraception use), and (d) model (c) + proteomic score. Model discrimination was assessed over 17 years of follow-up using C-index and area under the Receiver operating curve (AUC). Improvement from adding the proteomic score was quantified by comparing Receiver operating (ROC) curves using pROC [42] R package.

3. Results

3.1. Baseline Characteristics

Detectable abundances of 133 (30.3%) proteins and 306 (69.7%) 'protein groups' were found in 15,818 GS individuals. Protein groups are defined as mass spectrometry results that can be assigned to multiple proteins due to shared amino-acid sequences. Among identified protein groups, 199 (65.0%) originated from the same gene. Here, we refer to them as gene-derived protein groups (e.g., haptoglobin (G1) and haptoglobin (G2)). Of these, 61 (30.7%) protein groups occurred only once in the protein group subset of the dataset, such as selenoprotein P (G). Remaining protein groups (n=107, 35.0%) are labelled with a symbol "(PG)".

In our analysis sample of 8,343 participants (**Table 1**, detailed characteristics by outcome are provided in **Supplemental Table** 5), individuals who developed composite CVD (n=666) were more likely to be male and generally exhibited a more adverse risk factor profile. This included older age, elevated body mass index, lower HDL cholesterol and higher average systolic blood pressure. They also smoked more, spent less time in full-time education, were more likely to take cardiovascular medications, and had a higher prevalence of baseline diabetes or rheumatoid arthritis.

Table 1. Baseline Demographic Characteristics of the Study Cohort. *P* values correspond to independent sample *t* tests for continuous variables with normal distribution (reported as average ± SD), Mann–Whitney *U* test for continuous variables with skewed distribution (reported as median [quartile 1, quartile 3], or χ^2 test for categorical variables (reported as n [%]). BMI indicates body mass index; CVD, cardiovascular disease; HDL, high-density lipoprotein; SBP, systolic blood pressure; and SIMD, Scottish Index of Multiple Deprivation. BMI and smoking were log-transformed prior to running the statistical comparisons.

Characteristic	*No Composite CVD (n = 7677)*	*Composite CVD (n = 666)*	*P Value*
Age, years	53.7 (7.5)	58.1 (6.7)	<0.001
Male sex, n (%)	3041 (39.6)	428 (64.3)	<0.001
BMI, kg/m^2	27.15 (4.83)	28.60 (5.31)	<0.001
SIMD, rank	4593 [2708, 5554]	4178 [2213, 5483]	0.002
Diabetes, n (%)	222 (2.9)	68 (10.2)	<0.001
Rheumatoid arthritis, n (%)	142 (1.8)	18 (2.7)	0.164
Medication use, n (%)	1483 (19.3)	278 (41.7)	<0.001
Smoking, pack years	0 [0, 13]	5 [0, 34]	<0.001
SBP, mm Hg	133.8 (17.3)	140.6 (18.4)	<0.001
Total cholesterol, mmol/L	5.4 (1.0)	5.3 (1.2)	0.171
*Non-HDL cholesterol, mmol/L **	3.9 (1.0)	4.0 (1.1)	0.020
HDL cholesterol, mmol/L	1.4 [1.2, 1.7]	1.3 [1.1, 1.6]	<0.001
*Full-time education**	4 [3, 6]	4 [3, 5]	<0.001

* The medications variable indicates treatment with cardiovascular medications. Non-HDL cholesterol was calculated as total cholesterol minus HDL cholesterol concentration. Full-time education is a categorical variable, reported here as the median [quartile 1, quartile 3] for data privacy reasons. The categories are defined as follows: 0 (0 years), 1 (1–4 years), 2 (5–9 years), 3 (10–11 years), 4 (12–13 years), 5 (14–15 years), 6 (16–17 years), 7 (18–19 years), 8 (20–21 years), 9 (22–23 years), and 10 (24 or more years).

3.2. Protein Abundances are Associated with CVD Outcomes and Death

We tested whether abundances of 439 proteins and protein groups were associated with risk of cardiovascular outcomes and death over 17 years of follow-up. Risk was modelled using Cox PH regression, with one protein considered per model (**Table 2, Supplemental Tables 6** and **7**).

In the basic models, which were adjusted for age, sex, and the abundance of the specific protein, 243 proteins and protein groups showed significant associations with the health outcomes after Bonferroni correction for multiple testing ($P_{Bonferroni}<1.1x10^{-4}$). Out of these, 29 did not satisfy the PH assumption ($P_{local\_test}<0.05$) and were excluded from further analysis. 116 proteins were linked to multiple outcomes.

In the full models, which were further adjusted for pre-selected cardiovascular risk factors and the use of hormonal contraception (as detailed in the **Methods** section), the number of significant associations with health outcomes decreased to 48 ($P_{Bonferroni}<1.1x10^{-4}$). Among these, 41 met the PH assumption, with seven proteins linked to multiple outcomes. Twenty-four of the 41 associations between proteins and events had not been previously reported in the Open Targets database [43], which integrates a wide range of data sources on target-disease relationships. Adjusting for kinship did not affect the findings. Further details on previously unreported associations and the impact of adjusting for relatedness are provided in **Supplemental Table 8**.

Table 2. The Number of Proteins Associated with CVD and Death. Mean time to event is reported together with its SD. The basic models were adjusted for age, sex, and the abundance of the studied protein. The fully-adjusted models included these factors as well as additional covariates: average blood pressure, total cholesterol, HDL cholesterol, smoking status, rheumatoid arthritis, diabetes, years of education, SIMD score, and contraceptive use. Significant P values were corrected for multiple testing via the Benjamini-Hochberg method ($P_{FDR}<0.05$) and separately using the Bonferroni method ($P_{Bonferroni}<1.14\times10^{-4}$).

Event	Number of events	Time to Event (years)	Significant proteins (n)			
			Basic (P<0.05)	Full (P<0.05)	Full (P_{FDR})	Full ($P_{Bonferroni}$)
Composite CVD	666	7.6 (4.1)	194	109	42	9
CVD Death	245	9.6 (4.1)	164	109	46	4
All Cause Death	618	9.2 (4.1)	179	122	70	28
Coronary Heart Disease	329	7.2 (3.8)	164	56	0	0
Myocardial Infarction	233	7.1 (3.9)	140	48	0	0
Heart Failure	140	7.8 (3.9)	126	81	27	7
Ischaemic Stroke	99	8.3 (3.6)	53	27	0	0
Transient Ischaemic Attack	85	7.6 (3.7)	67	34	0	0

Figure 1 illustrates Bonferroni-significant associations after adjusting for risk factors in the full models. Multiple patterns were observed for these findings.

Firstly, increased abundances of haptoglobin isoforms (protein groups G1 and G3, HR range 1.17-1.18, $P<1.1x10^{-4}$) were associated with an increased hazard (HR>1) of composite CVD. In contrast, isoforms of selenoprotein P (G) were associated with a reduced hazard (HR<1) of CVD (HR=0.84 [0.78; 0.91], $P=9.1x10^{-6}$) and all-cause mortality (HR=0.83 [0.77, 0.90], $P=9.7x10^{-6}$). The association between Immunoglobulin heavy variable 3/OR16-9 (G) and both CVD (HR=0.85 [0.79, 0.92], $P=4.1\times10^{-5}$) and CVD-related death (HR=0.77 [0.68, 0.87], $P=2.3\times10^{-5}$) was not reported in the Open Targets database.

Secondly, Lumican (HR=1.41 [1.18,1.67], $P=1.1x10^{-4}$) was associated with an increased hazard of heart failure, while Angiotensinogen (HR=0.70 [0.59, 0.83], $P=4.3x10^{-5}$) was linked to a reduced hazard. The association between Angiotensinogen and heart failure was not driven by confounding due to HCU ($P_{HCU} > 0.05$). Previously unreported protein associations with heart failure risk included Apolipoprotein A-II (G2) (HR=0.70 [0.59, 0.84], $P=7.5x10^{-5}$), Corticosteroid-binding globulin (HR=0.67 [0.57, 0.80], $P=5.2x10^{-6}$) and Complement C1q subcomponent subunit B (G1) (HR=1.40 [1.18, 1.66], $P=1.1x10^{-4}$).

The three most statistically significant associations with an increased risk of all-cause mortality were unreported in Open Targets and included Alpha-1-antitrypsin (G2) (HR=1.27 [1.17, 1.38], $P=1.2x10^{-8}$), Lipopolysaccharide-binding protein (HR=1.24 [1.14, 1.35], $P=2.8x10^{-7}$) and Leucine-rich alpha-2-glycoprotein (HR=1.24 [1.14, 1.34], $P=4.7x10^{-7}$). Among the three proteins most significantly associated with a reduced risk of all-cause mortality, two were unreported - PG 35 (IGHG2, IGHG4) (HR=0.81 [0.74, 0.87], $P=2.2\times10^{-7}$) and Immunoglobulin heavy constant gamma 2 (G) (HR=0.82 [0.76, 0.89], $P=2.3\times10^{-6}$) - while one, Prothrombin (HR=0.81 [0.75, 0.88], $P=6.3\times10^{-7}$), was a previously known association.

Finally, while associations with coronary heart disease, myocardial infarction, ischaemic stroke, and transient ischaemic attack did not remain significant after correcting for multiple testing, proteins linked with: 1) coronary heart disease and myocardial infarction, 2) ischemic stroke and transient ischemic attack, as well as 3) CVD death and all-cause death exhibited similar effect directions and magnitudes across basic and fully-adjusted models (**Supplemental Figures 6, 7 and 8**).

Figure 1. Association Between Mass Spectrometry Protein Abundances, Death, and Cardiovascular Outcomes. The results are based on fully-adjusted models. Only the associations that met the Cox Proportional Hazards assumption are shown. Proteins and protein groups significant after applying Bonferroni correction are marked with circles, while non-significant associations are marked with triangles. Error bars represent 95% confidence intervals.

3.3. Sex-specific Effects

We also investigated sex-specific effects of proteins on the risk of the studied health outcomes. As before, we ran 439 Cox PH regressions, but now incorporating an interaction term (protein*sex). This analysis revealed a protein group with significant sex-specific effects on the risk of heart failure ($P_{Bonferroni}<1.1x10^{-4}$). A protein group containing complement factors B and C2 ($HR_{protein*sex}=2.04$ [1.45, 2.87], $P=4.1x10^{-5}$) was associated with a decreased risk of heart failure in females, but increased the risk in males (**Figure 2**).

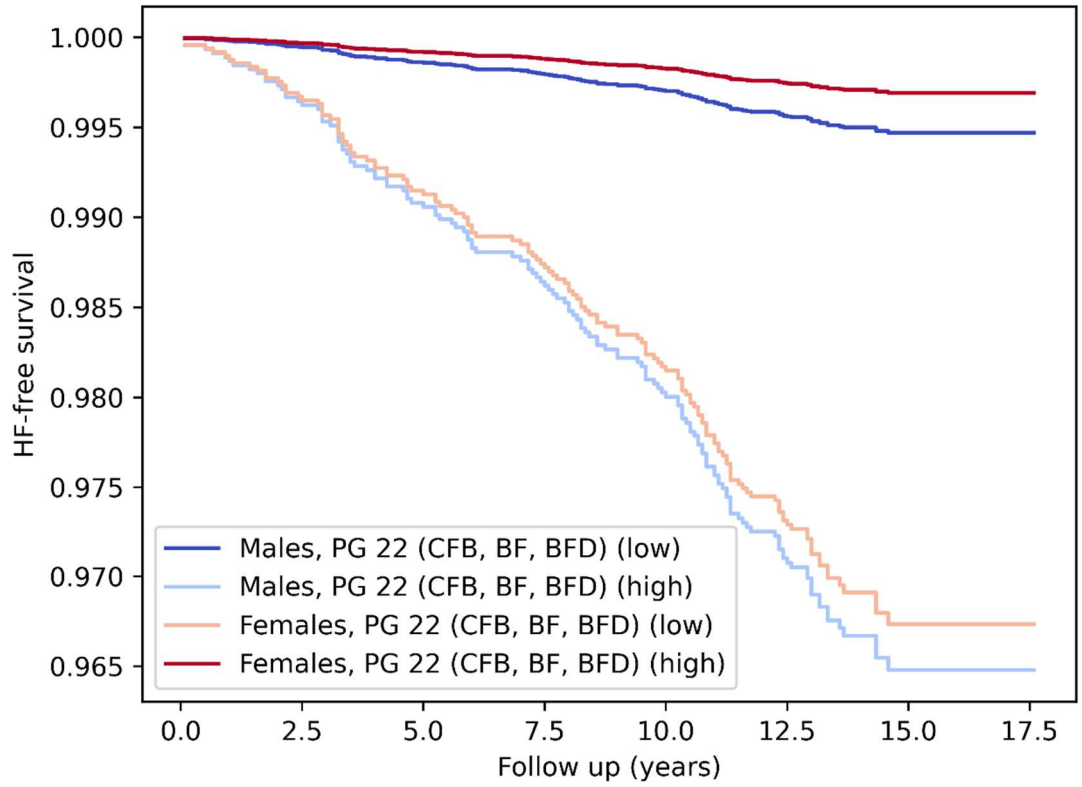


Figure 2. Predicted Sex-specific Effects of PG 22 (CFB, BF, BFD) on Heart Failure Risk. This protein group (containing complement factors B and C2) was associated with a decreased risk of heart failure in females, but with an increased risk in males. The high and low categories represent 3 SDs above and below the mean abundance of PG 22, respectively.

3.4. Risk Factors Driving Associations with Incident CVD Outcome

55 protein-incident CVD associations were statistically significant at a Bonferroni threshold ($P<1.1x10^{-4}$) in the basic model but not the fully-adjusted model. To determine which variables were potentially attenuating the associations (in terms of statistical significance), the basic model was augmented with each covariate in isolation (more details, including applied formulas, are available in **Supplemental Table 9**). The following number of associations with proteins were attenuated after the addition of: HDL cholesterol (24), diabetes status (13), pack years (11), years of education (9), SIMD score (8), average systolic blood pressure (5), rheumatoid arthritis (1). Of the 55 proteins, 17 have not previously been reported in the Open Targets database [43] in relation to CVD. Previously unreported biomarkers (n=9) included proteins such as Alpha-2-antiplasmin (G1), Immunoglobulin heavy variable 3-7 and PG 107 (VTN) (**Supplemental Table 9**).

Two proteins (Apolipoprotein B-100 and Apolipoprotein B-100 (G)) were not significant in the basic model of composite CVD but become significant at $P_{FDR}<0.05$ after adjusting the basic model for cardiovascular medication use. A full list of associations reaching $P<0.05$ in the age-, sex-, and medication-adjusted model is provided in **Supplemental Table 10**.

3.5. Proteomic CVD Risk Score

To assess whether mass spectrometry-derived protein abundances add predictive value beyond established CVD risk factors, we developed a proteomic CVD risk score. An elastic net regression model was trained on 5683 individuals with 437 events, selecting 50 of 439 measured proteins (coefficients in **Supplemental Table 11**).

The score was tested in a hold-out set of 2660 individuals (n=229 events) using four 17-year Cox proportional hazards models of increasing complexity. The hold-out set were both unrelated to each other and to all individuals in the training set, based on the family ID variable in GS. The minimally adjusted model (age and sex) achieved an AUC of 0.689 (C-index=0.685). Adding the proteomic score improved performance to AUC=0.729 (C-index=0.725), with the score showing independent predictive value (HR=1.47 [1.31-1.65], $P=4.1x10^{-11}$). The extended model with age, sex, and the nine lifestyle and clinical risk factors reached an AUC of 0.741 (C-index=0.737), while further inclusion of the proteomic score yielded the best discrimination (AUC=0.751, C-index=0.745). The proteomic score remained a significant predictor beyond established risk factors (HR=1.25 [1.09-1.44], P=0.002). Differences in ROC curves were significant for both the minimally adjusted model with vs. without the proteomic score (ΔAUC=0.040, ΔROC $P=4.4x10^{-5}$) and the extended model with vs. without the score (ΔAUC = 0.010, ROC P = 0.013).

4. Discussion

We conducted a large-scale, unbiased study using untargeted mass spectrometry to analyse the circulating proteome and its association with CVD and mortality. This approach focused on high-abundance proteins, which are critical to metabolism and immune function and have historically yielded successful biomarkers. We identified 48 proteins and protein groups that improve CVD and mortality risk prediction beyond traditional factors, with associations detectable up to 17 years before the event. Of these, 24 protein-disease links were not present in Open Targets database. Additionally, we explored sex-specific effects of the complement system in heart failure and developed a proteomic risk score that improved the prediction of incident CVD beyond established risk factors. These findings could inform the development of novel biomarkers for early CVD detection and personalized prevention strategies.

Our results align with the immuno-inflammatory theory of CVD, which suggests that inflammation and immune system dysregulation play a central role in the development and progression of CVD. Proteins associated with an increased hazard of the composite CVD outcome included haptoglobin, an acute-phase reactant whose concentrations fluctuate significantly during inflammation. [44]. Haptoglobin is also responsible for the clearance of toxic free haemoglobin released during red blood cell breakdown. By binding free haemoglobin, it mitigates tissues damage and oxidative stress, contributing to cardiovascular health [44]. Similarly, a selenium transporter called Selenoprotein P works to support the immune system and reduce oxidative stress [45]. It associated with a decreased risk of CVD. Additionally, isoforms of Immunoglobulin heavy variable 3/OR16-9 were associated with a decreased risk of CVD and were unreported in the Open Targets database.

Inflammation-related proteins were also identified in associations with all-cause and CVD death. While alpha-1-antitrypsin [46] modulates the immune response and protects tissues from inflammatory damage, Lipopolysaccharide-binding protein [47], and Leucine-rich alpha-2-glycoprotein [48] are acute phase reactants expressed in ongoing inflammation. In contrast, isoforms of immunoglobulin heavy constant gamma 2 are subclasses of IgG antibodies, which are critical components of the adaptive immune system. The inverse association between their abundance and all-cause mortality has not been reported before.

Proteins associated with heart failure seemed to track the progression of CVD. Apolipoprotein A-II is the second most abundant apolipoproteins of HDL. Although its role in HDL function and metabolism has been debated [49], epidemiological studies showed that it is inversely associated with risk of future coronary artery disease [50], possibly through its involvement in reverse cholesterol transport and removal from arteries [51]. Component C1q has been implicated in the initiation and progression of atherosclerotic plaques and promoting plaque instability [52], whereas Lumican [53] is involved cardiac remodelling [54].

The negative association of angiotensinogen with heart failure was unexpected. It did not reflect confounding by hormonal contraception [23]. Angiotensinogen is a part of the renin-angiotensin-aldosterone system, which regulates blood pressure [55]. Briefly, when the pressure drops, renin is released. Renin converts angiotensinogen to angiotensinogen I, which is then converted to angiotensin II. Angiotensin II stimulates aldosterone secretion in the adrenal glands, leading to salt and water reabsorption by the kidneys and constriction of small arteries, thereby increasing blood pressure. The protective effect of angiotensinogen may represent an

adaptive response to the heart's diminished capacity to meet the body's blood flow demands during heart failure [56]. This adaptation involves the upregulation of angiotensin-converting enzyme 2 (ACE2), which converts angiotensin II into angiotensin 1–7 [56]. Angiotensin 1–7 exerts vasodilatory effects, counterbalancing the actions of angiotensin II [57]. The upregulation of the ACE2/angiotensin 1–7 axis in heart failure serves as a compensatory response to mitigate the detrimental effects of an overactive renin-angiotensin system, thereby preserving cardiac function [57].

Our results suggest that the differential risk of heart failure between males and females is accounted for by protein groups associated with complement system. Specifically, complement factors B and C2 were associated with a decreased risk in females but an increased risk in males. Further research is warranted to explore the underlying mechanisms driving these sex-specific associations.

Whilst conducting Mendelian randomisation analyses was beyond the scope of this study, we have addressed this question as part of a broader genome-wide association study of the mass spectrometry–derived proteins [58]. That work identified three proteins - Apolipoprotein B, Apolipoprotein E, and Apolipoprotein - as causally linked to coronary artery disease. In our current analyses, Apolipoprotein and Apolipoprotein E showed significant associations at $P_{FDR}<0.05$ in the basic models of composite CVD and CVD death, respectively, but these associations attenuated after accounting for established risk factors, highlighting the influence of conventional clinical variables. Interestingly, Apolipoprotein B was not significant in the basic model of composite CVD, yet became significant at $P_{FDR}<0.05$ after adjusting for cardiovascular medication use, suggesting that therapeutic interventions may mask its true contribution to CVD risk. These observations underscore the importance of considering treatment effects when interpreting proteomic associations.

Our findings demonstrate that the newly developed proteomic risk score provides independent predictive value for CVD up to 17 years before the event. These results are consistent with previous studies using targeted affinity-based panels such as Olink and SOMAScan [14–17], which have also highlighted the utility of circulating proteins in risk prediction. By applying an untargeted mass spectrometry approach, our study captured proteins not typically included in targeted panels, which may represent novel contributors to long-term risk prediction. Furthermore, while most prior studies have concentrated on 10-year prediction horizons, our results suggest that proteomic profiling holds promise for identifying at-risk individuals much earlier in the disease course.

Study strengths include utilising one of the world’s largest mass spectrometry proteomics cohort studies, which also contained hundreds of incident CVD and death events. Many of the associations observed have not, to our knowledge, been previously studied in the context of CVD. Our unbiased approach, including rarely studied protein isoforms, adds depth and breadth to the underlying biology of CVD risk. The long follow-up period further supports early detection and risk assessment.

Despite these strengths, several limitations must be acknowledged. Firstly, the study cohort consists exclusively of individuals of European ancestries residing in Scotland, which may limit the generalisability of our findings to other populations. Secondly, while highly scalable and affordable, our mass spectrometry approach tends to measure the more abundant fraction of the proteome, excluding some known biomarkers such as cardiac troponin or C-reactive protein. Thirdly, in cases where peptides could not be uniquely assigned to a single protein, PGs were created. Each PG may contain several proteins that share peptides, meaning it is not always possible to identify which protein drives the measured signal. In addition, we identified previously unreported associations by cross-referencing with the Open Targets database. However, Open Targets is not fully comprehensive and may not include all published or emerging evidence. Finally, while external replication would strengthen the cardiovascular relevance and robustness of our findings, no comparable mass spectrometry-based dataset currently exists to support direct validation.

In conclusion, our findings demonstrate that circulating proteins and their isoforms are associated with incident CVD and mortality up to 17 years before the event, independent of traditional risk factors. By employing a mass spectrometry-based approach, we identified previously unreported associations that may have been overlooked in targeted studies. These findings could pave the way for the development of new biomarkers for early CVD detection and personalized prevention strategies. Future research should incorporate diverse cohorts and longitudinal sampling to deepen our understanding of the molecular mechanisms underlying CVD, enhance early disease prediction, and inform tailored interventions that improve patient outcomes.

5. Acknowledgments

The authors 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, health care assistants, and nurses.

6. Sources of Funding

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 Generation Scotland samples was carried out by the Genetics Core Laboratory at the Edinburgh Clinical Research Facility, Edinburgh, Scotland and was funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award STratifying Resilience and Depression Longitudinally (STRADL; Reference 104036/Z/14/Z). The DNA methylation data assayed for Generation Scotland was partially funded by a 2018 NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation (Ref: 27404; awardee: Dr David M Howard) and by a JMAS SIM fellowship from the Royal College of Physicians of Edinburgh (Awardee: Dr Heather C Whalley).

A.D.C. is supported by a Medical Research Council PhD Studentship in Precision Medicine with funding from the Medical Research Council Doctoral Training Program and the University of Edinburgh College of Medicine and Veterinary Medicine. R.E.M. and J.M. are supported by Alzheimer's Society major project grant AS-PG-19b-010. C.H. was funded by MRC Human Genetics Unit program (QTL in Health and Disease) (grant U.MC_UU_00007/10). J.A.R. is a University of Edinburgh Clinical Academic Track PhD student, supported by the Wellcome Trust (319878/Z/24/Z). H.M.S. is a student on the University of Edinburgh Translational Neuroscience PhD programme funded by the Wellcome Trust (218493/Z/19/Z).

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 221890/Z/20/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.

7. Disclosures

R.E.M has received a speaker fee from Illumina and is an advisor to the Epigenetic Clock Development Foundation. R.E.M. has received consultant fees from Optima partners. C.B., A.Z. and M.R. are co-founders and shareholders of Eliptica Ltd. C.B.M. is a consultant and shareholder of Eliptica Ltd. S.V., A.G. and S.A. are employed by Eliptica Ltd. All other authors declare no competing interests.

8. Author contributions

R.E.M., and A.D.C. were responsible for the conception and design of the study. R.E.M. and A.D.C. drafted the article. A.D.C and D.L.M. carried out the data analyses. J.M., J.R, H.M.S contributed to the analyses and methodology. C.D. and H.G. were involved in investigation; sample preparation, collating approximately 10000 samples and transferring into required format for mass spectrometry analysis, QC checks and helping coordinate shipping. P.A., M.W., C.B.M, A.G., S.A., C.B., A.Z. and M.R. were responsible for mass spectrometry analysis. A.C. facilitated data linkage. C.H., M.R., A.Z, K.L.E. and J.F.P. were involved in conceptualisation and provided consultation on the methodology. D.J.P contributed to data collection and preparation. All authors read and approved the final manuscript.

9. Supplemental Material

Supplemental Methods
Supplemental Tables 1-11
Supplemental Figures 1-8
Supplemental References 1-2

10. References

1. Ferrari, A. J. *et al.* Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet* **0**, (2024).
2. Reichlin, T. *et al.* Utility of Absolute and Relative Changes in Cardiac Troponin Concentrations in the Early Diagnosis of Acute Myocardial Infarction. *Circulation* **124**, 136–145 (2011).
3. Segraves, J. M. & Frishman, W. H. Highly Sensitive Cardiac Troponin Assays: A Comprehensive Review of Their Clinical Utility. *Cardiology in Review* **23**, 282 (2015).
4. Di Angelantonio, E. *et al.* B-Type Natriuretic Peptides and Cardiovascular Risk. *Circulation* **120**, 2177–2187 (2009).
5. Ridker, P. M., Rifai, N., Rose, L., Buring, J. E. & Cook, N. R. Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. *N Engl J Med* **347**, 1557–1565 (2002).
6. Ridker, P. M. C-Reactive Protein: Eighty Years from Discovery to Emergence as a Major Risk Marker for Cardiovascular Disease. *Clinical Chemistry* **55**, 209–215 (2009).
7. Lowe, G. D., Yarnell, J. W., Rumley, A., Bainton, D. & Sweetnam, P. M. C-reactive protein, fibrin D-dimer, and incident ischemic heart disease in the Speedwell study: are inflammation and fibrin turnover linked in pathogenesis? *Arterioscler Thromb Vasc Biol* **21**, 603–610 (2001).
8. Walldius, G. & Jungner, I. Apolipoprotein A-I versus HDL cholesterol in the prediction of risk for myocardial infarction and stroke. *Current Opinion in Cardiology* **22**, 359 (2007).
9. Mokou, M., Lygirou, V., Vlahou, A. & Mischak, H. Proteomics in cardiovascular disease: recent progress and clinical implication and implementation. *Expert Review of Proteomics* **14**, 117–136 (2017).
10. Gadd, D. A. *et al.* Blood protein assessment of leading incident diseases and mortality in the UK Biobank. *Nat Aging* **4**, 939–948 (2024).
11. Carrasco-Zanini, J. *et al.* Proteomic signatures improve risk prediction for common and rare diseases. *Nat Med* 1–10 (2024) doi:10.1038/s41591-024-03142-z.
12. Wajngarten, M. & Silva, G. S. Hypertension and Stroke: Update on Treatment. *Eur Cardiol* **14**, 111–115 (2019).
13. Lloyd-Jones, D. M. *et al.* Lifetime risk of coronary heart disease by cholesterol levels at selected ages. *Arch Intern Med* **163**, 1966–1972 (2003).
14. Gadd, D. A. *et al.* Blood protein levels predict leading incident diseases and mortality in UK Biobank. 2023.05.01.23288879 Preprint at https://doi.org/10.1101/2023.05.01.23288879 (2023).
15. Royer, P. *et al.* Large-scale plasma proteomics in the UK Biobank modestly improves prediction of major cardiovascular events in a population without previous cardiovascular disease. *Eur. J. Prev. Cardiol.* **31**, 1681–1689 (2024).
16. Mazidi, M. *et al.* Risk prediction of ischemic heart disease using plasma proteomics, conventional risk factors and polygenic scores in Chinese and European adults. *Eur J Epidemiol* **39**, 1229–1240 (2024).
17. Corlin, L. *et al.* Proteomic Signatures of Lifestyle Risk Factors for Cardiovascular Disease: A Cross-Sectional Analysis of the Plasma Proteome in the Framingham Heart Study. *J Am Heart Assoc* **10**, e018020 (2021).
18. Wang, Z. *et al.* A multiplex protein panel assay for severity prediction and outcome prognosis in patients with COVID-19: An observational multi-cohort study. *eClinicalMedicine* **49**, (2022).
19. Ryczkowska, K., Adach, W., Janikowski, K., Banach, M. & Bielecka-Dabrowa, A. Menopause and women's cardiovascular health: is it really an obvious relationship? *Arch Med Sci* **19**, 458–466 (2022).

20. Ae, S., Mm, W. & Ns, S. Sex differences in endothelial function important to vascular health and overall cardiovascular disease risk across the lifespan. *American journal of physiology. Heart and circulatory physiology* **315**, (2018).
21. Harding, A. T. & Heaton, N. S. The Impact of Estrogens and Their Receptors on Immunity and Inflammation during Infection. *Cancers* **14**, (2022).
22. Gusev, E. & Sarapultsev, A. Atherosclerosis and Inflammation: Insights from the Theory of General Pathological Processes. *Int J Mol Sci* **24**, 7910 (2023).
23. Dordevic, N. *et al.* Pervasive Influence of Hormonal Contraceptives on the Human Plasma Proteome in a Broad Population Study. 2023.10.11.23296871 Preprint at https://doi.org/10.1101/2023.10.11.23296871 (2023).
24. Messner, C. B. *et al.* Ultra-High-Throughput Clinical Proteomics Reveals Classifiers of COVID-19 Infection. *Cell Syst* **11**, 11-24.e4 (2020).
25. Smith, B. H. *et al.* Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med Genet* **7**, 74 (2006).
26. Vernardis, S. I. *et al.* The Impact of Acute Nutritional Interventions on the Plasma Proteome. *The Journal of Clinical Endocrinology and Metabolism* **108**, 2087 (2023).
27. Messner, C. B. *et al.* Ultra-fast proteomics with Scanning SWATH. *Nat Biotechnol* **39**, 846–854 (2021).
28. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* **17**, 41–44 (2020).
29. Bruderer, R. *et al.* Analysis of 1508 Plasma Samples by Capillary-Flow Data-Independent Acquisition Profiles Proteomics of Weight Loss and Maintenance. *Mol Cell Proteomics* **18**, 1242–1254 (2019).
30. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2023).
31. Rusilowicz, M., Dickinson, M., Charlton, A., O'Keefe, S. & Wilson, J. A batch correction method for liquid chromatography-mass spectrometry data that does not depend on quality control samples. *Metabolomics* **12**, 56 (2016).
32. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
33. 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* **42**, 2439–2454 (2021).
34. Dierks, C. *et al.* Menopause Hormone Replacement Therapy and Lifestyle Factors affect Metabolism and Immune System in the Serum Proteome of Aging Individuals. 2024.06.22.24309293 Preprint at https://doi.org/10.1101/2024.06.22.24309293 (2024).
35. Welsh, P. *et al.* Cardiac Troponin T and Troponin I in the General Population. *Circulation* **139**, 2754–2764 (2019).
36. Kuan, V. *et al.* A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *The Lancet Digital Health* **1**, e63–e77 (2019).
37. Rossum, G. van. *Python Tutorial*. (1995).
38. Davidson-Pilon, C. lifelines: survival analysis in Python. *Journal of Open Source Software* **4**, 1317 (2019).
39. Kassambara, A., Kosinski, M. & Biecek, P. *Survminer: Drawing Survival Curves Using 'Ggplot2'*. (2021).
40. Therneau, T. M. & Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model*. (Springer, New York, 2000).
41. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**, 1–22 (2010).
42. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
43. Buniello, A. *et al.* Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic Acids Research* **53**, D1467–D1475 (2025).

44. Somer, S. & Levy, A. P. The Role of Haptoglobin Polymorphism in Cardiovascular Disease in the Setting of Diabetes. *Int J Mol Sci* **22**, 287 (2020).
45. Schomburg, L., Orho-Melander, M., Struck, J., Bergmann, A. & Melander, O. Selenoprotein-P Deficiency Predicts Cardiovascular Disease and Death. *Nutrients* **11**, 1852 (2019).
46. Wanner, A. Alpha-1 Antitrypsin as a Therapeutic Agent for Conditions not Associated with Alpha-1 Antitrypsin Deficiency. *Alpha-1 Antitrypsin* 141–155 (2015) doi:10.1007/978-3-319-23449-6_8.
47. Ding, P.-H. & Jin, L. J. The role of lipopolysaccharide-binding protein in innate immunity: a revisit and its relevance to oral/periodontal health. *J Periodontal Res* **49**, 1–9 (2014).
48. Yang, F.-J. *et al.* Plasma Leucine-Rich α-2-Glycoprotein 1 Predicts Cardiovascular Disease Risk in End-Stage Renal Disease. *Sci Rep* **10**, 5988 (2020).
49. Maïga, S. F., Kalopissis, A.-D. & Chabert, M. Apolipoprotein A-II is a key regulatory factor of HDL metabolism as appears from studies with transgenic animals and clinical outcomes. *Biochimie* **96**, 56–66 (2014).
50. Birjmohun, R. S. *et al.* Apolipoprotein A-II Is Inversely Associated With Risk of Future Coronary Artery Disease. *Circulation* **116**, 2029–2035 (2007).
51. Melchior, J. T. *et al.* Apolipoprotein A-II alters the proteome of human lipoproteins and enhances cholesterol efflux from ABCA1. *Journal of Lipid Research* **58**, 1374–1385 (2017).
52. Sasaki, S. *et al.* Involvement of enhanced expression of classical complement C1q in atherosclerosis progression and plaque instability: C1q as an indicator of clinical outcome. *PLoS One* **17**, e0262413 (2022).
53. Mohammadzadeh, N. *et al.* The extracellular matrix proteoglycan lumican improves survival and counteracts cardiac dilatation and failure in mice subjected to pressure overload. *Scientific Reports* **9**, (2019).
54. Kv, E. *et al.* Lumican is increased in experimental and clinical heart failure, and its production by cardiac fibroblasts is induced by mechanical and proinflammatory stimuli. *The FEBS journal* **280**, (2013).
55. Iravanian, S. & Dudley, S. C. The Renin-Angiotensin-Aldosterone System (RAAS) and Cardiac Arrhythmias. *Heart Rhythm* **5**, s12–s17 (2008).
56. Tanvir Kahlon, M. D. *et al.* Angiotensinogen: More Than its Downstream Products: Evidence From Population Studies and Novel Therapeutics. *Heart Failure* (2022) doi:10.1016/j.jchf.2022.06.005.
57. Patel, V. B., Zhong, J.-C., Grant, M. B. & Oudit, G. Y. Role of the ACE2/Angiotensin 1–7 axis of the Renin-Angiotensin System in Heart Failure. *Circ Res* **118**, 1313–1326 (2016).
58. Richmond, A. *et al.* Genome-wide analysis of 439 mass spectrometry-based proteomic profiles in a population of 15,035 Scottish individuals. 2025.08.14.25333677 Preprint at https://doi.org/10.1101/2025.08.14.25333677 (2025).

7.3. Conclusion

Associations between the abundances of individual proteins and protein groups and the risk of CVD, its subtypes, and mortality revealed 48 potential biomarkers that significantly improved risk prediction beyond established clinical factors, with associations detectable up to 17 years before disease onset ($P<0.05/439$, Bonferroni-corrected). Of these, 24 protein – disease associations were not represented in the Open Targets database. Sex – specific analyses further revealed differential effects of complement system components in HF, suggesting distinct immune – related pathways between men and women. Finally, a composite proteomic risk score integrating multiple protein concentrations into a single measure yielded improved discrimination of incident CVD beyond established risk factors ($\Delta AUC = 0.010$, ROC $P = 0.013$).

This work demonstrates that MS-based profiling of the serum proteome provides a valuable source of candidate biomarkers for incident CVD and death. Collectively, these findings underscore the potential of large-scale proteomic analyses to advance understanding of disease mechanisms and to support more precise strategies for CVD prevention. A comprehensive discussion of the empirical results presented in **Chapters 5** to **7** is provided in the following chapter.

8. Discussion

8.1. Overview and Integration of the Thesis Aims

This chapter provides an integrated interpretation of the findings presented across **Chapters 5 – 7**, which collectively examined molecular biomarkers of CVD using epigenetic and proteomic data from large population-based cohorts. The overarching aim of this work was to identify and evaluate molecular signatures that improve understanding of CVD biology and enhance risk prediction beyond established clinical factors.

Firstly, I conducted a blood- and brain-based EWAS of smoking pack-years (n > 17,000 individuals, **Chapter 5**). This work addressed the first three aims of the thesis: to identify CpG sites associated with smoking pack-years in blood and brain using array- and sequencing-based DNAm data (**Aim 1**), to train a smoking EpiScore in >10,000 blood DNAm samples and compare its predictive performance with existing smoking EpiScores (**Aim 2**), and compare the DNA signal of self-reported smoking with that of epigenetic smoking via GWAS. (**Aim 3**). The study revealed robust methylation signatures of tobacco exposure, captured a broad molecular signature of smoking through the EpiScore, and linked these epigenetic patterns to underlying genetic determinants of smoking behaviour.

Secondly, I examined associations between 109 DNAm-derived protein EpiScores and incident CVD, including evaluation of a composite CVD EpiScore for its ability to enhance prediction beyond established clinical risk scores such as ASSIGN and SCORE2 (**Chapter 6**). This study addressed **Aim 4** of the thesis and demonstrated that methylation-based protein proxies capture subclinical disease processes and provide complementary information to conventional risk factors, modestly improving 10-year CVD risk prediction.

Thirdly, I applied untargeted MS to profile 439 highly abundant serum proteins in over 8,000 individuals and investigated their associations with incident CVD and mortality (**Chapter 7**). Addressing **Aim 5**, the study identified both established biomarkers and proteins not previously reported in the Open Targets database that predicted CVD and death, with associations detectable up to 17 years before disease onset. These findings illuminate molecular pathways involved in early disease development and suggest that untargeted proteomic profiling may help identify novel targets for risk stratification or prevention.

In the following sections, I will integrate the key findings of this thesis, from novel insights in EWASs to the development of EpiScores and identification of protein biomarkers for CVD. I will then discuss the potential for multi-omic integration and clinical translation, reflect on the limitations of the work, and present the final conclusions and implications for future research.

8.2. EWAS-Based Discovery

The EPIC array-based Bayesian EWAS of pack years presented in **Chapter 5** replicated the majority of previously reported smoking-associated methylation signals, including the well-established loci in *AHRR*, *F2RL3*, and *ALPPL2*, while also identifying several novel CpGs not catalogued in the EWAS Catalog [170]. It is important to note that the catalog is not comprehensive; associations from studies that did not submit their results are not included, underscoring the need for complementary manual literature review to capture the full landscape of reported findings.

The Bayesian model identified 42 loci associated with smoking pack-years with a PIP exceeding 80%. This figure is smaller than the number of significant associations reported at FDR-corrected thresholds in the largest frequentist EWASs of comparable smoking phenotypes (e.g., current vs never smoker comparisons: EPIC array – 65,857 CpGs [171]; 450k array – 18,760 CpGs [183]; see **Section 2.3.3**). This reduction in discovery breadth likely reflects several inherent characteristics of the Bayesian models. First, the informative priors induce sparsity, shrinking weaker or less certain associations toward zero. Second, the Bayesian framework adjusts for measured and unmeasured covariates within the model, reducing spurious associations driven by confounding factors without requiring explicit separate regression steps (see comparison between adjusted and unadjusted EWAS results published as a Supplemental Material of smoking EWAS – **Chapter 5**). Together, these features prioritise effect size precision and model stability over sheer discovery breadth, producing a more conservative set of associations that are likely to be more robust and biologically interpretable – that is, less influenced by noise or confounding and more reflective of true underlying effects.

This phenomenon has been observed in other contexts. For example, Smith *et al*. [336] compared Bayesian and frequentist EWASs of BMI and found that while the frequentist approach produced a large and difficult-to-interpret set of associations, the Bayesian framework yielded a smaller, more interpretable subset of robust signals. Similarly, in the present analysis, although the number of identified CpGs was smaller, these sites likely represent reproducible associations that are less susceptible to overfitting or cohort-specific artefacts. Together, these results highlight how analytical choices – particularly modelling assumptions – profoundly influence discovery, interpretation, and downstream biological inference in large-scale EWASs.

Interestingly, although the NGS-based EWAS identified a greater number of CpG sites associated with smoking when compared to the array-based EWAS, these CpGs mapped to a smaller set of unique genes, as reflected by the lower number of distinct association peaks observed in the NGS Manhattan plot. This pattern occurred despite the NGS platform offering theoretically superior genomic coverage and single-base resolution [166]. This discrepancy likely arises from a combination of technical and analytical challenges that remain inherent to NGS-based methylation profiling. One key limitation is variability in sequencing depth and coverage across CpG sites [337]. In NGS data, each CpG is measured based on the number of sequencing reads that cover it, but this coverage is often uneven – some sites are represented by many reads, while others by only a few. CpGs with low read depth yield less precise methylation estimates, which increases measurement noise and reduces statistical power to detect associations [337]. These issues are particularly pronounced in genomic regions containing abundant repetitive sequences, where alignment is difficult [338]. Consequently, true biological signals may be obscured by technical variability rather than being absent altogether.

Despite these challenges, the strongest signals were consistently replicated across the EPIC array and NGS platforms, reinforcing their reliability and biological validity. This overlap suggests that although NGS currently offers lower statistical efficiency, it provides valuable validation and the ability to investigate CpG sites not represented on commercial arrays. Furthermore, NGS enables the quantification of non-CpG methylation [339], which may provide deeper insight into the molecular mechanisms underlying smoking-related epigenetic variation.

Both the NGS and array-based analyses excluded CpGs on the X and Y chromosomes due to several technical and biological reasons. Technical reasons include unreliable hybridisation on arrays, as well as uneven sequencing coverage and alignment difficulties in repetitive or homologous regions, which compromise data quality and reproducibility when using NGS [338]. Biological reasons include sex-specific copy number differences (XX vs. XY) and the complex effects of X-chromosome inactivation in females, which generate highly variable methylation patterns that cannot be directly compared between sexes [340]. The resulting mosaic methylation pattern on the X chromosome and the sparse, repeat-rich nature of the Y chromosome make direct comparison between sexes statistically and biologically challenging. This exclusion highlights a persistent gap in current EWAS designs, as sex chromosomes likely contribute to interindividual variation in many phenotypes, including smoking response. Future advances in long-read and single-molecule sequencing technologies – offering improved basecalling accuracy, uniform coverage, and haplotype-level resolution – may overcome these limitations.

It is likely that in the future, NGS data will be more common in epidemiological studies. Recent advancements, such as those highlighted by UK Biobank's announcement to adopt ONT sequencing to profile 50,000 human samples [341], will enable genome-wide epigenetic analyses with unprecedented resolution. However, several gaps remain. Future research will need to systematically evaluate the technical comparability and reproducibility of NGS-derived DNAm estimates relative to established array-based platforms, particularly across diverse tissues, populations, and experimental conditions. Analytical pipelines must be adapted and optimised for the unique characteristics of NGS data, while maintaining computational scalability for studies with tens of thousands of samples. Moreover, the biological interpretation of NGS methylation data, including regional aggregation of CpGs, non-CpG methylation, and sex-chromosome analyses, requires careful methodological development to ensure robust and generalisable findings. Addressing these challenges will enable NGS to realise its full potential for high-resolution epigenome-wide association studies and the development of more accurate and transferable epigenetic predictors of disease risk.

8.3. From EWAS to Prediction: EpiScores

In **Chapter 5**, I evaluated several strategies for developing EpiScores, including elastic net [326] and Bayesian EWAS-based approach [168]. The strongest predictive performance was achieved using elastic net applied to a subset of CpGs pre-selected based on EWAS significance. While this strategy produced robust predictors, it has inherent limitations: restricting training to previously identified suggestive CpGs may exclude additional informative sites, whereas including all CpGs could capture more signals but carries a high risk of overfitting. Bayesian models partially mitigated this issue through probabilistic shrinkage and automatic regularisation, reducing overfitting, but they still underperformed compared with the elastic net approach applied to EWAS-informed CpGs.

These methodological choices directly influenced the performance of the smoking EpiScore (mCigarette) trained as part of **Chapter 5**. This predictor performed comparably to the best existing smoking EpiScores and achieved higher R^2 values with pack years, likely reflecting its training on pack years as a continuous outcome rather than binary smoking status. It demonstrated robust predictive ability across age groups and European cohorts. However, cross-ancestry validation was limited by incomplete phenotypic information in the Health for Life in Singapore study, where replication of the mCigarette predictor was not possible.

EpiScores can also be constructed by integrating information from multiple individual predictors into a single composite score (see **Chapter 6**). Such composite predictors capture patterns that may be missed by individual EpiScores by aggregating signals across the CpG sets underlying each predictor. This approach enhances predictive power and reduces dataset-specific noise. Composite scores are also less prone to overfitting and tend to provide more stable, generalisable measures of molecular risk, making them well suited for applications across diverse populations and for multi-omic integration. However, their performance ultimately depends on the relevance of their constituent predictors – if the input EpiScores lack informative features, the composite score is unlikely to perform well.

In this study, the composite CVD EpiScore was constructed from 109 available protein EpiScores (Gadd *et al.*[1], **Chapter 6**), which did not include several key CVD biomarkers such as GDF15 or ApoB. The modest improvement in prediction beyond established clinical risk models, including ASSIGN and SCORE2, likely reflects both this limited coverage and the fact that the majority of CVD risk is already captured by established risk factors. Expanding the EpiScore library to encompass a broader range of proteins will therefore be essential to enhance predictive utility. Alternatively, instead of aggregating existing protein EpiScores, a CVD EpiScore could be trained directly on genome-wide CpG data. To mitigate overfitting, CpGs could first be prioritised based on findings from large-scale EWASs of CVD, similar to the feature-selection strategy applied in smoking-related predictors. However, the feasibility of this approach is currently limited by the scarcity of well-powered EWASs of CVD outcomes, and future studies should revisit this strategy as larger and more comprehensive datasets become available.

One of the strongest predictors of CVD is age, reflecting the cumulative biological processes that drive atherosclerosis over time [342]. However, chronological age alone does not fully capture an individual's health or physiological state [343]. Measures of biological ageing, such as DNAm-based clocks (e.g., GrimAge [197] or PhenoAge [344]), provide additional information about systemic ageing processes and disease susceptibility. For example, a 40-year-old individual may have a biological age closer to 50, reflecting an elevated risk of age-related diseases including CVD. Clocks such as GrimAge have been consistently associated with an increased risk of CVD [345]. Incorporating DNAm-based biological age proxies into clinical risk tools could therefore refine risk stratification and improve predictive accuracy.

Furthermore, a more integrative modelling framework could be developed that combines EpiScores for established risk factors (e.g., DNAm age, smoking exposure EpiScore such as miCigarette, etc.), comorbidities (e.g., EpiScores for type 2 diabetes developed by Cheng *et al.* [346]), and environmental exposures (e.g., DNAm signature of air pollution currently studied by Robertson *et al.* [347] and others [348]). Such a composite score would more comprehensively capture the multifactorial nature of CVD risk and offer a systems-level approach to prediction. Rigorous validation across large and diverse populations will, however, be essential to ensure its robustness and translational relevance.

Improving EpiScores also requires increasing the size of the training dataset, as demonstrated in **Chapter 5**. For example, the mCigarette predictor was trained using data from 17,865 GS individuals, whereas earlier pack-years predictor developed by McCartney *et al*. [155] used the same cohort and a similar elastic net approach but was trained on a smaller sample of 5,087 GS participants. Despite these methodological similarities, mCigarette achieved substantially greater predictive power, underscoring the critical influence of sample size. Consequently, future large-scale efforts leveraging resources such as UK Biobank DNAm data could yield even more accurate EpiScores. Nevertheless, the comparable performance observed across existing smoking EpiScores suggests that this proxy may be nearing predictive saturation, beyond which further gains are likely to be marginal.

Another important consideration is tissue specificity. DNAm patterns are highly tissue-dependent. In **Chapter 5**, I showed that although individual CpGs could effectively discriminate smokers from never-smokers, the significant sites differed between blood and brain. As a result, the mCigarette score was not predictive in brain tissue. These findings indicate that separate EWASs and tissue-specific EpiScores will be necessary to study the effects of smoking – or other exposures – across organs. Similar constraints are likely to apply to protein EpiScores, as blood methylation may not always reflect molecular changes occurring in disease-relevant tissues such as the heart. Emerging approaches, including liquid biopsies and analysis of circulating cell-free DNA, may help capture tissue-specific signals from otherwise inaccessible organs [349], although their application for disease prevention remains in early development.

The presence of a detectable biological signal is crucial for accurate EpiScore generation. Not all proteins are amenable to EpiScore construction: for example, it was not possible to generate an EpiScore for cTnT, and the EpiScore for cTnI was not significantly associated with outcomes (**Chapter 6**). This likely reflects biological constraints, as elevations in troponins signal myocardial injury rather than stable, long-term epigenetic regulation detectable in blood. Consequently, while some proteins are best captured through direct measurement, others – particularly those with chronic roles – may be more effectively proxied by DNAm [191]. In some cases, EpiScores can outperform measured protein levels when the latter are transient or noisy, whereas for proteins such as troponins, direct quantification remains the gold standard. Future predictive models should therefore integrate both measured and methylation-derived data to optimise performance for specific endpoints.

Finally, even the best proxy is not equivalent to the measured phenotype. GWAS of DNAm GrimAge identified loci distinct from those found in GWAS of self-reported smoking (**Chapter 5**), reflecting differences in both the type of data captured and the underlying error structures. Self-reported smoking captures conscious behaviour but is prone to recall and social-desirability bias, whereas EpiScores are objective and continuous but may include noise from non-smoking exposures such as air pollution. These differences can generate distinct association signals even when both measures aim to represent the same underlying exposure. Accordingly, proxy-based and direct measures should be regarded as complementary.

8.4. Proteins and EpiScores as Potential Biomarkers

Thirty-six protein EpiScores and 48 mass spectrometry-derived proteins were significantly associated with CVD and mortality outcomes after applying a conservative Bonferroni correction for multiple testing (**Chapters 6** and **7**). Several markers replicated well-established proteomic biomarkers, while others pointed to novel candidates.

Interestingly, all proteins which showed evidence of a causal link to CAD in a recent MR study [350] – Apolipoprotein B, Apolipoprotein E, and Apolipoprotein A – were associated with established CV risk factors (**Chapter 7**). This pattern raises the possibility that proteins with causal roles in disease may exert their effects through upstream pathways already captured by risk factors, and therefore may not emerge as independent markers in models designed to enhance risk prediction beyond conventional scores.

More broadly, these findings emphasise the importance of studying individual risk factors to elucidate the biological mechanisms underlying CVD and of carefully considering potential sources of confounding, such as medication use. For example, the association between Apolipoprotein B and composite CVD became significant only after adjustment for CV medications, which may suggest that pharmacological interventions mask underlying disease biology. Such observations underscore the necessity of accounting for treatment effects – and other potential sources of confounding – when interpreting multi-omic associations. Medication use was not considered in the protein EpiScore analysis (**Chapter 6**), which primarily focused on enhancing clinical CVD risk prediction tools. Future studies using proteins and protein EpiScores to investigate CVD biology should explicitly account for medication use, to help disentangle intrinsic biological effects from treatment-related influences and to maximise mechanistic insight into CVD pathophysiology.

Importantly, several protein-CVD associations identified in **Chapter 7** appeared potentially novel, as they were not listed in the Open Targets database. However, it should be acknowledged that Open Targets is not fully comprehensive and may not capture all published or emerging evidence. For instance, some proteins initially identified as novel in this analysis – such as Immunoglobulin heavy variable 3 and Complement C1q – have in fact been previously reported to associate with subclinical CVD [351,352].This highlights the importance of integrating multiple data sources and up-to-date literature to accurately interpret the novelty of protein – disease associations.

Although both measured proteins and their corresponding EpiScores were associated with CVD in **Chapters 6** and **7**, they seem to capture complementary aspects of the same underlying biological processes. Circulating proteins reflect the current physiological and inflammatory state, marking active biochemical events such as oxidative stress, immune activation, and tissue remodelling within the vessel wall. For example, elevated plasma levels of acute-phase reactants including haptoglobin, alpha-1-antitrypsin, and complement components may indicate ongoing endothelial injury and inflammatory responses [353]. In contrast, protein EpiScores capture longer-term, more stable regulatory influences encoded in the methylome. EpiScores for markers such as CRP, CCL18, CXCL9, and VCAM1 highlight sustained inflammatory and endothelial dysfunction pathways [354], while VEGFA, HGF, and FGF21 may reflect chronic vascular remodelling and tissue adaptation [355]. In other words, while both layers reflect similar pathophysiological processes, protein levels likely indicate what is happening in the moment, whereas EpiScores reflect cumulative exposures or regulatory predispositions that shape risk over time. Integrating these complementary layers therefore provides a more complete view of CVD biology, combining the temporal stability of methylation-derived measures with the acute sensitivity of proteomic data, which may enhance long-term risk prediction and mechanistic understanding.

It is important to highlight that the identified associations between circulating proteins, protein EpiScores, and incident CVD provide promising leads for biomarker development, but establishing their clinical and biological relevance requires further analyses beyond association testing. Replication in independent datasets would be the most direct way to confirm robustness and generalisability; however, to my knowledge, no other cohort currently provides MS-based proteomic data of comparable scale to that used in **Chapter 7**, limiting the possibility of external validation of these data at present. Expanding MS resources across large population studies will therefore be essential to enable replication and cross-platform comparisons in the future. Analytical validation using other proteomic technologies such as immunoassays would help confirm that the observed signals are not platform-specific artefacts [356]. In addition, complementary wet-lab experiments could further support biomarker validation by mapping their presence in relevant tissues, testing their effects in cultured cells, and assessing their behaviour in established atherosclerosis mouse models [357,358].

Beyond validation, future studies could also refine the statistical approaches used to identify associations. In **Chapter 7**, I examined CVD both as a composite outcome and through its individual components, reflecting the distinct pathophysiological drivers underlying different subtypes of CVD. Analysing sub-types allows the identification of proteins linked to specific disease processes, while a composite outcome captures systemic risk spanning multiple event types, providing a more integrated view of CV biology. However, modelling outcomes separately can obscure shared biological mechanisms that contribute to several conditions simultaneously. For instance, a protein involved in systemic inflammation may influence both MI and HF, yet its association may not reach significance for either outcome in isolation. Traditional univariate approaches may therefore underestimate biomarkers that reflect broader CV risk. Although such multivariate tools were not available when this project began, new Bayesian methods such as MAJA (Multivariate Adaptive Joint Association analysis) [359] now enable the simultaneous testing of biomarkers across correlated outcomes. By accounting for shared biological variation rather than treating each outcome independently, these approaches can better capture underlying disease processes and offer a more integrated understanding of CVD risk architecture.

Furthermore, in **Chapter 6**, multiple testing was addressed using the Bonferroni correction, a stringent method that controls the error rate by dividing the significance threshold by the number of tests. While this approach effectively reduces the risk of false positives, it can be overly conservative, particularly when the underlying data has a complex correlation structure, as is often the case with proteomic data. In **Chapter 7**, I observed that several proteins with established causal links to CVD were significant only when using a Benjamini-Hochberg threshold [334], rather than a Bonferroni correction. This highlights the importance of selecting an appropriate multiple-testing correction method that balances stringency with power to detect true biological associations.

To further maximise discovery, MS data from **Chapter 7** could be combined with affinity-based measurements, as the two approaches tend to capture different parts of the proteome [360,361]. This difference reflects both the distinct detection principles of each method and the underlying complexity of the proteome (see **Section 2.4**).

Another strategy to enhance proteomic coverage is to perform MS experiments in modes optimised for detecting the low-abundance fraction of the proteome. While technically feasible, these high-sensitivity approaches require longer acquisition times, greater instrument stability, and often additional sample fractionation or enrichment [362,363]. As a result, they are typically several times more costly per sample – often in the range of hundreds rather than tens of pounds – and substantially lower in throughput compared with standard workflows. These factors currently limit their feasibility for large population cohorts. Future research should therefore assess the scalability and cost-effectiveness of low-abundance MS, while also exploring cross-platform integration as a practical way to maximise coverage, discover novel biomarkers, and improve predictive modelling for CVD risk.

The importance of extending proteomic coverage is underscored by the present findings: the CVD protein score derived from mass spectrometry-based measurements significantly improved CVD risk prediction beyond established clinical factors, although the predictive gains were more modest than those observed for affinity-based scores (ΔAUC ≈ 0.010 in **Chapter 7** vs. ΔAUC ≈ 0.025 in Olink-based studies discussed in **Section 2.4.2**). This discrepancy is expected given the lower sensitivity of MS for detecting low-abundance proteins. Clinically established cardiac biomarkers such as troponins and natriuretic peptides circulate at concentrations below the detection limit of the mass spectrometry platform used here (see **Section 2.4**), meaning that important components of CVD risk are likely to have been missed.

Finally, future studies could expand beyond conventional Cox PH models to explore a wider range of machine learning (ML) approaches for constructing composite multi-omic scores and predicting CVD risk. In **Chapter 6**, I applied two methods to derive composite CVD EpiScores: a Cox PH model with elastic net regularisation and a Random Forest. For downstream CVD risk modelling, I primarily relied on the conventional Cox PH models. Other ML approaches currently under investigation include gradient-boosted trees [364], support vector machines [365], and neural networks (more recently, Generative Pre-trained Transformer (GPT)-based predictive frameworks) [366]. Systematic reviews indicate that ML models can yield improvements in discrimination (e.g., higher C-index or AUC) compared to Cox-based models when large, heterogeneous datasets are available – for example, a recent meta-analysis reported AUCs of approximately 0.87 for ML models versus 0.77 for traditional models [367]. Nevertheless, many ML models continue to face challenges with generalisability, interpretability, and external validation, which likely explains why Cox-based approaches remain the preferred standard for clinical implementation. Thus, while ML holds promise for high-dimensional and complex data contexts, the simpler and more interpretable Cox framework continues to offer the most practical route for risk score integration in current clinical settings.

8.5. Towards Clinical Translation

While multiple multi-omic markers of CVD have been identified to date, their clinical translation remains challenging [368]. At present, the improvement in predictive performance offered by these biomarkers - whether assessed individually or as composite scores - beyond established risk factors is typically modest (see **Sections 2.3.4** and **2.4.2**). The identification of more informative markers in the future may enhance their utility and enable more effective risk prediction.

Another key barrier is the uncertainty about how multi-omic tools could be implemented in practice to improve patient outcomes [369]. Even if such models identify individuals at higher risk, it remains unclear whether this information would lead to different or more effective clinical management than current approaches focused on established risk factors such as smoking, hypertension, and hypercholesterolaemia (see **Section 1.7**). Demonstrating that targeted interventions based on multi-omic risk profiles improve outcomes beyond standard prevention strategies is therefore an essential next step.

In addition to these challenges, cost remains a major consideration. NGS using high-throughput platforms such as Illumina or Oxford Nanopore, although increasingly accessible, remains expensive. For context, the National Human Genome Research Institute reports that sequencing a single human genome currently costs approximately US$1,000 [370], depending on coverage and platform, with additional expenses for library preparation, reagents, and bioinformatics analysis. Applying NGS at the population level for prevention screening therefore involves substantial cumulative costs.

In the epigenetic domain, widely used platforms such as the Infinium MethylationEPIC BeadChips typically cost approximately US$250 – 500 per sample [371]. Targeted epigenetic arrays, which focus on specific CpG sites, are generally less expensive. In proteomics, large-scale multiplexed affinity assays such as those offered by Olink span a broad range of costs depending on panel size and throughput. For instance, Olink Reveal, which quantifies approximately 1,000 proteins, is priced at approximately US$98 per sample, making it one of the more affordable options [372]. On the MS side, traditional discovery workflows (shotgun/LC-MS/MS) are typically in the region of US$150 – US$300 per sample for simpler workflows, and can exceed US$1,000 per sample for more complex or deep-coverage proteomics. For example, a micro-costing study of MS-based proteomics in diagnostics estimated a total per-patient cost around US$897 when applied to rare disease testing [373]. Notably, recent advances in ultra-high-throughput MS workflows have demonstrated that direct costs can be reduced to nearly US$10 per sample [239] – an approach that was employed in **Chapter 7**.

Taken together, these cost estimates highlight several strategies to improve the clinical feasibility of multi-omic CVD risk assessment. First, targeted panels that focus on the most predictive genetic variants, CpG sites, or protein markers can substantially reduce per-sample costs by limiting testing to the features with the highest clinical utility, rather than performing genome-, epigenome-, or proteome-wide assays for every patient. This approach preserves predictive power while minimising unnecessary expenditure on less informative regions or analytes. Second, high-throughput processing and batching of samples take advantage of economies of scale, as running large numbers of samples simultaneously reduces reagent waste, machine idle time, and per-sample labour costs. Core facilities or centralised laboratories that implement standardised workflows are particularly effective in achieving these efficiencies. Finally, formal cost-effectiveness modelling is essential to determine when testing is both economically viable and clinically valuable. In practice, multi-omic testing would be most useful for individuals in whom traditional risk calculators perform less reliably, such as younger adults with family history of early-onset CVD, older adults where short-term risk is overestimated, or individuals from underrepresented ethnic groups. By focusing on these subpopulations, clinicians can use multi-omic data to refine risk estimates, guide preventive interventions, and avoid unnecessary testing in clearly low- or high-risk individuals. This targeted, evidence-based approach maximises clinical benefit while addressing cost and ethical considerations, providing a roadmap for translating multi-omic CVD risk assessment into routine practice.

Beyond cost, significant technical barriers must be overcome before multi-omic assays can be implemented reliably in the clinic. Assay reproducibility and standardisation are critical, as measurements can differ depending on the laboratory, platform, or workflow. For example, protein quantification using SOMAScan, Olink, and MS are only modestly correlated, reflecting differences in detection chemistry and analytical sensitivity, which complicates the interpretation of absolute values and the establishment of clinical thresholds [374,375]. In addition, issues with protein group interpretation and limited data for low-abundance proteins further hinder the translation of these assays into clinically actionable tools (see **Section 8.6.3**).

Similarly, EpiScores presents challenges: these scores are currently scaled to the cohort in which they were derived, meaning that a “high-risk” value in one population may not correspond to the same level of risk in another. Such cohort-specific scaling limits the use of universal clinical cut-offs for both protein and epigenetic markers, complicating risk stratification across diverse populations. Age is another critical consideration. Protein EpiScores exhibit age-dependent associations with measured protein levels, which may influence their predictive performance [211]. This suggests that incorporating age-specific adjustments or using age-specific EpiScores could improve the precision and clinical relevance of risk models. Similarly, comorbidities and physiological conditions – such as obesity, diabetes, or chronic inflammation – may influence both protein levels and EpiScore associations with CVD. These conditions can alter biomarker distributions, potentially modifying effect sizes and predictive performance. For example, chronic inflammation can elevate inflammatory proteins (such as CRP or IL-6) or alter DNAm at inflammation-sensitive CpG sites (see **Section 2.3.4**). In these cases, the measured EpiScore or protein level may reflect the comorbidity rather than the true CVD risk. Therefore, evaluating multi-omic tests across a broad spectrum of health statuses remains essential to ensure accurate and equitable CVD risk prediction.

Moreover, the associations between multi-omic markers and CVD risk may differ across populations. This point is discussed in more detail in the next section (**Section 8.6.1**). Most research on EpiScores has been conducted in European-ancestry cohorts (see **Section 2.3.4**), and the findings presented in this thesis are no exception. Variations in genetic background (e.g., allele frequencies of *PCSK9* variants, see **Section 2.3.3**), environmental exposures (such as air pollution or dietary patterns), lifestyle factors (including physical activity, smoking, or alcohol consumption), and disease prevalence (for example, higher rates of hypertension or type 2 diabetes in certain populations) can influence both the distribution of multi-omic markers and their associations with CVD [183]. As a result, predictive performance established in one population may not be directly generalisable to another. Should future studies demonstrate that EpiScores predict CVD in non-European populations, multi-omic risk models may require adaptation to ensure accuracy and clinical relevance.

Finally, an additional challenge lies in the processing and integration of multi-omic outputs into actionable clinical risk scores. Overcoming this requires the development of streamlined pipelines and clinical-grade software that automate data processing, reporting, and quality control, thereby reducing bioinformatics burden. Clinician-friendly dashboards that translate molecular measurements into interpretable, actionable scores are also essential for adoption in routine practice.

8.6. Limitations

I have outlined the specific limitations of each study in previous sections (**Sections 8.2 – 8.5**) and in the published articles (**Chapters 5–7**). In this section, I discuss the broader limitations of the thesis, with particular focus on the cohorts, omics measurements, and statistical methodologies employed.

8.6.1. Cohorts

The cohort studies used in this thesis have several limitations. Firstly, GS, the LBC1936, and ALSPAC are all British cohorts, representing high-income populations with relatively good access to healthcare and preventive services. However, the majority of CVD cases occur in low- and middle-income countries, which account for ~80% of global CVD deaths (see **Section 1.6.1**) [51]. Therefore, while the findings from these studies provide valuable insight into disease mechanisms, their generalisability to populations with different environmental exposures, healthcare systems, and risk factor profiles may be limited. Furthermore, accessibility of biomarker testing remains a major challenge in low-resource settings. In this thesis, I focused on blood-based biomarkers, as they are minimally invasive and suitable for large-scale population screening. Yet, even blood sampling may be difficult to implement widely in low-income regions. Future studies should explore biomarkers that can be measured in more accessible and non-invasively collected samples, such as saliva obtained via postal kits.

All large population-based cohorts, including those used in this thesis, are subject to selection biases. Participants in GS, the LBC1936, and ALSPAC tend to have higher educational attainment and socioeconomic status than the population from which they were recruited, reflecting well-documented patterns of non-representativeness across volunteer cohorts [320,376]. This introduces uncertainty regarding the generalisability of the findings, as molecular and phenotypic associations observed in relatively healthy, well-educated individuals may differ from those in more socioeconomically diverse or disadvantaged groups. For example, a study of participation bias – a specific form of selection bias – in patients with HF found that non-participation was associated with a two-fold increase in mortality [377]. In the context of this thesis, associations between blood-based omics markers and CVD outcomes may therefore be specific to this healthier, more advantaged subset of participants.

Furthermore, all cohorts are affected by survivorship bias [378]. Individuals who remain active in long-term studies are typically healthier and more engaged with healthcare services, whereas those with more severe disease or higher mortality risk are more likely to withdraw over time. This is particularly relevant in LBC1936, where participants are extensively screened at multiple time points (see **Section 4.2**). As a result, data from the most vulnerable segments of the population may be underrepresented, which could influence the observed associations between molecular markers and health outcomes.

Using data from individuals of European ancestry resulted in limited genetic diversity. This is an important consideration, as associations between phenotypes and molecular markers of health and disease may vary across populations with different ancestral backgrounds (see **Section 8.5**). For example, the cross-ancestry meta-analysis of smoking by Joehanes *et al*. [183] reported a high overall correlation of smoking-associated methylation loci between individuals of European and African ancestry (Spearman r=0.89 for current vs never smokers and Spearman r=0.75 for former vs never smokers), but also identified several CpG sites with ancestry-specific effects. One such site, cg00706683 (mapped to *ECEL1P2*), exhibited persistent differential methylation in individuals of European ancestry but not in those of African ancestry. In individuals of European ancestry, that site did not revert to never-smoker levels even 30 years after smoking cessation. This illustrates that ancestry can influence both the presence and persistence of smoking-associated DNAm changes.

Several variables used in my analyses were derived from self-reported questionnaire data, which are inherently prone to recall bias and misclassification [379]. For example, smoking status and aspects of medical history (such as self-reported type 2 diabetes) were based on participant recall and may therefore be under- or misreported. While linkage to electronic health records in GS helps to mitigate some of these limitations, inaccuracies in self-reported data can still compromise the precision of phenotype definitions and, consequently, the strength and reproducibility of molecular associations.

8.6.2. DNA and DNAm Quantification

The only form of epigenetic variation considered in this thesis was cytosine methylation, specifically 5mC. It should be recognised that standard bisulfite conversion, which underlies array-based DNAm profiling, does not distinguish 5-hydroxymethylcytosine (5hmC) from 5mC; both modifications are read as "methylated", meaning that array-based measurements in this thesis may overestimate 5mC levels in the presence of 5hmC [380]. In addition, arrays do not capture other potentially relevant DNA modifications, such as 6-methyladenine, which may play a role in gene regulation and disease [381,382]. The Oxford Nanopore sequencing data generated as part of this thesis provide a valuable opportunity to study these additional modifications, although the sample size is small (n=46 individuals). Beyond DNA modifications, other epigenetic mechanisms, including RNA modifications and histone modifications, may also serve as biomarkers of CVD [383,384]; however, these are not yet routinely measured in large population-based cohort studies.

8.6.3. Proteomics

A key challenge in large-scale proteomic studies, including those presented in this thesis, is the issue of non-uniquely mapped peptides [385]. In MS-based proteomics, peptide fragments are sometimes shared between multiple proteins, making it difficult to confidently assign them to a single protein. This ambiguity can introduce false-positive signals, dilute true associations and reduce the likelihood of replication in independent cohorts. Is also complicates biological interpretation as it is not always clear which protein is being quantified [386]. Interpretation challenges arising from protein group ambiguity, combined with the previously discussed limited detection of low-abundance proteins – including established biomarkers such as cTnI and NT-proBNP – constrain the immediate applicability of these findings for clinical risk prediction. Addressing these challenges will be essential for translating MS – derived proteomic discoveries into reliable tools for CVD prevention and patient care.

8.6.4. Statistical Methods

Most EpiScores in this thesis were trained using elastic net regression, which is well-suited to high-dimensional, correlated DNAm data. Alternative methods, including LASSO, Ridge regression, and Bayesian regression, were also tested for the mCigarette score (**Chapter 5** and the associated peer review file available online), but elastic net consistently showed the best predictive performance. Cross-validation was used to optimise the regularisation parameter lambda, which controls the overall strength of shrinkage applied to the coefficients, while the mixing parameter alpha, which balances the contributions of LASSO and Ridge penalties, was fixed at 0.5. Although this choice is common for EpiScores [343], simultaneous optimisation of both alpha and lambda could potentially improve predictive accuracy and robustness.

All statistical models rely on underlying assumptions, and violations of these assumptions can affect the reliability of the results. The vast majority of the models applied in this thesis assume a linear relationship between predictors and outcomes, which may overlook potential non-linear effects [387,388]. Assessing such assumptions across high-dimensional datasets, such as individual CpG sites, is challenging and often not feasible on a large scale. Nevertheless, non-linear relationships did not appear to substantially influence the findings in **Chapter 6**, where composite CVD EpiScores trained using linear elastic net and non-linear Random Forest approaches demonstrated comparable predictive performance.

The Cox models used in **Chapters 6** and **7** to analyse time-to-event outcomes (CVD, its subtypes, or all-cause mortality) rely on the proportional hazards assumption. This assumption states that the hazard ratio between groups defined by a predictor remains constant over time – in other words, the relative risk of the event for one group compared with another does not change during follow-up [327]. To evaluate this assumption, I reported both local (biomarker-specific) and global (full-model) p-values from the cox.zph() function in R, which tests whether the Schoenfeld residuals for each predictor are correlated with time [328]. Significant correlations indicate violations of the proportional hazards assumption. In **Chapters 6** and **7**, this assumption did not hold for all biomarkers (see visualisations in **Section 4.2.2**), and all such cases were explicitly reported. Associations where the assumption was violated should therefore be interpreted with caution.

Lastly, while **Chapters 6** and **7** provide robust observational analyses of associations between protein EpiScores, measured proteins, and CVD outcomes, these findings are inherently correlational and cannot establish causality. One approach that could strengthen causal inference in this context is MR. MR uses genetic variants associated with an exposure of interest – here, protein levels or EpiScores – as instrumental variables to estimate the causal effect of that exposure on an outcome, under the assumption that these variants are randomly allocated at conception and are not influenced by confounding factors [389]. Applying MR to the associations identified in **Chapters 6** and **7** could help determine whether changes in specific proteins or protein EpiScores contribute causally to CVD risk, rather than simply being correlated with it. While the findings from **Chapter 7** were supported by a recent MR analysis reporting that proteins and protein groups annotated to three genes (*APOE*, *APOB* and *LPA*) may be causally linked to CAD [350], no such analysis was conducted for EpiScores (**Chapter 6**). MR would provide stronger evidence for prioritising particular proteins as potential therapeutic targets or preventive biomarkers.

MR findings should ideally be complemented by additional lines of evidence, such as longitudinal analyses that can clarify temporal relationships and dynamic patterns of risk [390]. For example, repeated biomarker measurements could reveal whether within-person increases in specific proteins precede CVD onset, supporting a causal interpretation, or instead follow early disease processes, suggesting reverse causation. Together, triangulating evidence from MR, longitudinal studies, and experimental approaches – including studies in animal models [357] – would provide a more rigorous framework for distinguishing causal biomarkers from correlational signals

8.7. Recommendations

The overarching aim of this thesis was to identify biomarkers of CVD that both deepen understanding of its underlying biology and improve risk prediction beyond established clinical factors. The work focused on two major classes of blood-based biomarkers: DNAm-derived EpiScores and directly measured protein levels. In this section, I outline several recommendations for future research directions in both domains.

In **Chapter 5**, I performed an analysis of epigenetic loci associated with smoking by conducting three EWASs of smoking pack years: one using data from the Illumina EPIC array, and two using NGS data – one generated with the TWIST Human Methylome kit, representing targeted short-read sequencing, and the other using ONT, representing untargeted long-read sequencing. This comparative analysis was designed as a first step toward understanding the implications of transitioning from array-based to sequencing-based methylation profiling in large-scale epidemiological studies. However, this represents only an initial exploration.

I recommend that future research systematically evaluates the comparability of DNAm estimates obtained using array- and NGS-based technologies to better understand how methodological differences influence methylation quantification. This could be achieved through correlation analyses at overlapping CpG sites to identify regions of high and low concordance across technologies. Such analyses will be crucial for determining which loci are reliably measurable using both array and sequencing approaches.

In parallel, I recommend that the transferability and robustness of EpiScores derived from different methylation platforms be carefully assessed. Benchmarking EpiScores generated using EPIC array data against those obtained from NGS-based methylation calls will help determine whether predictive models trained on one technology generalise to another. These analyses should also guide the development of normalisation or harmonisation strategies, thereby facilitating integrative analyses across evolving methylation profiling platforms.

In the same chapter, I developed an EpiScore for smoking pack years, called mCigarette. This score captures DNAm-based signatures of cumulative smoking exposure and provides an objective biomarker that may complement or refine self-reported smoking measures. I recommend that future studies evaluate the predictive utility of mCigarette in relation to CVD outcomes by comparing its performance with self-reported smoking status or pack years. In particular, I recommend testing whether the inclusion of mCigarette improves risk prediction when added to established clinical models such as ASSIGN or SCORE2. Such analyses could determine whether DNAm-derived measures of smoking capture residual risk not fully explained by self-report, potentially reflecting misreporting or exposure misclassification.

Next, I recommend that future studies expand the repertoire of protein EpiScores and assess their added predictive value beyond established clinical CVD risk tools. In **Chapter 6**, I developed a composite CVD EpiScore integrating multiple protein EpiScores; its main limitation was the absence of several key CVD biomarkers in the training data, which likely limited its performance. Although the composite CVD EpiScore remained statistically significant when added to the ASSIGN model, the observed improvement in predictive accuracy was modest. Therefore, I recommend that future efforts focus on generating high-quality EpiScores for proteins most

strongly implicated in CVD pathophysiology – particularly inflammatory, metabolic, and vascular markers – to enhance model coverage [391,392]. Incorporating these additional protein proxies into the composite CVD EpiScore is expected to yield greater predictive gains, potentially approaching those achieved by models based on directly measured protein concentrations.

In **Chapter 7**, I reported multiple associations between MS-based protein levels and CVD; however, these findings could not be externally validated, as no other cohort currently provides MS proteomic data of comparable scale. Therefore, I recommend that: a) cohort studies expand their multi-omic libraries by measuring the proteome using MS, and b) future studies validate the associations identified in **Chapter 7** using these external datasets. I further recommend that existing protein measurements obtained through affinity-based technologies be used alongside MS data to jointly identify and confirm CVD biomarkers. Combining the complementary strengths of both approaches may also facilitate the development of more robust and predictive composite protein scores.

Finally, I recommend that future research directly compare the predictive value of protein EpiScores with that of their measured counterparts for CVD risk. Although some studies suggest that certain protein EpiScores outperform measured protein levels in association analyses [191,207], this may not hold true for all proteins. For example, in **Chapter 6**, I demonstrated that it is not possible to generate an EpiScore for cTnT, highlighting that for some key cardiac biomarkers, measured protein concentrations may be preferable in predictive models.

8.8. Final Summary

CVD is the leading cause of death worldwide and often develops silently over many years, making early detection essential. The findings in this thesis indicate that DNAm- and protein-based biomarkers can provide valuable insights into the biological pathways underlying CVD and modestly improve risk prediction beyond established risk factors. Future research should prioritise validating these biomarkers in diverse, multi-ancestry populations, expanding the range of protein and DNAm markers examined, and integrating multi-omic data to develop more robust and clinically relevant predictive tools.

Bibliography

1. Gadd, D. A. *et al.* Epigenetic scores for the circulating proteome as tools for disease prediction. *eLife* **11**, e71802 (2022).
2. Naghavi, M. *et al.* Global burden of 288 causes of death and life expectancy decomposition in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet* **403**, 2100–2132 (2024).
3. Martin, S. S. *et al.* 2025 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. *Circulation* **151**, e41–e660 (2025).
4. Kapuku, G. K. & Kop, W. J. Classification of Cardiovascular Diseases: Epidemiology, Diagnosis, and Treatment. in *Handbook of Cardiovascular Behavioral Medicine* (eds. Waldstein, S. R., Kop, W. J., Suarez, E. C., Lovallo, W. R. & Katzel, L. I.) 45–80 (Springer, New York, NY, 2022). doi:10.1007/978-0-387-85960-6_3.
5. Steindel, S. J. International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. *J Am Med Inform Assoc* **17**, 274–282 (2010).
6. Sanchis-Gomar, F., Perez-Quilis, C., Leischik, R. & Lucia, A. Epidemiology of coronary heart disease and acute coronary syndrome. *Ann Transl Med* **4**, 256 (2016).
7. Sharry, J. M. *et al.* Delay in Seeking Medical Help following Transient Ischemic Attack (TIA) or “Mini-Stroke”: A Qualitative Study. *PLOS ONE* **9**, e104434 (2014).
8. Easton, J. D. *et al.* Definition and Evaluation of Transient Ischemic Attack. *Stroke* **40**, 2276–2293 (2009).

9. Nordanstig, J. *et al.* Peripheral arterial disease (PAD) – A challenging manifestation of atherosclerosis. *Preventive Medicine* **171**, 107489 (2023).
10. Gillespie, H. S., Lin, C. C. H. & Prutkin, J. M. Arrhythmias in structural heart disease. *Curr Cardiol Rep* **16**, 510 (2014).
11. Solomon, E. P., Berg, L. P., & Diana W. Martin. *Biology*. (MULTICO Oficyna Wydawnicza).
12. Libby, P. *et al.* Atherosclerosis. *Nat Rev Dis Primers* **5**, 1–18 (2019).
13. Rajendran, P. *et al.* The Vascular Endothelium and Human Diseases. *Int J Biol Sci* **9**, 1057–1069 (2013).
14. Jang, E., Robert, J., Rohrer, L., von Eckardstein, A. & Lee, W. L. Transendothelial transport of lipoproteins. *Atherosclerosis* **315**, 111–125 (2020).
15. Witztum, J. L. & Steinberg, D. The oxidative modification hypothesis of atherosclerosis: does it hold for humans? *Trends Cardiovasc Med* **11**, 93–102 (2001).
16. Amento, E. P., Ehsani, N., Palmer, H. & Libby, P. Cytokines and growth factors positively and negatively regulate interstitial collagen gene expression in human vascular smooth muscle cells. *Arteriosclerosis and Thrombosis: A Journal of Vascular Biology* **11**, 1223–1230 (1991).
17. Schade, D. S., Gonzales, K. & Eaton, R. P. Stop Stenting; Start Reversing Atherosclerosis. *The American Journal of Medicine* **134**, 301–303 (2021).
18. Wilkins, J. T., Gidding, S. S. & Robinson, J. G. Can Atherosclerosis Be Cured? *Curr Opin Lipidol* **30**, 477–484 (2019).
19. Morris, P. *Eureka: Cardiovascular Medicine*. (Scion Publishing Ltd, Banbury, 2015).

20. Manfredi, R. *et al.* Angina in 2022: Current Perspectives. *J Clin Med* **11**, 6891 (2022).
21. Noronha, B., Duncan, E. & Byrne, J. A. Optimal medical management of angina. *Curr Cardiol Rep* **5**, 259–265 (2003).
22. DeVon, H. A., Ryan, C. J., Ochs, A. L. & Shapiro, M. Symptoms across the continuum of acute coronary syndromes: differences between women and men. *Am J Crit Care* **17**, 14–24 (2008).
23. McSweeney, J. C. *et al.* Women's Early Warning Symptoms of Acute Myocardial Infarction. *Circulation* **108**, 2619–2623 (2003).
24. Thygesen, K. *et al.* Fourth Universal Definition of Myocardial Infarction (2018). *Circulation* **138**, e618–e651 (2018).
25. van Oosterhout, R. E. M. *et al.* Sex Differences in Symptom Presentation in Acute Coronary Syndromes: A Systematic Review and Meta-analysis. *Journal of the American Heart Association* **9**, e014733 (2020).
26. Rathore, S. S. *et al.* Association of door-to-balloon time and mortality in patients admitted to hospital with ST elevation myocardial infarction: national cohort study. *BMJ* **338**, b1807 (2009).
27. Benamer, H. *et al.* Longer pre-hospital delays and higher mortality in women with STEMI: the e-MUST Registry. *EuroIntervention* **12**, e542-549 (2016).
28. Welsh, P. *et al.* Cardiac Troponin T and Troponin I in the General Population. *Circulation* **139**, 2754–2764 (2019).
29. Chiam, K. *et al.* Brain PET and Cerebrovascular Disease. *PET Clinics* **18**, 115–122 (2023).

30. Portegies, M. L. P., Koudstaal, P. J. & Ikram, M. A. Chapter 14 - Cerebrovascular disease. in *Handbook of Clinical Neurology* (eds. Aminoff, M. J., Boller, F. & Swaab, D. F.) vol. 138 239–261 (Elsevier, 2016).
31. Albers, G. W. *et al.* Transient Ischemic Attack — Proposal for a New Definition. *New England Journal of Medicine* **347**, 1713–1716 (2002).
32. Hart, R. G., Pearce, L. A. & Koudstaal, P. J. Transient Ischemic Attacks in Patients With Atrial Fibrillation. *Stroke* **35**, 948–951 (2004).
33. Sacco, R. L. *et al.* An Updated Definition of Stroke for the 21st Century. *Stroke* **44**, 2064–2089 (2013).
34. Bamford, J., Sandercock, P., Dennis, M., Warlow, C. & Burn, J. Classification and natural history of clinically identifiable subtypes of cerebral infarction. *The Lancet* **337**, 1521–1526 (1991).
35. Muir, K. W. Stroke. *Medicine* **41**, 169–174 (2013).
36. Lewandowski, C. A., Rao, C. P. V. & Silver, B. Transient Ischemic Attack: Definitions and Clinical Presentations. *Annals of Emergency Medicine* **52**, S7–S16 (2008).
37. Orfei, M. D. *et al.* Anosognosia for hemiplegia after stroke is a multifaceted phenomenon: a systematic review of the literature. *Brain* **130**, 3075–3090 (2007).
38. Nielsen, J. A., Zielinski, B. A., Ferguson, M. A., Lainhart, J. E. & Anderson, J. S. An Evaluation of the Left-Brain vs. Right-Brain Hypothesis with Resting State Functional Connectivity Magnetic Resonance Imaging. *PLOS ONE* **8**, e71275 (2013).
39. Vossel, S., Weiss, P. H., Eschenbeck, P. & Fink, G. R. Anosognosia, neglect, extinction and lesion site predict impairment of daily living after right-hemispheric stroke. *Cortex* **49**, 1782–1789 (2013).

40. Bailey, E. L., Smith, C., Sudlow, C. L. M. & Wardlaw, J. M. Pathology of lacunar ischemic stroke in humans - A systematic review. *Brain Pathology* **22**, 583–591 (2012).
41. Qureshi, A. I. *et al.* Spontaneous Intracerebral Hemorrhage. *New England Journal of Medicine* **344**, 1450–1460 (2001).
42. Evenson, K. R., Rosamond, W. D. & Morris, D. L. Prehospital and In-Hospital Delays in Acute Stroke Care. *Neuroepidemiology* **20**, 65–76 (2001).
43. Mair, G. & Wardlaw, J. M. Imaging of acute stroke prior to treatment: current practice and evolving techniques. *British Journal of Radiology* **87**, 20140216 (2014).
44. Babić, A. *et al.* Blood Biomarkers in Ischemic Stroke Diagnostics and Treatment—Future Perspectives. *Medicina* **61**, 514 (2025).
45. Bozkurt, B. *et al.* Universal definition and classification of heart failure: a report of the Heart Failure Society of America, Heart Failure Association of the European Society of Cardiology, Japanese Heart Failure Society and Writing Committee of the Universal Definition of Heart Failure. *European Journal of Heart Failure* **23**, 352–380 (2021).
46. Adams, K. F. New epidemiologic perspectives concerning mild-to-moderate heart failure. *The American Journal of Medicine* **110**, 6–13 (2001).
47. Ames, M. K., Atkins, C. E. & Pitt, B. The renin-angiotensin-aldosterone system and its suppression. *J Vet Intern Med* **33**, 363–382 (2019).
48. Watson, R. D. S., Gibbs, C. R. & Lip, G. Y. H. ABC of Heart Failure: Clinical features and complications. *BMJ* **320**, 236–239 (2000).

49. British Heart Foundation. Global Heart & Circulatory Diseases Factsheet. https://www.bhf.org.uk/-/media/files/for-professionals/research/heart-statistics/bhf-cvd-statistics-global-factsheet.pdf (2025).
50. Lopez, A. D. & Adair, T. Is the long-term decline in cardiovascular-disease mortality in high-income countries over? Evidence from national vital statistics. *Int J Epidemiol* **48**, 1815–1823 (2019).
51. World Heart Federation. *World Heart Report 2023: Confronting the World's Number One Killer*. https://world-heart-federation.org/wp-content/uploads/World-Heart-Report-2023.pdf (2023).
52. Wise, J. Early deaths from cardiovascular disease reach 14 year high in England. *BMJ* **384**, q176 (2024).
53. Borland, S. Why are UK cardiovascular deaths in under 65s rising again? *BMJ* **389**, r1015 (2025).
54. National Records of Scotland. Vital Events Reference Tables. https://www.nrscotland.gov.uk/publications/vital-events-reference-tables-2023/ (2025).
55. Gao, T., Wang, X. C., Chen, R., Ngo, H. H. & Guo, W. Disability adjusted life year (DALY): A useful tool for quantitative assessment of environmental pollution. *Science of The Total Environment* **511**, 268–287 (2015).
56. Megan Lindstrom, P. *et al.* Global Burden of Cardiovascular Diseases and Risks Collaboration, 1990-2021. *Journal of the American College of Cardiology* https://doi.org/10.1016/j.jacc.2022.11.001 (2022) doi:10.1016/j.jacc.2022.11.001.
57. Bhatnagar, A. Environmental Determinants of Cardiovascular Disease. *Circ Res* **121**, 162–180 (2017).

58. NICE. Risk factors for CVD. https://cks.nice.org.uk/topics/cvd-risk-assessment-management/background-information/risk-factors-for-cvd/ (2025).
59. Kartiosuo, N. *et al.* Cardiovascular Risk Factors in Childhood and Adulthood and Cardiovascular Disease in Middle Age. *JAMA Netw Open* **7**, e2418148 (2024).
60. Avis, S. R., Vernon, S. T., Hagström, E. & Figtree, G. A. Coronary artery disease in the absence of traditional risk factors: a call for action. *European Heart Journal* **42**, 3822–3824 (2021).
61. Khot, U. N. Prevalence of Conventional Risk Factors in Patients With Coronary Heart Disease. *JAMA* **290**, 898 (2003).
62. Figtree, G. A. *et al.* Mortality in STEMI patients without standard modifiable risk factors: a sex-disaggregated analysis of SWEDEHEART registry data. *The Lancet* **397**, 1085–1094 (2021).
63. Vishram, J. K. K. *et al.* Impact of Age on the Importance of Systolic and Diastolic Blood Pressures for Stroke Risk. *Hypertension* **60**, 1117–1123 (2012).
64. Le, N. N. *et al.* Unravelling the Distinct Effects of Systolic and Diastolic Blood Pressure Using Mendelian Randomisation. *Genes* **13**, (2022).
65. Public Health England. Health matters: combating high blood pressure. *GOV.UK* https://www.gov.uk/government/publications/health-matters-combating-high-blood-pressure/health-matters-combating-high-blood-pressure (2017).
66. Prospective Studies Collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *The Lancet* **360**, 1903–1913 (2002).
67. Lonn, E. M. *et al.* Blood-Pressure Lowering in Intermediate-Risk Persons without Cardiovascular Disease. *New England Journal of Medicine* **374**, 2009–2020 (2016).

68. SPRINT Research Group *et al.* A Randomized Trial of Intensive versus Standard Blood-Pressure Control. *N Engl J Med* **373**, 2103–2116 (2015).
69. Ettehad, D. *et al.* Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. *The Lancet* **387**, 957–967 (2016).
70. Ference, B. A. *et al.* Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel. *European Heart Journal* **38**, 2459–2472 (2017).
71. Valdes-Marquez, E. *et al.* Relative effects of LDL-C on ischemic stroke and coronary disease. *Neurology* **92**, e1176–e1187 (2019).
72. Cholesterol Treatment Trialists' (CTT) Collaboration *et al.* Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170,000 participants in 26 randomised trials. *Lancet* **376**, 1670–1681 (2010).
73. Pirie, K. *et al.* The 21st century hazards of smoking and benefits of stopping: a prospective study of one million women in the UK. *Lancet* **381**, 133–141 (2013).
74. Doll, R. & Hill, A. B. The mortality of doctors in relation to their smoking habits; a preliminary report. *Br Med J* **1**, 1451–1455 (1954).
75. Calle, E. E. *et al.* The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. *Cancer* **94**, 2490–2501 (2002).
76. Garfinkel, L. Selection, follow-up, and analysis in the American Cancer Society prospective studies. *Natl Cancer Inst Monogr* **67**, 49–52 (1985).
77. Doll, R., Peto, R., Boreham, J. & Sutherland, I. Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ* **328**, 1519 (2004).

78. Bolego, C., Poli, A. & Paoletti, R. Smoking and gender. *Cardiovascular Research* **53**, 568–576 (2002).
79. Duncan, M. S. *et al.* Association of Smoking Cessation With Subsequent Risk of Cardiovascular Disease. *JAMA* **322**, 642–650 (2019).
80. Ishida, M., Sakai, C., Kobayashi, Y. & Ishida, T. Cigarette Smoking and Atherosclerotic Cardiovascular Disease. *J Atheroscler Thromb* **31**, 189–200 (2024).
81. Yamaguchi, N. H. Smoking, immunity, and DNA damage. *Transl Lung Cancer Res* **8**, S3–S6 (2019).
82. Anstey, K. J., von Sanden, C., Salim, A. & O'Kearney, R. Smoking as a Risk Factor for Dementia and Cognitive Decline: A Meta-Analysis of Prospective Studies. *Am J Epidemiol* **166**, 367–378 (2007).
83. Li, R., Luo, L., Yuan, C. & Zhu, Q. Association of smoke exposure with cognitive function trajectories among middle and old-aged adults: evidence from the China Health and Retirement Longitudinal Study. *J Glob Health* **15**, 04150.
84. Toda, N. & Okamura, T. Cigarette smoking impairs nitric oxide-mediated cerebral blood flow increase: Implications for Alzheimer's disease. *Journal of Pharmacological Sciences* **131**, 223–232 (2016).
85. Elbejjani, M. *et al.* Cigarette smoking and cerebral blood flow in a cohort of middle-aged adults. *J Cereb Blood Flow Metab* **39**, 1247–1257 (2019).
86. Rajeev, V. *et al.* Chronic cerebral hypoperfusion: a critical feature in unravelling the etiology of vascular cognitive impairment. *Acta Neuropathol Commun* **11**, 93 (2023).
87. Power, M. C. *et al.* Smoking and white matter hyperintensity progression: the ARIC-MRI Study. *Neurology* **84**, 841–848 (2015).

88. Alber, J. *et al.* White matter hyperintensities in vascular contributions to cognitive impairment and dementia (VCID): Knowledge gaps and opportunities. *Alzheimers Dement (N Y)* **5**, 107–117 (2019).
89. Meysami, S. *et al.* Smoking predicts brain atrophy in 10,134 healthy individuals and is potentially influenced by body mass index. *npj Dement.* **1**, 17 (2025).
90. Gray, J. C. *et al.* Associations of cigarette smoking with gray and white matter in the UK Biobank. *Neuropsychopharmacol.* **45**, 1215–1222 (2020).
91. Karama, S. *et al.* Cigarette smoking and thinning of the brain's cortex. *Mol Psychiatry* **20**, 778–785 (2015).
92. Gons, R. A. R. *et al.* Cigarette smoking is associated with reduced microstructural integrity of cerebral white matter. *Brain* **134**, 2116–2124 (2011).
93. Zuo, W., Peng, J. & Wu, J. Relationship of smoking cessation duration and cognitive function among middle-aged and older adults in China: a national cross-sectional study. *Front. Public Health* **12**, (2025).
94. Bahorik, A. L. *et al.* Early to Midlife Smoking Trajectories and Cognitive Function in Middle-Aged US Adults: the CARDIA Study. *J Gen Intern Med* **37**, 1023–1030 (2022).
95. Park, C. *et al.* Fasting Glucose Level and the Risk of Incident Atherosclerotic Cardiovascular Diseases. *Diabetes Care* **36**, 1988–1993 (2013).
96. Shaye, K., Amir, T., Shlomo, S. & Yechezkel, S. Fasting glucose levels within the high normal range predict cardiovascular outcome. *Am Heart J* **164**, 111–116 (2012).
97. Dal Canto, E. *et al.* Diabetes as a cardiovascular risk factor: An overview of global trends of macro and micro vascular complications. *European Journal of Preventive Cardiology* **26**, 25–32 (2019).

98. Winters-Miner, L. A. *et al.* Chapter 13 - Personalized Medicine. in *Practical Predictive Analytics and Decisioning Systems for Medicine* (eds. Winters-Miner, L. A. et al.) 176–204 (Academic Press, 2015). doi:10.1016/B978-0-12-411643-6.00013-2.

99. Carlsson, A. C. *et al.* Neighbourhood socioeconomic status and coronary heart disease in individuals between 40 and 50 years. *Heart* **102**, 775–782 (2016).

100. Loucks, E. B. *et al.* Life-course socioeconomic position and incidence of coronary heart disease: the Framingham Offspring Study. *Am J Epidemiol* **169**, 829–836 (2009).

101. Lee, M., Khan, M. M. & Wright, B. Is Childhood Socioeconomic Status Related to Coronary Heart Disease? Evidence From the Health and Retirement Study (1992-2012). *Gerontology and Geriatric Medicine* **3**, 2333721417696673 (2017).

102. Marshall, I. J. *et al.* The effects of socioeconomic status on stroke risk and outcomes. *Lancet Neurol* **14**, 1206–1218 (2015).

103. Shakoor, A. *et al.* Socio-economic inequalities and heart failure morbidity and mortality: A systematic review and data synthesis. *ESC Heart Failure* **12**, 927–941 (2025).

104. Kivimäki, M. *et al.* Association between socioeconomic status and the development of mental and physical health conditions in adulthood: a multi-cohort study. *Lancet Public Health* **5**, e140–e149 (2020).

105. Woodward, M., Brindle, P., Tunstall-Pedoe, H., & for the SIGN group on risk estimation*. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart* **93**, 172–176 (2005).

106. Johri, N. *et al.* Association of cardiovascular risks in rheumatoid arthritis patients: Management, treatment and future perspectives. *Health Sciences Review* **8**, 100108 (2023).

107. Sun, X. *et al.* Characterizing the polygenic overlap and shared loci between rheumatoid arthritis and cardiovascular diseases. *BMC Medicine* **22**, 152 (2024).

108. Piepoli, M. F. *et al.* 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts)Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). *Eur Heart J* **37**, 2315–2381 (2016).

109. D'Agostino, R. B. *et al.* General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* **117**, 743–753 (2008).

110. Badawy, M. A. E. M. D. *et al.* Evaluation of cardiovascular diseases risk calculators for CVDs prevention and management: scoping review. *BMC Public Health* **22**, 1742 (2022).

111. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* **357**, j2099 (2017).

112. Scottish Intercollegiate Guidelines Network. SIGN 149 • Risk estimation and the prevention of cardiovascular disease. https://www.sign.ac.uk/assets/sign149.pdf (2017).

113. 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* **42**, 2439–2454 (2021).

114. Cooney, M. T., Dudina, A. L. & Graham, I. M. Value and Limitations of Existing Scores for the Assessment of Cardiovascular Risk: A Review for Clinicians. *Journal of the American College of Cardiology* **54**, 1209–1227 (2009).

115. 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* **42**, 2439–2454 (2021).

116. Welsh, P., Kimenai, D. M. & Woodward, M. Updating the Scottish national cardiovascular risk score: ASSIGN version 2.0. *Heart* **111**, 557–564 (2025).

117. Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).

118. National Human Genome Research Institute. Human Genomic Variation. https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genomic-variation (2023).

119. Rotimi, C. N. & Jorde, L. B. Ancestry and Disease in the Age of Genomic Medicine. *New England Journal of Medicine* **363**, 1551–1558 (2010).

120. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

121. Illumina. *HumanExome BeadChips Data Sheet*. https://www.bioresource.nihr.ac.uk/media/43dpibvc/datasheet_humanexome_beadchips.pdf (2011).

122. Illumina. *Infinium™ Global Screening Array-24 v3.0 BeadChip*. https://emea.illumina.com/content/dam/illumina-

marketing/documents/products/datasheets/infinium-global-screening-array-data-sheet-370-2016-016.pdf.

123. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499–511 (2010).

124. Illumina. *Infinium Assay Workflow*. https://www.illumina.com/content/dam/illumina-marketing/documents/products/workflows/workflow_infinium_ii.pdf (2025).

125. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* **15**, 121–132 (2014).

126. Pei, X. M. *et al.* Targeted Sequencing Approach and Its Clinical Applications for the Molecular Diagnosis of Human Diseases. *Cells* **12**, 493 (2023).

127. Dijk, E. L. van, Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends in Genetics* **30**, 418–426 (2014).

128. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *Am J Hum Genet* **110**, 179–194 (2023).

129. Cerezo, M. *et al.* The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. *Nucleic Acids Research* **53**, D998–D1005 (2025).

130. Altshuler, D., Donnelly, P., & The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).

131. Uffelmann, E. *et al.* Genome-wide association studies. *Nat Rev Methods Primers* **1**, 59 (2021).

132. Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* **47**, 1121–1130 (2015).

133. McPherson, R. & Tybjaerg-Hansen, A. Genetics of Coronary Artery Disease. *Circulation Research* **118**, 564–578 (2016).

134. Wei, B., Liu, Y., Li, H., Peng, Y. & Luo, Z. Effect of 9p21.3 (lncRNA and CDKN2A/2B) variant on lipid profile. *Front. Cardiovasc. Med.* **9**, (2022).

135. Brænne, I. *et al.* Prediction of causal candidate genes in coronary artery disease loci. *Arterioscler Thromb Vasc Biol* **35**, 2207–2217 (2015).

136. Zanetti, D., Carreras-Torres, R., Esteban, E., Via, M. & Moral, P. Potential Signals of Natural Selection in the Top Risk Loci for Coronary Artery Disease: 9p21 and 10q11. *PLOS ONE* **10**, e0134840 (2015).

137. Peterson, A. S., Fong, L. G. & Young, S. G. PCSK9 function and physiology. *J Lipid Res* **49**, 1152–1156 (2008).

138. Cohen, J. C., Boerwinkle, E., Mosley, T. H. & Hobbs, H. H. Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. *New England Journal of Medicine* **354**, 1264–1272 (2006).

139. Erzurumluoglu, A. M. *et al.* Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci. *Mol Psychiatry* **25**, 2392–2409 (2020).

140. Liu, J. Z. *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* **42**, 436–440 (2010).

141. Thorgeirsson, T. E. *et al.* Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet* **42**, 448–453 (2010).

142. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* **42**, 441–447 (2010).

143. Wu, W. *et al.* Associations between smoking behavior-related alleles and the risk of melanoma. *Oncotarget* **7**, 47366–47375 (2016).

144. Li, K., Liang, S. & Mi, H. Preliminary exploration of the potential biological functions and prognosis values of RAB4B in pan-cancer combing with experimental validation in BLCA. *Transl Cancer Res* **13**, 613–633 (2024).

145. Teruo, N., Akemi, T., Izumi, Y. & Michie, Y. Biphasic effects of smoking on human serum dopamine-β-hydroxylase activity. *Toxicology Letters* **60**, 325–328 (1992).

146. Chen, L.-S. & Bierut, L. J. Genomics and personalized medicine: *CHRNA5-CHRNA3-CHRNB4* and smoking cessation treatment. *Journal of Food and Drug Analysis* **21**, S87–S90 (2013).

147. Bestor, T. H., Edwards, J. R. & Boulard, M. Notes on the role of dynamic DNA methylation in mammalian development. *PNAS* **112**, 6796–6799 (2015).

148. Beck, S. & Rakyan, V. K. The methylome: approaches for global DNA methylation profiling. *Trends in Genetics* **24**, 231–237 (2008).

149. Russo, G., Tramontano, A., Iodice, I., Chiariotti, L. & Pezone, A. Epigenome Chaos: Stochastic and Deterministic DNA Methylation Events Drive Cancer Evolution. *Cancers* **13**, 1800 (2021).

150. Villicaña, S. & Bell, J. T. Genetic impacts on DNA methylation: research findings and future perspectives. *Genome Biol* **22**, 127 (2021).

151. McCartney, D. L. *et al.* Epigenetic signatures of starting and stopping smoking. *EBioMedicine* **37**, 214–220 (2018).

152. Rider, C. F. & Carlsten, C. Air pollution and DNA methylation: effects of exposure in humans. *Clinical Epigenetics* **11**, 131 (2019).

153. Fang, M., Chen, D. & Yang, C. S. Dietary polyphenols may affect DNA methylation. *J Nutr* **137**, 223S-228S (2007).

154. Niculescu, M. D. & Zeisel, S. H. Diet, Methyl Donors and DNA Methylation: Interactions between Dietary Folate, Methionine and Choline. *The Journal of Nutrition* **132**, 2333S-2335S (2002).

155. McCartney, D. L. *et al.* Epigenetic prediction of complex traits and death. *Genome Biology* **19**, 136 (2018).

156. Illumina. *Automated Bisulfite Conversion for Infinium ™ Methylation BeadChips*. https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/automated-bisulfite-infinium-methylation-tech-note-m-gl-00144/automated-bisulfite-Infinium-methylation-tech-note-m-gl-00144.pdf (2021).

157. Illumina. *Illumina Methylation BeadChips Achieve Breadth of Coverage Using Two Infinium Chemistries*. (2022).

158. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* **17**, 208 (2016).

159. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).

160. Illumina. *Infinium HumanMethylation450 BeadChip*. https://support.illumina.com/content/dam/illumina-marketing/documents/products/product_information_sheets/product_info_hm450.pdf (2011).

161. Illumina. *Infinium™ MethylationEPIC BeadChip*. https://support.illumina.com/content/dam/illumina-support/documents/downloads/productfiles/methylationepic/infinium-methylation-epic-ds-1070-2015-008.pdf (2019).

162. Carreras-Gallo, N. *et al.* Creation and validation of the first infinium DNA methylation array for the human imprintome. *Epigenetics Communications* **4**, 5 (2024).

163. Illumina. *Infinium MethylationEPIC v2.0 BeadChip*. (2022).

164. Edwards, J. R. *et al.* Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Res* **20**, 972–980 (2010).

165. Bock, C. *et al.* Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. *Nat Biotechnol* **34**, 726–737 (2016).

166. Twist Bioscience. Twist Human Methylome Panel. https://www.twistbioscience.com/products/ngs/fixed-panels/human-methylome-panel?tab=data (2025).

167. Saffari, A. *et al.* Estimation of a significance threshold for epigenome-wide association studies. *Genet Epidemiol* **42**, 20–33 (2018).

168. Trejo Banos, D. *et al.* Bayesian reassessment of the epigenetic architecture of complex traits. *Nat Commun* **11**, 2865 (2020).

169. Brooks, S. Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society: Series D (The Statistician)* **47**, 69–100 (1998).

170. Battram, T. *et al.* The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Res* **7**, 41 (2022).

171. Hoang, T. T. *et al.* Comprehensive evaluation of smoking exposures and their interactions on DNA methylation. *eBioMedicine* **100**, 104956 (2024).

172. Pfeiffer, L. *et al.* DNA methylation of lipid-related genes affects blood lipid levels. *Circ Cardiovasc Genet* **8**, 334–342 (2015).

173. Braun, K. V. E. *et al.* Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. *Clin Epigenetics* **9**, 15 (2017).

174. Sayols-Baixeras, S. *et al.* Identification and validation of seven new loci showing differential DNA methylation related to serum lipid profile: an epigenome-wide approach. The REGICOR study. *Hum Mol Genet* **25**, 4556–4565 (2016).

175. Hedman, Å. K. *et al.* Epigenetic Patterns in Blood Associated With Lipid Traits Predict Incident Coronary Heart Disease Events and Are Enriched for Results From Genome-Wide Association Studies. *Circ Cardiovasc Genet* **10**, e001487 (2017).

176. Chambers, J. C. *et al.* Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. *Lancet Diabetes Endocrinol* **3**, 526–534 (2015).

177. Cardona, A. *et al.* Epigenome-Wide Association Study of Incident Type 2 Diabetes in a British Population: EPIC-Norfolk Study. *Diabetes* **68**, 2315–2326 (2019).

178. Hillary, R. F. *et al.* Blood-based epigenome-wide analyses of 19 common disease states: A longitudinal, population-based linked cohort study of 18,413 Scottish individuals. *PLoS Med* **20**, e1004247 (2023).

179. Juvinao-Quintero, D. L. *et al.* DNA methylation of blood cells is associated with prevalent type 2 diabetes in a meta-analysis of four European cohorts. *Clin Epigenetics* **13**, 40 (2021).

180. Richard, M. A. *et al.* DNA Methylation Analysis Identifies Loci for Blood Pressure Regulation. *Am J Hum Genet* **101**, 888–902 (2017).

181. Huang, Y. *et al.* Identification, Heritability, and Relation With Gene Expression of Novel DNA Methylation Loci for Blood Pressure. *Hypertension* **76**, 195–205 (2020).

182. Zeilinger, S. *et al.* Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation. *PLOS ONE* **8**, e63812 (2013).

183. Joehanes, R. *et al.* Epigenetic Signatures of Cigarette Smoking. *Circulation: Cardiovascular Genetics* **9**, 436–447 (2016).

184. Dugué, P.-A. *et al.* Smoking and blood DNA methylation: an epigenome-wide association study and assessment of reversibility. *Epigenetics* **15**, 358–368 (2020).

185. Sikdar, S. *et al.* Comparison of smoking-related DNA methylation between newborns from prenatal exposure and adults from personal smoking. *Epigenomics* **11**, 1487–1500 (2019).

186. Marzi, S. J. *et al.* Analysis of DNA Methylation in Young People: Limited Evidence for an Association Between Victimization Stress and Epigenetic Variation in Blood. *Am J Psychiatry* **175**, 517–529 (2018).

187. Barcelona, V. *et al.* Novel DNA methylation sites associated with cigarette smoking among African Americans. *Epigenetics* **14**, 383–391 (2019).

188. Dogan, M. V., Beach, S. R. H. & Philibert, R. A. Genetically contextual effects of smoking on genome wide DNA methylation. *Am J Med Genet B Neuropsychiatr Genet* **174**, 595–607 (2017).

189. Domingo-Relloso, A. *et al.* Cadmium, Smoking, and Human Blood DNA Methylation Profiles in Adults from the Strong Heart Study. *Environ Health Perspect* **128**, 067005 (2020).

190. Christiansen, C. *et al.* Novel DNA methylation signatures of tobacco smoking with trans-ethnic effects. *Clin Epigenetics* **13**, 36 (2021).

191. Stevenson, A. J. *et al.* Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. *Clinical Epigenetics* **12**, 113 (2020).

192. Conole, E. L. S. *et al.* DNA Methylation and Protein Markers of Chronic Inflammation and Their Associations With Brain and Cognitive Aging. *Neurology* **97**, e2340–e2352 (2021).

193. Stevenson, A. J. *et al.* Creating and Validating a DNA Methylation-Based Proxy for Interleukin-6. *J Gerontol A Biol Sci Med Sci* **76**, 2284–2292 (2021).

194. Connor Gorber, S., Schofield-Hurwitz, S., Hardt, J., Levasseur, G. & Tremblay, M. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. *Nicotine Tob Res* **11**, 12–24 (2009).

195. Buccafusco, J. J. & Terry, A. V. The potential role of cotinine in the cognitive and neuroprotective actions of nicotine. *Life Sciences* **72**, 2931–2942 (2003).

196. Bollepalli, S., Korhonen, T., Kaprio, J., Anders, S. & Ollikainen, M. EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data. *Epigenomics* **11**, 1469–1486 (2019).

197. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY)* **11**, 303–327 (2019).

198. Thompson, S. G., Stone, R., Nanchahal, K. & Wald, N. J. Relation of urinary cotinine concentrations to cigarette smoking and to exposure to other people's smoke. *Thorax* **45**, 356–361 (1990).

199. Maas, S. C. E. *et al.* Validated inference of smoking habits from blood with a finite DNA methylation marker set. *Eur J Epidemiol* **34**, 1055–1074 (2019).

200. Willinger, C. M. *et al.* A MicroRNA Signature of Cigarette Smoking and Evidence for a Putative Causal Role of MicroRNAs in Smoking-Related Inflammation and Target Organ Damage. *Circ Cardiovasc Genet* **10**, e001678 (2017).

201. Morales, N. A. *et al.* Accuracy of self-reported tobacco use in newly diagnosed cancer patients. *Cancer Causes Control* **24**, 1223–1230 (2013).

202. Kent, B. A. *et al.* Circadian lipid and hepatic protein rhythms shift with a phase response curve different than melatonin. *Nat Commun* **13**, 681 (2022).

203. Chen, T. *et al.* Long-term C-Reactive Protein Variability and Prediction of Metabolic Risk. *Am J Med* **122**, 53–61 (2009).

204. Sherwani, S. I., Khan, H. A., Ekhzaimy, A., Masood, A. & Sakharkar, M. K. Significance of HbA1c Test in Diagnosis and Prognosis of Diabetic Patients. *Biomark Insights* **11**, 95–104 (2016).

205. Ligthart, S. *et al.* DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome Biol* **17**, 255 (2016).

206. Wielscher, M. *et al.* DNA methylation signature of chronic low-grade inflammation and its role in cardio-respiratory diseases. *Nat Commun* **13**, 2408 (2022).

207. Hillary, R. F. *et al.* Blood-based epigenome-wide analyses of chronic low-grade inflammation across diverse population cohorts. *Cell Genomics* **4**, (2024).

208. Mckinnon, K. *et al.* Epigenetic scores derived in saliva are associated with gestational age at birth. *Clinical Epigenetics* **16**, 84 (2024).

209. Gadd, D. A. *et al.* DNAm scores for serum GDF15 and NT-proBNP levels associate with a range of traits affecting the body and brain. *Clinical Epigenetics* **16**, 124 (2024).

210. Smith, H. M. *et al.* Epigenetic scores of blood-based proteins as biomarkers of general cognitive function and brain health. *Clinical Epigenetics* **16**, 46 (2024).

211. Waterfield, S., Yousefi, P. & Suderman, M. DNA methylation models of protein abundance across the lifecourse. *Clin Epigenetics* **16**, 189 (2024).

212. Amaral, P. *et al.* The status of the human gene catalogue. *Nature* **622**, 41–47 (2023).

213. Deutsch, E. W. *et al.* Advances and Utility of the Human Plasma Proteome. *J Proteome Res* **20**, 5241–5263 (2021).

214. Jaros, J. A. J., Guest, P. C., Bahn, S. & Martins-de-Souza, D. Affinity Depletion of Plasma and Serum for Mass Spectrometry-Based Proteome Analysis. in *Proteomics for Biomarker Discovery* (eds. Zhou, M. & Veenstra, T.) 1–11 (Humana Press, Totowa, NJ, 2013). doi:10.1007/978-1-62703-360-2_1.

215. Kaur, G. *et al.* Extending the Depth of Human Plasma Proteome Coverage Using Simple Fractionation Techniques. *J Proteome Res* **20**, 1261–1279 (2021).

216. Lam, M. P. Y., Ping, P. & Murphy, E. Proteomics Research in Cardiovascular Medicine and Biomarker Discovery. *J Am Coll Cardiol* **68**, 2819–2830 (2016).

217. Lequin, R. M. Enzyme Immunoassay (EIA)/Enzyme-Linked Immunosorbent Assay (ELISA). *Clinical Chemistry* **51**, 2415–2418 (2005).

218. Wik, L. *et al.* Proximity Extension Assay in Combination with Next-Generation Sequencing for High-throughput Proteome-wide Analysis. *Mol Cell Proteomics* **20**, 100168 (2021).

219. Norman, M. *et al.* Toward Identification of Markers for Brain-Derived Extracellular Vesicles in Cerebrospinal Fluid: A Large-Scale, Unbiased Analysis Using Proximity Extension Assays. *J Extracell Vesicles* **14**, e70052 (2025).

220. Olink. Olink Explore HT. *Olink Explore HT* https://olink.com/products/olink-explore-ht (2025).

221. Gold, L. *et al.* Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. *PLOS ONE* **5**, e15004 (2010).

222. Somalogic. *SomaScan® Assay v4.1 - Technical Note*. https://somalogic.com/wp-content/uploads/2023/03/SomaScan-Assay-v4.1-Technical-Note.pdf (2023).

223. Somalogic. *SomaScan® 11K Assay v5.0 Technical Note*. https://somalogic.com/wp-content/uploads/2023/12/SL00000919-Rev-1-2023-12-SomaScan-11K-Assay-v5.0-1.pdf (2023).

224. Qian, W.-J., Jacobs, J. M., Liu, T., Camp, D. G. & Smith, R. D. Advances and Challenges in Liquid Chromatography-Mass Spectrometry-based Proteomics Profiling for Clinical Applications*. *Molecular & Cellular Proteomics* **5**, 1727–1744 (2006).

225. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).

226. TARIQ, M. U. *et al.* Methods for Proteogenomics Data Analysis, Challenges, and Scalability Bottlenecks: A Survey. *IEEE Access* **9**, 5497–5516 (2021).

227. Finehout, E. J. & Lee, K. H. An introduction to mass spectrometry applications in biological research. *Biochemistry and Molecular Biology Education* **32**, 93–100 (2004).

228. Domon, B. & Aebersold, R. Mass Spectrometry and Protein Analysis. *Science* **312**, 212–217 (2006).

229. Tu, C. *et al.* Depletion of Abundant Plasma Proteins and Limitations of Plasma Proteomics. *J Proteome Res* **9**, 4982–4991 (2010).

230. Gadd, D. A. *et al.* Blood protein assessment of leading incident diseases and mortality in the UK Biobank. *Nat Aging* **4**, 939–948 (2024).

231. Royer, P. *et al.* Large-scale plasma proteomics in the UK Biobank modestly improves prediction of major cardiovascular events in a population without previous cardiovascular disease. *Eur. J. Prev. Cardiol.* **31**, 1681–1689 (2024).

232. Lind, L., Mazidi, M., Clarke, R., Bennett, D. A. & Zheng, R. Measured and genetically predicted protein levels and cardiovascular diseases in UK Biobank and China Kadoorie Biobank. *Nat Cardiovasc Res* **3**, 1189–1198 (2024).

233. Mazidi, M. *et al.* Risk prediction of ischemic heart disease using plasma proteomics, conventional risk factors and polygenic scores in Chinese and European adults. *Eur J Epidemiol* **39**, 1229–1240 (2024).

234. Corlin, L. *et al.* Proteomic Signatures of Lifestyle Risk Factors for Cardiovascular Disease: A Cross-Sectional Analysis of the Plasma Proteome in the Framingham Heart Study. *J Am Heart Assoc* **10**, e018020 (2021).

235. Kuku, K. O. *et al.* Development and Validation of a Protein Risk Score for Mortality in Heart Failure: A Community Cohort Study. *Ann Intern Med* **177**, 39–49 (2024).

236. Girerd, N. *et al.* Protein Biomarkers of New-Onset Heart Failure: Insights From the Heart Omics and Ageing Cohort, the Atherosclerosis Risk in Communities Study, and the Framingham Heart Study. *Circ Heart Fail* **16**, e009694 (2023).

237. Florijn, B. W. *et al.* Non-coding RNAs versus protein biomarkers to diagnose and differentiate acute stroke: Systematic review and meta-analysis. *Journal of Stroke and Cerebrovascular Diseases* **32**, (2023).

238. Mirza, S. P. Quantitative Mass Spectrometry-Based Approaches in Cardiovascular Research. *Circulation: Cardiovascular Genetics* **5**, 477–477 (2012).

239. Messner, C. B. *et al.* Ultra-High-Throughput Clinical Proteomics Reveals Classifiers of COVID-19 Infection. *Cell Syst* **11**, 11-24.e4 (2020).

240. Wang, J. *et al.* Novel biomarkers for cardiovascular risk prediction. *J Geriatr Cardiol* **14**, 135–150 (2017).

241. Stubbs, P. *et al.* Lipoprotein(a) as a risk predictor for cardiac mortality in patients with acute coronary syndromes. *Eur Heart J* **19**, 1355–1364 (1998).

242. McQueen, M. J. *et al.* Lipids, lipoproteins, and apolipoproteins as risk markers of myocardial infarction in 52 countries (the INTERHEART study): a case-control study. *Lancet* **372**, 224–233 (2008).

243. Emerging Risk Factors Collaboration *et al.* C-reactive protein, fibrinogen, and cardiovascular disease prediction. *N Engl J Med* **367**, 1310–1320 (2012).

244. Puleo, P. R. *et al.* Early diagnosis of acute myocardial infarction based on assay for subforms of creatine kinase-MB. *Circulation* **82**, 759–764 (1990).

245. Shlipak, M. G. *et al.* Cystatin C and the risk of death and cardiovascular events among elderly persons. *N Engl J Med* **352**, 2049–2060 (2005).

246. Adam, S. S., Key, N. S. & Greenberg, C. S. D-dimer antigen: current concepts and future prospects. *Blood* **113**, 2878–2887 (2009).

247. Daniels, L. B. *et al.* Lipoprotein-Associated Phospholipase A2 Is an Independent Predictor of Incident Coronary Heart Disease in an Apparently Healthy Older Population. *JACC* **51**, 913–919 (2008).

248. Brennan, M.-L. *et al.* Prognostic value of myeloperoxidase in patients with chest pain. *N Engl J Med* **349**, 1595–1604 (2003).

249. Kavsak, P. A. *et al.* Effects of contemporary troponin assay sensitivity on the utility of the early markers myoglobin and CKMB isoforms in evaluating patients with possible acute myocardial infarction. *Clin Chim Acta* **380**, 213–216 (2007).

250. de Lemos, J. A., McGuire, D. K. & Drazner, M. H. B-type natriuretic peptide in cardiovascular disease. *Lancet* **362**, 316–322 (2003).

251. Johnson, B. D. *et al.* Serum amyloid A as a predictor of coronary artery disease and cardiovascular outcome in women: the National Heart, Lung, and Blood Institute-Sponsored Women's Ischemia Syndrome Evaluation (WISE). *Circulation* **109**, 726–732 (2004).

252. Adams, J. E. *et al.* Cardiac troponin I. A marker with high specificity for cardiac injury. *Circulation* **88**, 101–106 (1993).

253. Reichlin, T. *et al.* Early diagnosis of myocardial infarction with sensitive cardiac troponin assays. *N Engl J Med* **361**, 858–867 (2009).

254. Gami, B. N. *et al.* Utility of Heart-type Fatty Acid Binding Protein as a New Biochemical Marker for the Early Diagnosis of Acute Coronary Syndrome. *J Clin Diagn Res* **9**, BC22-24 (2015).

255. McMahon, C. G. *et al.* Diagnostic accuracy of heart-type fatty acid-binding protein for the early diagnosis of acute myocardial infarction. *Am J Emerg Med* **30**, 267–274 (2012).

256. Kabekkodu, S. P., Mananje, S. R. & Saya, R. P. A Study on the Role of Heart Type Fatty Acid Binding Protein in the Diagnosis of Acute Myocardial Infarction. *J Clin Diagn Res* **10**, OC07-10 (2016).

257. Cotter, G. *et al.* Growth differentiation factor 15 (GDF-15) in patients admitted for acute heart failure: results from the RELAX-AHF study. *Eur J Heart Fail* **17**, 1133–1143 (2015).

258. Wollert, K. C. *et al.* Growth differentiation factor 15 for risk stratification and selection of an invasive treatment strategy in non ST-elevation acute coronary syndrome. *Circulation* **116**, 1540–1548 (2007).

259. Wollert, K. C. *et al.* Prognostic value of growth-differentiation factor-15 in patients with non-ST-elevation acute coronary syndrome. *Circulation* **115**, 962–971 (2007).

260. Bonaca, M. P. *et al.* Prospective evaluation of pregnancy-associated plasma protein-a and outcomes in patients with acute coronary syndromes. *J Am Coll Cardiol* **60**, 332–338 (2012).

261. Bayes-Genis, A. *et al.* Pregnancy-associated plasma protein A as a marker of acute coronary syndromes. *N Engl J Med* **345**, 1022–1029 (2001).

262. Dhillon, O. S. *et al.* Matrix metalloproteinase-2 predicts mortality in patients with acute coronary syndrome. *Clin Sci (Lond)* **118**, 249–257 (2009).

263. Kelly, D. *et al.* Plasma matrix metalloproteinase-9 and left ventricular remodelling after acute myocardial infarction in man: a prospective cohort study. *Eur Heart J* **28**, 711–718 (2007).

264. Wang, L.-X. *et al.* Comparision of high sensitivity C-reactive protein and matrix metalloproteinase 9 in patients with unstable angina between with and

without significant coronary artery plaques. *Chin Med J (Engl)* **124**, 1657–1661 (2011).

265. Mallat, Z. *et al.* Circulating secretory phospholipase A2 activity and risk of incident coronary events in healthy men and women: the EPIC-Norfolk study. *Arterioscler Thromb Vasc Biol* **27**, 1177–1183 (2007).

266. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat Genet* **47**, 856–860 (2015).

267. Liebetrau, C. *et al.* Release kinetics of inflammatory biomarkers in a clinical model of acute myocardial infarction. *Circ Res* **116**, 867–875 (2015).

268. Li, J. *et al.* Soluble CD40L Is a Useful Marker to Predict Future Strokes in Patients With Minor Stroke and Transient Ischemic Attack. *Stroke* **46**, 1990–1992 (2015).

269. Schönbeck, U., Varo, N., Libby, P., Buring, J. & Ridker, P. M. Soluble CD40L and cardiovascular risk in women. *Circulation* **104**, 2266–2268 (2001).

270. Möckel, M. & Searle, J. Copeptin—Marker of Acute Myocardial Infarction. *Curr Atheroscler Rep* **16**, 421 (2014).

271. Greisenegger, S. *et al.* Copeptin and Long-Term Risk of Recurrent Vascular Events After Transient Ischemic Attack and Ischemic Stroke: Population-Based Study. *Stroke* **46**, 3117–3123 (2015).

272. Boeckel, J.-N. *et al.* Analyzing the Release of Copeptin from the Heart in Acute Myocardial Infarction Using a Transcoronary Gradient Model. *Sci Rep* **6**, 20812 (2016).

273. Neumann, J. T. *et al.* Association of MR-proadrenomedullin with cardiovascular risk factors and subclinical cardiovascular disease. *Atherosclerosis* **228**, 451–459 (2013).

274. Klip, I. T. *et al.* Prognostic value of mid-regional pro-adrenomedullin in patients with heart failure after an acute myocardial infarction. *Heart* **97**, 892–898 (2011).

275. Gottsäter, M. *et al.* Adrenomedullin is a marker of carotid plaques and intima-media thickness as well as brachial pulse pressure. *J Hypertens* **31**, 1959–1965 (2013).

276. Dhillon, O. S. *et al.* Pre-discharge risk stratification in unselected STEMI: is there a role for ST2 or its natural ligand IL-33 when compared with contemporary risk markers? *Int J Cardiol* **167**, 2182–2188 (2013).

277. Demyanets, S. *et al.* Soluble ST2 and interleukin-33 levels in coronary artery disease: relation to disease activity and adverse outcome. *PLoS One* **9**, e95055 (2014).

278. Manzano-Fernández, S., Mueller, T., Pascual-Figal, D., Truong, Q. A. & Januzzi, J. L. Usefulness of soluble concentrations of interleukin family member ST2 as predictor of mortality in patients with acutely decompensated heart failure relative to left ventricular ejection fraction. *Am J Cardiol* **107**, 259–267 (2011).

279. Ky, B. *et al.* High-sensitivity ST2 for prediction of adverse outcomes in chronic heart failure. *Circ Heart Fail* **4**, 180–187 (2011).

280. Sabatine, M. S. *et al.* Evaluation of multiple biomarkers of cardiovascular stress for risk prediction and guiding medical therapy in patients with stable coronary disease. *Circulation* **125**, 233–240 (2012).

281. Khan, S. Q. *et al.* C-terminal pro-endothelin-1 offers additional prognostic information in patients after acute myocardial infarction: Leicester Acute Myocardial Infarction Peptide (LAMP) Study. *Am Heart J* **154**, 736–742 (2007).

282. Perez, A. L. *et al.* Increased mortality with elevated plasma endothelin-1 in acute heart failure: an ASCEND-HF biomarker substudy. *Eur J Heart Fail* **18**, 290–297 (2016).

283. Chen, Y.-S. *et al.* Using the galectin-3 test to predict mortality in heart failure patients: a systematic review and meta-analysis. *Biomark Med* **10**, 329–342 (2016).

284. Lemmens, K., Doggen, K. & De Keulenaer, G. W. Role of neuregulin-1/ErbB signaling in cardiovascular physiology and disease: implications for therapy of heart failure. *Circulation* **116**, 954–960 (2007).

285. Geisberg, C. A. *et al.* Circulating Neuregulin-1β Levels Vary According to the Angiographic Severity of Coronary Artery Disease and Ischemia. *Coron Artery Dis* **22**, 577–582 (2011).

286. Puspitasari, V., Gunawan, P. Y., Wiradarma, H. D. & Hartoyo, V. Glial Fibrillary Acidic Protein Serum Level as a Predictor of Clinical Outcome in Ischemic Stroke. *Open Access Maced J Med Sci* **7**, 1471–1474 (2019).

287. Kumar, A. *et al.* Role of glial fibrillary acidic protein as a biomarker in differentiating intracerebral haemorrhage from ischaemic stroke and stroke mimics: a meta-analysis. *Biomarkers* **25**, 1–8 (2020).

288. Anogianakis, G. *et al.* Current Trends in Stroke Biomarkers: The Prognostic Role of S100 Calcium-Binding Protein B and Glial Fibrillary Acidic Protein. *Life (Basel)* **14**, 1247 (2024).

289. Rossi, R. *et al.* S100b in acute ischemic stroke clots is a biomarker for post-thrombectomy intracranial hemorrhages. *Front Neurol* **13**, 1067215 (2023).

290. Poislane, P.-A. *et al.* Diagnostic performance of S100B assay for intracranial hemorrhage detection in patients with mild traumatic brain injury under antiplatelet or anticoagulant therapy. *Sci Rep* **15**, 5741 (2025).

291. Smith, B. H. *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* **42**, 689–700 (2013).

292. Smith, B. H. *et al.* Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med Genet* **7**, 74 (2006).

293. Milbourn, H. *et al.* Generation Scotland: an update on Scotland's longitudinal family health study. *BMJ Open* **14**, e084719 (2024).

294. Nagy, R. *et al.* Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants. *Genome Medicine* **9**, 23 (2017).

295. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007).

296. Walker, R. M. *et al.* Data Resource Profile: Whole-Blood DNA Methylation Resource in Generation Scotland (MeGS). *Int J Epidemiol* **54**, dyaf091 (2025).

297. Navrady, L. B. *et al.* Cohort Profile: Stratifying Resilience and Depression Longitudinally (STRADL): a questionnaire follow-up of Generation Scotland: Scottish Family Health Study (GS:SFHS). *Int J Epidemiol* **47**, 13–14g (2018).

298. Fortin, J.-P., Fertig, E. & Hansen, K. shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R. *F1000Res* **3**, 175 (2014).

299. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14**, 293 (2013).

300. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).

301. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).

302. Min, J. L., Hemani, G., Davey Smith, G., Relton, C. & Suderman, M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics* **34**, 3983–3989 (2018).

303. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res* **45**, e22 (2017).

304. McCartney, D. L. *et al.* Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. *Genomics Data* **9**, 22–24 (2016).

305. Chybowska, A. D. *et al.* A blood- and brain-based EWAS of smoking. *Nat Commun* **16**, 3210 (2025).

306. Ho, D., Imai, K., King, G. & Stuart, E. A. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* **42**, 1–28 (2011).

307. TWIST Bioscience. *Twist Targeted Methylation Sequencing Protocol*. https://www.twistbioscience.com/resources/protocol/twist-targeted-methylation-sequencing-protocol (2023).

308. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* **35**, 316–319 (2017).

309. Ewels, P. *et al.* nf-core/methylseq: nf-core/methylseq version 1.6.1 [Nauseous Serpent]. Zenodo https://doi.org/10.5281/zenodo.4744708 (2021).

310. Mayakonda, A. *et al.* Methrix: an R/Bioconductor package for systematic aggregation and analysis of bisulfite sequencing data. *Bioinformatics* **36**, 5524–5525 (2020).

311. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).

312. Vernardis, S. I. *et al.* The Impact of Acute Nutritional Interventions on the Plasma Proteome. *The Journal of Clinical Endocrinology and Metabolism* **108**, 2087 (2023).

313. Messner, C. B. *et al.* Ultra-fast proteomics with Scanning SWATH. *Nat Biotechnol* **39**, 846–854 (2021).

314. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* **17**, 41–44 (2020).

315. Bruderer, R. *et al.* Analysis of 1508 Plasma Samples by Capillary-Flow Data-Independent Acquisition Profiles Proteomics of Weight Loss and Maintenance. *Mol Cell Proteomics* **18**, 1242–1254 (2019).

316. Rusilowicz, M., Dickinson, M., Charlton, A., O'Keefe, S. & Wilson, J. A batch correction method for liquid chromatography-mass spectrometry data that does not depend on quality control samples. *Metabolomics* **12**, 56 (2016).

317. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).

318. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158–D169 (2017).

319. Deary, I. J., Gow, A. J., Pattie, A. & Starr, J. M. Cohort profile: the Lothian Birth Cohorts of 1921 and 1936. *Int J Epidemiol* **41**, 1576–1584 (2012).

320. Taylor, A. M., Pattie, A. & Deary, I. J. Cohort Profile Update: The Lothian Birth Cohorts of 1921 and 1936. *International Journal of Epidemiology* **47**, 1042–1042r (2018).

321. Corley, J. *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* **9**, 248 (2019).

322. Shah, S. *et al.* Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Res* **24**, 1725–1733 (2014).

323. Stevenson, A. J. *et al.* A comparison of blood and brain-derived ageing and inflammation-related DNA methylation signatures and their association with microglial burdens. *Eur J Neurosci* **56**, 5637–5649 (2022).

324. Harris, P. A. *et al.* Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* **42**, 377–381 (2009).

325. Relton, C. L. *et al.* Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *International Journal of Epidemiology* **44**, 1181–1190 (2015).

326. Zou, H. & Hastie, T. Regularization and Variable Selection Via the Elastic Net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67**, 301–320 (2005).

327. Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 187–202 (1972).

328. Therneau, T. M. & Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model*. (Springer, New York, 2000).

329. Therneau, T. M. *Coxme: Mixed Effects Cox Models*. (2024).

330. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**, 1–22 (2010).

331. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).

332. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *Ann. Appl. Stat.* **2**, (2008).

333. Dunn, O. J. Multiple Comparisons Among Means. *Journal of the American Statistical Association* **56**, 52–64 (1961).

334. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995).

335. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53**, 457–481 (1958).

336. Smith, H. M. *et al.* DNA methylation-based predictors of metabolic traits in Scottish and Singaporean cohorts. *Am J Hum Genet* **112**, 106–115 (2025).

337. Flynn, R. *et al.* Evaluation of nanopore sequencing for epigenetic epidemiology: a comparison with DNA methylation microarrays. *Hum Mol Genet* **31**, 3181–3190 (2022).

338. Sigurpalsdottir, B. D. *et al.* A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes. *Genome Biology* **25**, 69 (2024).

339. Goldsmith, C. *et al.* Low biological fluctuation of mitochondrial CpG and non-CpG methylation at the single-molecule level. *Sci Rep* **11**, 8032 (2021).

340. Inkster, A. M., Wong, M. T., Matthews, A. M., Brown, C. J. & Robinson, W. P. Who's afraid of the X? Incorporating the X and Y chromosomes into the analysis of DNA methylation array data. *Epigenetics & Chromatin* **16**, 1 (2023).

341. Landmark genetics partnership to probe causes of cancer and dementia. *GOV.UK* https://www.gov.uk/government/news/landmark-genetics-partnership-to-probe-causes-of-cancer-and-dementia.

342. Dhingra, R. & Vasan, R. S. Age as a Cardiovascular Risk Factor. *Med Clin North Am* **96**, 87–91 (2012).

343. Bernabeu, E. *et al.* Refining epigenetic prediction of chronological and biological age. *Genome Medicine* **15**, 12 (2023).

344. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)* **10**, 573–591 (2018).

345. Joyce, B. T. *et al.* Epigenetic Age Acceleration Reflects Long-Term Cardiovascular Health. *Circ Res* **129**, 770–781 (2021).

346. Cheng, Y. *et al.* Development and validation of DNA methylation scores in two European cohorts augment 10-year risk prediction of type 2 diabetes. *Nat Aging* **3**, 450–458 (2023).

347. Robertson, J. A. Board 8038T: Molecular signatures of ambient air pollution exposure: a multivariate and multi-omic analysis in Generation Scotland. (2025).

348. Nguyen, M.-N. *et al.* DNA methylation: a potential mediator between air pollution exposures and asthma control. *Clinical Epigenetics* **17**, 175 (2025).

349. Locke, W. J. *et al.* DNA Methylation Cancer Biomarkers: Translation to the Clinic. *Front Genet* **10**, 1150 (2019).

350. Richmond, A. *et al.* Genome-wide analysis of 439 mass spectrometry-based proteomic profiles in a population of 15,035 Scottish individuals. 2025.08.14.25333677 Preprint at https://doi.org/10.1101/2025.08.14.25333677 (2025).

351. Beck, H. C. *et al.* A Mass Spectrometry-Based Proteome Study of Twin Pairs Discordant for Incident Acute Myocardial Infarction within Three Years after Blood Sampling Suggests Novel Biomarkers. *Int J Mol Sci* **25**, 2638 (2024).

352. LI, Q.-X. *et al.* Association of serum complement C1q with cardiovascular outcomes among patients with acute coronary syndrome undergoing percutaneous coronary intervention. *J Geriatr Cardiol* **19**, 949–959 (2022).

353. Borlak, J., Chatterji, B., Londhe, K. B. & Watkins, P. B. Serum acute phase reactants hallmark healthy individuals at risk for acetaminophen-induced liver injury. *Genome Med* **5**, 86 (2013).

354. Wienke, J. *et al.* Biomarker profiles of endothelial activation and dysfunction in rare systemic autoimmune diseases: implications for cardiovascular risk. *Rheumatology (Oxford)* **60**, 785–801 (2021).

355. Huang, S. *et al.* An overview of the multi-dimensional mechanisms of exercise-regulated hormones and growth factors in cardiac physiological adaptation. *Front Physiol* **16**, 1642389 (2025).

356. Boja, E. S., Fehniger, T. E., Baker, M. S., Marko-Varga, G. & Rodriguez, H. Analytical Validation Considerations of Multiplex Mass-Spectrometry-Based Proteomic Platforms for Measuring Protein Biomarkers. *J Proteome Res* **13**, 5325–5332 (2014).

357. Liang, W., Wang, Q., Ma, H., Yan, W. & Yang, J. Knockout of Low Molecular Weight FGF2 Attenuates Atherosclerosis by Reducing Macrophage Infiltration and Oxidative Stress in Mice. *Cell Physiol Biochem* **45**, 1434–1443 (2018).

358. Langley, S. R. *et al.* Extracellular matrix proteomics identifies molecular signature of symptomatic carotid plaques. *J Clin Invest* **127**, 1546–1560 (2017).

359. Krätschmer, I. *et al.* Discovery of shared epigenetic pathways across human phenotypes. 2024.04.15.589547 Preprint at https://doi.org/10.1101/2024.04.15.589547 (2024).

360. Sissala, N. *et al.* Comparative evaluation of Olink Explore 3072 and mass spectrometry with peptide fractionation for plasma proteomics. *Commun Chem* **8**, 327 (2025).

361. Kirsher, D. Y. *et al.* Current landscape of plasma proteomics from technical innovations to biological insights and biomarker discovery. *Commun Chem* **8**, 279 (2025).

362. Mann, M. & Kelleher, N. L. Precision proteomics: The case for high resolution and high mass accuracy. *Proceedings of the National Academy of Sciences* **105**, 18132–18138 (2008).

363. Han, X., Aslanian, A. & Yates, J. R. Mass Spectrometry for Proteomics. *Curr Opin Chem Biol* **12**, 483–490 (2008).

364. Kok, T. F. *et al.* High-dimensional machine learning models for prediction of heart failure in more than 400 000 men and women from the UK Biobank. *Eur Heart J Digit Health* https://doi.org/10.1093/ehjdh/ztaf118 (2025) doi:10.1093/ehjdh/ztaf118.

365. Krittanawong, C. *et al.* Machine learning prediction in cardiovascular diseases: a meta-analysis. *Sci Rep* **10**, 16057 (2020).

366. Shmatko, A. *et al.* Learning the natural history of human disease with generative transformers. *Nature* 1–9 (2025) doi:10.1038/s41586-025-09529-3.

367. Liu, T., Krentz, A., Lu, L. & Curcin, V. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. *Eur Heart J Digit Health* **6**, 7–22 (2024).

368. Neumann, J. T., de Lemos, J. A., Apple, F. S. & Leong, D. P. Cardiovascular biomarkers for risk stratification in primary prevention. *Eur Heart J* **46**, 3823–3843 (2025).

369. Elliott, P. *et al.* Development, validation, and implementation of biomarker testing in cardiovascular medicine state-of-the-art: proceedings of the European Society of Cardiology—Cardiovascular Round Table. *Cardiovasc Res* **117**, 1248–1256 (2021).

370. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.

371. Illumina Methylation Microarrays | University of Minnesota Genomics Center. https://genomics.umn.edu/service/illumina-methylation-microarrays.

372. Olink. *Olink® Reveal*. https://7074596.fs1.hubspotusercontent-na1.net/hubfs/7074596/000-documents/7-Brochure/1586-Olink-Reveal-Brochure.pdf (2025).

373. Santos Gonzalez, F. *et al.* A micro-costing study of mass-spectrometry based quantitative proteomics testing applied to the diagnostic pipeline of mitochondrial and other rare disorders. *Orphanet Journal of Rare Diseases* **19**, 443 (2024).

374. Bahai, A. *et al.* High-throughput plasma proteomic platforms: Insights from a multi-ethnic Asian cohort. 2025.10.24.682486 Preprint at https://doi.org/10.1101/2025.10.24.682486 (2025).

375. Rooney, M. R. *et al.* Comparison of proteomic measurements across platforms in the Atherosclerosis Risk in Communities (ARIC) Study. *Clin Chem* **69**, 68–79 (2023).

376. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* **186**, 1026–1034 (2017).

377. Simsek, I. *et al.* Participation Bias in a Survey of Community Patients with Heart Failure. *Mayo Clin Proc* **95**, 911–919 (2020).

378. Banack, H. R., Kaufman, J. S., Wactawski-Wende, J., Troen, B. R. & Stovitz, S. D. Investigating and Remediating Selection Bias in Geriatrics Research: The Selection Bias Toolkit. *J Am Geriatr Soc* **67**, 1970–1976 (2019).

379. Althubaiti, A. Information bias in health research: definition, pitfalls, and adjustment methods. *J Multidiscip Healthc* **9**, 211–217 (2016).

380. Jin, S.-G., Kadam, S. & Pfeifer, G. P. Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res* **38**, e125 (2010).

381. Wu, K.-J. The epigenetic roles of DNA N6-Methyladenine (6mA) modification in eukaryotes. *Cancer Lett* **494**, 40–46 (2020).

382. Bizet, M. *et al.* Improving Infinium MethylationEPIC data processing: re-annotation of enhancers and long noncoding RNA genes and benchmarking of normalization methods. *Epigenetics* **17**, 2434–2454 (2022).

383. Lorenzen, J. M., Martino, F. & Thum, T. Epigenetic modifications in cardiovascular disease. *Basic Res Cardiol* **107**, 245 (2012).

384. Zhang, L. *et al.* DNA methylation and histone post-translational modifications in atherosclerosis and a novel perspective for epigenetic therapy. *Cell Communication and Signaling* **21**, 344 (2023).

385. Claassen, M. Inference and Validation of Protein Identifications. *Molecular & Cellular Proteomics* **11**, 1097–1104 (2012).

386. Huang, T., Wang, J., Yu, W. & He, Z. Protein inference: a review. *Brief Bioinform* **13**, 586–614 (2012).

387. Lefebvre, F. & Giorgi, R. A strategy for optimal fitting of multiplicative and additive hazards regression models. *BMC Medical Research Methodology* **21**, 100 (2021).

388. Heinze, G., Wallisch, C. & Dunkler, D. Variable selection – A review and recommendations for the practicing statistician. *Biom J* **60**, 431–449 (2018).

389. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* **23**, R89–R98 (2014).

390. Burgess, S., Woolf, B., Mason, A. M., Ala-Korpela, M. & Gill, D. Addressing the credibility crisis in Mendelian randomization. *BMC Med* **22**, 374 (2024).

391. Ho, J. E. *et al.* Protein Biomarkers of Cardiovascular Disease and Mortality in the Community. *Journal of the American Heart Association* **7**, e008108 (2018).

392. Zhang, J. Biomarkers of endothelial activation and dysfunction in cardiovascular diseases. *Rev Cardiovasc Med* **23**, 73 (2022).

Appendix – Publications

First author publications

Published:

- **Chybowska AD**, Gadd DA, *et al.* Epigenetic Contributions to Clinical Risk Prediction of Cardiovascular Disease, *Circulation: Genomic and Precision Medicine.* 2024
- **Chybowska AD**, Bernabeu E, *et al.* A blood- and brain-based EWAS of smoking, *Nature Communications*. 2025

Submitted:

- **Chybowska AD**, Vernardis S, *et al.* Untargeted Proteomic Profiling Identifies Candidate Biomarkers for Early detection of Cardiovascular Disease Outcomes and Mortality.

Middle author publications:

Published:

- Bernabeu E, **Chybowska AD**, *et al.* Blood-based DNA methylation study of alcohol consumption, *Clinical Epigenetics.* 2025.
- Smith HM, Moodie J, [3 authors], **Chybowska AD**, *et al.* Epigenetic scores of blood-based proteins as biomarkers of general cognitive function and brain health, *Clinical Epigenetics.* 2024.
- Gadd DA, Hannah M. Smith, Donncha Mullin, **Chybowska AD**, *et al.* DNAm scores for serum GDF15 and NT-proBNP levels associate with a range of traits affecting the body and brain, *Clinical Epigenetics.* 2024
- Hillary RF, McCartney DL, [3 authors], **Chybowska AD**, *et al.* Blood-based epigenome-wide analyses of 19 common disease states: A longitudinal, population-based linked cohort study of 18,413 Scottish individuals, *PLOS Medicine.* 2023.

Submitted:

- Richmond A, Robertson JA, [4 authors], **Chybowska AD**, *et al.* Genome-wide analysis of 439 mass spectrometry-based proteomic profiles in a population of 15,035 Scottish individuals.
- Mur J, **Chybowska AD**, *et al.* Risk factors for dementia reflected in the serum proteome: a study using mass spectrometry data in Generation Scotland.
- Robertson JA, Bakzik J, Vernardis S, **Chybowska AD**, *et al.* Methylome-wide association studies and epigenetic biomarker development for 133 mass spectrometry-assessed circulating proteins in 14,761 Generation Scotland participants.
- Smith HM, Moodie JE, [4 authors], **Chybowska AD**, *et al.* Proteomic biomarkers of cognitive function and dementia in Generation Scotland.