Epigenetic scores for the circulating proteome replicate protein-disease predictions as tools for biomarker discovery

Danni A Gadd[1,*], Robert F Hillary[1,*], Daniel L McCartney[1,*], Shaza B Zaghlool[2,3*,], Anna J Stevenson[1], Cliff Nangle[1], Archie Campbell[1], Robin Flaig[1], Sarah E Harris[4,5], Rosie M Walker[6], Liu Shi[7], Elliot M Tucker-Drob[8,9], Christian Gieger[10,11,12,13], Annette Peters[11,12,13], Melanie Waldenberger[10,11,12], Johannes Graumann[14,15], Allan F McRae[16], Ian J Deary[4,5], David J Porteous[1], Caroline Hayward[1,17], Peter M Visscher[16], Simon R Cox[4,5], Kathryn L Evans[1], Andrew M McIntosh[1,18], Karsten Suhre[2], Riccardo E Marioni[1,†]

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

[2] Department of Physiology and Biophysics, Weill Cornell Medicine-Qatar, Education City, PO 24144 Doha, Qatar

[3] Computer Engineering Department, Virginia Tech, Blacksburg, VA, USA.

[4] Lothian Birth Cohorts, University of Edinburgh, Edinburgh, EH8 9JZ

[5] Department of Psychology, University of Edinburgh, Edinburgh, EH8 9JZ

[6] Centre for Clinical Brain Sciences, Chancellor's Building, 49 Little France Crescent, Edinburgh BioQuarter, Edinburgh, EH16 4SB

[7] Department of Psychiatry, University of Oxford, UK

[8] Department of Psychology, The University of Texas at Austin, United States

[9] Population Research Center, The University of Texas at Austin, United States

[10] Research Unit Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Bavaria, Germany

[11] Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Bavaria, Germany

[12] German Centre for Cardiovascular Research (DZHK), partner site Munich Heart Alliance, Munich, Germany

[13] German Center for Diabetes Research (DZD), Neuherberg, Germany

[14] Scientific Service Group Biomolecular Mass Spectrometry, Max Planck Institute for Heart and Lung Research, W.G. Kerckhoff Institute, Bad Nauheim, Germany

[15] German Centre for Cardiovascular Research (DZHK), Partner Site Rhine-Main, Max Planck Institute of Heart and Lung Research, Bad Nauheim, Germany

[16] Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD, Australia

[17] Medical Research Council Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, EH4 2XU

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

* These authors contributed equally

[†] Corresponding author: Riccardo Marioni, riccardo.marioni@ed.ac.uk

Abstract

Polygenic scores capture the genetic predisposition to complex disease and are powerful tools for risk stratification. However, such diseases are also largely influenced by environmental insults and lifestyle factors, which are thought to be captured by epigenetic modifications. DNA methylation encodes information on the epigenetic landscape of an individual, regulates gene transcription, and can be reflected in the variability of the proteome. Protein biomarkers that predict or correlate with disease onset have been identified across many age-related morbidities; however, the extent of epigenetic influence in these pathways is often unclear. Here, we report a comprehensive epigenome-wide study of the circulating proteome. Using data from four independent cohort studies, we train and test epigenetic scores (EpiScores) for 953 plasma proteins, identifying 109 robust signatures that explain between 1% and 58% of the variance after adjusting for genetic effects. These EpiScores are then linked to the incidence of 12 leading morbidities over 14 years of follow up in the Generation Scotland cohort (n=9,537). After adjustment for common risk factors, 137 EpiScore - incident disease associations are revealed, several of which are confirmed by previously identified associations with measured protein levels. Notably, 21 EpiScores that associate with incident diabetes replicate previous protein associations with incident (n=12 proteins) and prevalent (n=21 proteins) diabetes. Complementary to polygenic scores, these epigenetic scores for protein levels are therefore likely to capture the environmental components of disease risk and can be a valuable resource for disease prediction, stratification and biomarker discovery.

1 Introduction

Chronic morbidities place longstanding burdens on our health as we age. Stratifying an individual's risk prior to overt symptom presentation is therefore critical [1]. Complex diseases, such as type 2 diabetes, cancer, cardiovascular disease and Alzheimer's disease, are known to be partially driven by genetic factors that underpin the risk of onset [2–4]. Polygenic risk scores (PRS) can be useful tools to stratify disease risk [5–7]. However, PRS do not assess the environmental part of this risk. Epigenetic modifications are known to associate with complex disease [8,9]. DNA methylation can reflect genetic and phenotypic exposures and, in many cases, may also regulate gene expression in response to physiological stress [10]. Blood-based methylation signatures have been shown to strongly predict all-cause mortality and epigenetic scores have been developed as tools to predict smoking status and other lifestyle-related exposures [11–13].

Advances in genetic and epigenetic studies of the plasma proteome have uncovered pathways through which DNA methylation may influence both protein expression and disease outcomes [4,10,13–17]. Low-grade inflammation, which is thought to exacerbate many age-related morbidities, is particularly well-captured through DNAm studies of the plasma proteome [18]. As proteins are important intermediate phenotypes connecting the genome, epigenome and onset of disease, understanding the molecular regulation of protein markers for disease may help to target interventions.

Epigenetic predictors have utilised DNAm from the blood to estimate a person's 'biological age' [19], measure their exposure to lifestyle and environmental exposures [12,20,21] and predict circulating levels of inflammatory proteins [22,23]. A leading epigenetic predictor of biological aging, the GrimAge epigenetic clock, incorporates methylation scores for seven proteins along with smoking and chronological age, and is associated with numerous incident disease outcomes [19,24]. However, focusing on measurements of individual proteins, instead of composite measures may offer greater insights into specific DNAm-protein pathways pertinent in the earliest stages of disease pathogenesis. DNAm scores for Interleukin-6 and C-Reactive protein associate with a range of phenotypes independently of measured protein levels, show more stable longitudinal trajectories than repeated protein measurements, and, in some cases, outperform blood-based proteomic associations with brain morphology [23,25].

Here, we report a comprehensive association study of blood-based DNAm with proteomics and disease (**Figure 1**). We train epigenetic scores – referred to as EpiScores – for 953 plasma proteins (with sample size ranging from 725 – 944 individuals) and validate them using two independent cohorts with 778 and 162 participants; Demographic information for all cohorts is described in **Supplementary Table 1**. We regress out genetic effects (protein quantitative trait loci; pQTLs) from the protein levels prior to generating the EpiScores so that they reflect environmental and lifestyle contributions to individual differences. Finally, we examine whether the most robust predictors (n=109 EpiScores) associate with the incidence of 12 major morbidities, over a follow up period of up to 14 years in the Generation Scotland cohort (n = 9,537). We control for common risk factors for disease and assess the replication of EpiScore-disease relationships against previously reported protein-disease associations.

2 Results

2.1 Selecting the most robust EpiScores for protein levels

To generate epigenetic scores for a comprehensive set of plasma proteins, we ran elastic net penalised regression models using protein measurements from the SOMAscan (aptamer-based) and Olink (antibody-based) platforms. We used two cohorts: the German population study KORA (n=944, with a subset of 793 SOMAscan proteins present across train and test cohorts) and the Lothian Birth Cohort 1936 (LBC1936) study (between 725 and 875 individuals in the training cohort, with a total of 160 Olink neurology and inflammatory panel proteins). Demographic information is available for all cohorts in **Supplementary Table 1**.

Prior to running the elastic net models, we rank-based inverse normalised protein levels and adjusted for age, sex, cohort-specific variables and, where present, *cis* and *trans* pQTL effects identified from previous analyses [17,26,27] (**Methods**). Of a possible 793 proteins in KORA, 84 EpiScores had Pearson $r > 0.1$ and $P < 0.05$ when tested in an independent subset of Generation Scotland (The Stratifying Resilience and Depression Longitudinally:STRADL study, n=778) (**Supplementary Table 2**). We selected these for EpiScore-disease analyses. Of the 160 Olink proteins trained in LBC1936, there were 21 with $r > 0.1$ and $P < 0.05$ in independent test sets (STRADL, n=778, Lothian Birth Cohort 1921: LBC1921, n=162) (**Supplementary Table 3**). Independent test set data were not available for four Olink proteins. However, they were included based on their performance ($r > 0.1$ and $P < 0.05$) in a holdout sample of 150 individuals who were left out of the training set. We then retrained these four predictors on the full training sample. A total of 109 EpiScores (84 SOMAscan-based and 25 Olink-based) were brought forward to EpiScore-disease analyses (**Supplementary Table 4)**. Predictor weights, positional information and *cis/trans* status for CpG sites contributing to these EpiScores are available in **Supplementary Table 5**. The number of CpG features selected for EpiScores ranged from one (Lyzozyme) to 395 (Aminoacylase-1), with a mean of 96 (**Supplementary Table 6**). The most frequently selected CpG was the smoking-related site cg05575921 (mapping to the *AHRR* gene), which was included in 25 EpiScores. Counts for each CpG site are summarised in **Supplementary Table 7.** This table includes the set of protein EpiScores that each CpG contributes to, along with phenotypic annotations (traits) from the MRC-IEU EWAS catalog [28] for each CpG site having genome-wide significance ($P < 3.6 \times 10^{-8}$) [29].

2.2 EpiScore-disease associations in Generation Scotland

The Generation Scotland dataset contains extensive electronic health data from GP and hospital records available as well as DNA methylation data for 9,537 individuals. This makes it uniquely positioned to test whether EpiScore signals can predict disease onset. As the 109 EpiScores were trained on protein levels after adjustment for known pQTLs, they are likely to reflect only non-genetic contributions to the proteome. We ran nested mixed effects Cox proportional hazards models (**Figure 2**) to determine whether the levels of each EpiScore at baseline associated with the incidence of 12 morbidities over a maximum of 14 years of follow up. The correlation structures for the 109 EpiScore measures used for Cox modelling are presented in **Supplementary Figure 1**.

The Cox proportional hazard assumption dictates that hazard ratios for EpiScore – disease associations should remain constant over time. We correlated the Schoenfeld residuals from the models with time to test this. Two associations in the basic model adjusting for age and sex failed to satisfy the global assumption (across all covariates) and were excluded. There were 294 remaining EpiScore-disease associations with a False Discovery Rate (FDR)-adjusted $P < 0.05$ in the basic model. After further adjustment for common risk factor covariates (smoking, social deprivation status, educational attainment, body mass index (BMI) and alcohol consumption), 137 of the 294 EpiScore-disease associations from the basic model had $P < 0.05$ in the fully-adjusted model (**Supplementary Tables 8-9**). Eleven of the 137 fully-adjusted associations failed the Cox proportional hazards assumption for the EpiScore variable ($P < 0.05$ for the association between the Schoenfeld residuals and time; **Supplementary Table 10**). When we restricted the time-to-event/censor period by possible years of follow-up, there were minimal differences in the EpiScore - disease hazard ratios between follow-up periods that did not violate the assumption and those that did (**Supplementary Table 11**). The 137 associations were therefore retained as the primary results.

The 137 associations found in the fully-adjusted model comprised 78 unique EpiScores that were related to the incidence of 11 of the 12 morbidities studied. Diabetes and chronic obstructive pulmonary disease (COPD) had the greatest number of associations, with 33 and 41, respectively. **Figure 3** presents the EpiScore-disease relationships for COPD and the remaining nine morbidities: stroke, lung cancer, ischaemic heart disease, inflammatory bowel disease, rheumatoid arthritis, depression, bowel cancer, pain and Alzheimer's dementia. Of the EpiScore signatures contributing to each trait, there were between one and four independent signals, as determined by components with eigenvalues > 1 in principal components analyses **(Supplementary Figures 2-4)**. There were 13 EpiScores that associated with the onset of three or more morbidities. **Figure 4** presents relationships

for these EpiScores in the fully-adjusted Cox model results. Of note is the EpiScore for Complement 5 (C5), which associated with five outcomes: stroke, diabetes, ischaemic heart disease, rheumatoid arthritis and COPD.

Of the 29 SOMAscan-derived EpiScore associations with incident diabetes, 21 replicated previously reported protein associations [30,31] with incident or prevalent diabetes in one or more cohorts (**Figure 5 and Supplementary Table 12**).

2.3 Immune cell and GrimAge sensitivity analyses

Correlations of the 109 EpiScores with covariates suggested interlinked relationships with both immune cells and GrimAge acceleration (**Supplementary Figure 5)**. These covariates were therefore added incrementally to the fully-adjusted Cox models (**Figure 2**). There were 117 associations that remained statistically significant (FDR $P < 0.05$ in the basic model and $P < 0.05$ in the fully-adjusted model) after adjustment for immune cell proportions, of which 89 remained significant when GrimAge acceleration scores were added to this model **(Supplementary Table 9 and Supplementary Figure 6)**. In a further sensitivity analysis, relationships between estimated white blood cell proportions and incident diseases were assessed in the Cox model structure independently of EpiScores. Of the 60 possible relationships between WBC measures and the morbidities assessed, four were statistically significant (FDR-adjusted $P < 0.05$) in the basic model and remained significant with $P < 0.05$ in the fully-adjusted model **(Supplementary Table 13**). A higher proportion of Natural Killer cells was linked to decreased risk of incident COPD, rheumatoid arthritis, diabetes and pain.

3 Discussion

Here, we report a comprehensive DNA methylation scoring study of 953 circulating proteins. We define 109 robust EpiScores for plasma protein levels that are independent of pQTL effects. By projecting these EpiScores into a large cohort with extant data linkage, we show that 78 EpiScores associate with the incidence of 11 leading causes of morbidity (137 EpiScore – disease associations in total). Finally, we describe EpiScore - disease associations that replicate previously measured protein - disease relationships. The bulk of the EpiScore-disease associations are independent of

common risk factors, differences in immune cell composition and GrimAge acceleration. EpiScores therefore provide a deeper view into the methylation-proteomic signature of an individual that may uncover early pathways associated with morbidity.

Whereas polygenic scores capture genetic predisposition to disease, epigenetic scores capture the body's response to a range of environmental factors [32]. Our replication of previously identified protein – diabetes relationships [30,31] suggests that epigenetic scores can identify disease-protein pathways in the absence of direct protein measurements, most likely because the EpiScores capture relevant protein regulatory patterns that are determined by the diseases state. The two studies used for replication represent the largest candidate protein characterisations of diabetes to date and the top markers identified included aminoacylase-1 (ACY-1), sex hormone binding globulin (SHBG), growth hormone receptor (GHR) and Insulin-like growth factor-binding protein 2 (IGFBP-2) [30,31]. We replicate the associations and direction of effect observed with these measured proteins, as well as several other top markers reported in the studies. There is a growing body of evidence suggesting that type 2 diabetes is mediated by a complex interaction between genetic and epigenetic regulators [8,33,34]. As proteins such as ACY-1 and GHR are thought to influence a range of diabetes-associated metabolic mechanisms [35,36], EpiScores for these proteins may provide an important bridge between epigenetic influences on protein levels and the onset of diabetes. This may facilitate early risk tracking for individuals in the absence of proteomic measurement.

With modest test set performances (for example, SHBG r = 0.18 and ACY-1 r = 0.25), it is perhaps surprising that such strong replication is observed in place of the measured protein. Nonetheless, DNA methylation scores for CRP and IL6 have been shown to perform modestly in test sets (r ~ 0.2, equivalent to ~ 4% explained variance in protein level), but augment and often outperform the measured protein related to a range of phenotypes [22,23]. Though not directly comparable to protein-specific scores, polygenic scores for disease show utility in risk prediction and stratification with similar performance thresholds to our EpiScores in independent test sets [37,38]. As EpiScores capture a range of influences on DNA methylation, they may therefore highlight a different, but complementary signal to the *in vivo* protein that is important to disease prediction. In future work, epigenetic protein score profiles could be created for specific diseases; the most extreme ranges of which could be used to stratify risk, similar to recent polygenic scoring approaches [39,40]. Our framework substantially expands the existing repertoire of methylomic predictors for protein levels.

The EpiScores we provide may therefore also augment existing composite epigenetic scores, given that many of the observed associations were independent of DNAm GrimAge.

Thirteen EpiScores were associated with multiple outcomes and the EpiScore for Complement Component 5 (C5) at baseline was associated with the onset of five morbidities, the highest number for any EpiScore. EpiScores that occupy central hubs in the disease-prediction framework may provide evidence of early, upstream pathways acting across morbidities that could be targeted in preventative approaches. C5 is cleaved into C5a and C5b as part of the immune response [41] and elevated levels of these peptides have been associated with severe inflammatory, autoimmune and neurodegenerative states [424344]. A range of C5-targetting therapeutic approaches are currently in development [44–49]. EpiScores for acute inflammatory markers have been shown to offer stable longitudinal measurements of the inflammatory state of an individual, which is likely due to DNA methylation reflecting a more consistent profile of stress in the body than protein measurements [23,25]. Given that there are issues surrounding complement peptide assay sensitivity [50], our EpiScores may therefore add value, as many are trained on proteins that occupy roles in the innate and adaptive immune response. A multi-protein signature of the inflammasome of an individual is likely to have important applications for understanding and stratifying risk for multiple immune-mediated morbidities.

For conditions such as Alzheimer's dementia, for which the early molecular regulators causing protein aggregates in the brain are still relatively unknown, the acid sphingomyelinase (ASM) EpiScore association we identified highlights disease pathways that may occur years prior to diagnosis. Acid sphingomyelinase is encoded by the SMPD1 gene and converts sphingomyelin to ceramide as part of a lysosomal pathway [51]; the association we describe supports previous work characterising aberrations to lysosomal pathways in the early stages of Alzheimer's disease [52–54]. In murine models of Alzheimer's disease pathology, excessive lysosomal ASM disrupts autophagic protein degradation and associates with accumulation of amyloid-beta and other cytotoxic proteins [52,54]. ASM has therefore been discussed as a therapeutic candidate for aging and Alzheimer's disease [54]. Several of the EpiScore protein pathways we identify are already being pursued as therapeutic candidates in various stages of development. For example, NEP (an Olink-derived EpiScore) associated with incident diabetes; a recent trial demonstrated a reduction in the rate of decline in renal function in individuals that had type 2 diabetes and heart failure resulting from the inhibition of

NEP [55,56]. NEP inhibition has also been shown to associate with improved insulin sensitivity in those with obesity and hypertension [57], which has led to this pathway being proposed as a candidate for type 2 diabetes therapy [58]. EpiScore pathways may therefore help to target therapeutic investigations.

This study has a number of limitations. First, although pQTL effects from GWAS analyses in the samples were regressed out of protein levels prior to training, there may be rare cases of unknown pQTLs that can still influence the predictors. Second, due to missing risk factor covariate information, cases were excluded from the fully-adjusted models; however, effect sizes for the 137 reported associations showed minimal attenuation in the fully-adjusted models. This indicated that missing covariate data had only a marginal effect on the findings. Third, the epitope nature of the protein measurement in the SOMAscan panel may incur probe cross-reactivity and non-specific binding [16]. Fourth, the associations present between EpiScore measures and disease incidence may have been influenced by external factors such as prescription medications for comorbid conditions and comorbid disease prevalence, which should be investigated in future analyses. Fifth, due to known heterogeneity between blood and brain methylation [59], it is unlikely that EpiScores capture the minutiae of processes in tissues with high proteomic specificity, such as the brain.

Our findings suggest that DNA methylation phenotyping approaches and data linkage to electronic health records in large, population-based studies have the potential to (1) Capture the inter-individual, environmental influences on protein levels; (2) Replicate disease associations in the absence of the measured protein; and (3) Augment risk prediction many years prior to morbidity onset. Although the EpiScore-disease associations replicated some known protein-disease findings, validation of other EpiScore-disease findings are still required. Specifically, this should entail the direct measurement of proteins, as inference from EpiScores alone, while useful for disease risk stratification, is not sufficient to determine pathways and mechanisms. Though it was only possible to assess replication for diabetes associations, several of the markers we identify for COPD and IHD have also been identified in previous studies using measured proteins [60,61]. Strengths of this study are the inclusion of cohorts that ranged in age and consisted of both Scottish and German ancestries. As EpiScores trained in these cohorts performed well when applied to cohorts of differing distributions, it is likely that our scores will generalise to other cohorts comprising healthy ageing individuals. However, these are both cohorts with European ancestries and there is evidence that performance of polygenic scores is poorer when applied to non-European ancestries [40]. It is therefore critical to

expand the diversity of training samples in future studies. Another strength is the extensive data linkage capacity in Generation Scotland that allowed us to investigate time-to-event for several common disease outcomes. Given the increasingly widespread assessment of DNAm in cohort studies [62,63], EpiScores offer an affordable and consistent (i.e. array-based) way to measure multiple biomarkers pertinent to ageing and disease. They may also provide important insights into quantifying the impact of environmental insults [21]. We provide the weights for our EpiScores, such that any study can project them into cohorts with Illumina DNA methylation data.

___

In conclusion, we show that EpiScores for circulating protein levels have the capacity to replicate protein-disease associations and identify novel biomarkers for disease, up to 14 years prior to diagnosis. EpiScore weights are adjusted for known genetic effects and are provided for future studies to utilise. This information is likely to be important in risk stratification and prevention of age-related morbidities.

4 Methods

4.1 The KORA sample population

The KORA F4 study includes 3,080 participants who reside in Southern Germany. Individuals were between 32 and 81 years of age when recruited to the study from 2006 and 2008. In the current study, there were 944 individuals with methylation, proteomics and genetic data available. The Infinium HumanMethylation450 BeadChip platform was used to generate DNA methylation data for these individuals. The Affymetrix Axiom array was used to generate genotyping data and the SOMAscan platform was used to generate proteomic measurements in the sample.

4.2 DNA methylation in KORA

Methylation data were generated for 1,814 individuals in the KORA F4 study [64]. There were 944 participants that also had protein and genotype measurements available. Briefly, during preprocessing measurements from 65 probes targeting SNPs were excluded and the R package minfi was used to perform background correction. Samples with a detection rate of less than 95%

were excluded. Next, the minfi R package was used to perform normalization on the intensity methylation measures[65], with a method consistent with the Lumi:QN +BMIQ pipeline. White blood cell proportions were estimated through the Houseman method, which utilised 473 of the 500 cell-specific CpG sites that were available on the methylation array. Of the 485,512 CpG sites available, those present on sex chromosomes (11,231 for X and 416 for Y) were excluded, along with 35 sites that had fewer than 100 measures available and 2,993 CpG sites which had non-cg status. A total of 470,837 sites were therefore available for analyses.

4.3 Proteomics in KORA

The SOMAscan platform was used to quantify protein levels in undepleted plasma. SOMAscan allows for simultaneous measurement of proteins through an aptamer-based assay [66]. The analysis pipeline was performed by SomaLogic Inc. (Boulder Colorado, USA) and 1129 SOMAmer probes were obtained using the SOMAscan assay V3.2 for samples included in the present study, which is described elsewhere [17]. Briefly, the undepleted EDTA-plasma samples were divided into 0.05, 1 and 40% dilutions and allowed to bind to their respective target SOMAmers during an incubation step in 96-well plates. Following several washes, the bead-bound target proteins underwent biotinylation and flouresence-labelled SOMAmers were photo cleaved, such that these complexes were separated from the bead and pooled. Customised arrays with oligonucleotides that were complementary to the SOMAmers were then used to quantify SOMAmer levels once they were captured on streptavidin beads. The raw intensities quantified through hybridisation between the SOMAmer and oligonucleotides produced a proxy for protein concentration. Several quality controls steps were employed to account for hybridization normalization, signal calibration and median signal normalization to control for inter-plate variation. These used standard samples which were included on each processing plate as comparative measures.

Of the 1,000 samples provided for analysis, two samples were excluded due to errors in bio-bank sampling and one based on quality control measures. Of the 997 samples available, there were 944 individuals with methylation and genotypic data. This cohort were therefore included in the current analyses. Of the 1,129 probes available, five failed the QC, leaving a total of 1,124 probes for the subsequent analysis. Protein measurements were transformed by rank-based inverse normalisation and regressed onto age, sex, known pQTLs and 20 genetic principal components of ancestry derived from the Affymetrix Axiom Array to control for population structure. pQTLs for each protein were taken from a previous GWAS in the sample [17].

4.4 The LBC1936 and LBC1921 sample populations

The Lothian Birth Cohorts of 1921 (LBC1921; N = 550) and 1936 (LBC1936; N = 1091) are longitudinal studies of healthy ageing in individuals who reside in Edinburgh and the surrounding areas [67,68]. Participants completed an intelligence test at age 11 years and were recruited for these cohorts at mean ages of 79 (LBC1921) and 70 (LBC1936). They have been followed up triennially for a series of cognitive, clinical, physical and social data, along with blood donations that have been used for genetic, epigenetic, and proteomic measurement. In the present study, there were DNAm, proteomic (Olink® platform) and genetic data for up to 875 individuals from waves 1 and 2 of the LBC1936 (at mean ages 70 and 73 years) and 162 individuals at wave 3 of the LBC1921 (at mean age 87 years).

4.5 DNAm in LBC1936 and LBC1921

DNA from whole blood was assessed using the Illumina 450 K methylation array at the Edinburgh Clinical Research Facility. Details of quality control have been described elsewhere [69,70]. Raw intensity data were background-corrected and normalised using internal controls. Following background correction, manual inspection resulted in the removal of low quality samples that presented issues related to bisulphite conversion, staining signal, inadequate hybridisation or nucleotide extension. Probes with low detection rate <95% at P < 0.01 were removed. Samples with low call rates (samples with <450,000 probes detected at P < 0.01) were also removed. Additionally, samples were removed if they had a poor match between genotype and SNP control probes, or incorrect DNA methylation-predicted sex.

4.6 Proteomics in LBC1936 and LBC1921

To generate Olink® neurology panel protein measures, plasma was extracted from 816 blood samples collected in citrate tubes at mean age 72.5 ± 0.7 years in the Lothian Birth Cohort 1936 (Wave 2) and mean age 86.7 ± 0.4 years in the Lothian Birth Cohort 1921 group (Wave 3). For the Olink® inflammatory panel, plasma was extracted from 1,047 samples collected in lithium heparin tubes at an earlier time point: mean age of 69.8 ± 0.8 years in the Lothian Birth Cohort 1936

(Wave 1). Plasma samples were analysed using either the Olink® neurology 92-plex proximity extension assay or the Olink® inflammation 92-plex proximity extension assay (Olink® Bioscience, Uppsala Sweden). From the inflammation panel, one protein (BDNF) failed quality control and was removed from the study. For a further 21 proteins, over 40% of samples fell below the lowest limit of detection. These proteins were removed from analyses leaving a final set of 70 inflammatory proteins. Two neurology proteins, MAPT and HAGH, were excluded due to >40% of observations being below the lower limit of detection. This resulted in 160 protein measurements in the LBC1936 cohort across both panels for use (n≤875). In the Lothian Birth Cohort 1921, analyses were performed in 162 individuals using only the Olink® neurology panel, resulting in 92 plasma protein measurements available for use.

In brief, 1 µL of sample was incubated in the presence of proximity antibody pairs linked to DNA reporter molecules. Upon binding of an antibody pair to their corresponding antigen, the respective DNA tails form an amplicon by proximity extension, which can be quantified by high-throughput real-time PCR. This method limits the reporting of cross-reactive events. The data were pre-processed by Olink® using NPX Manager software. Protein levels were transformed by rank-based inverse normalisation. Normalised plasma protein levels were then regressed onto age, sex, four genetic components of ancestry derived from multidimensional scaling of the Illumina 610-Quadv1 genotype array (to control for population structure) and Olink® array plate. In the LBC1936 cohort, known pQTLs were adjusted for as additional covariates, through reference to previous GWAS studies in the samples [26,27].

4.7 Generation Scotland and STRADL sample populations

Generation Scotland: the Scottish Family Health Study (GS) is a large, family-structured, population-based cohort study of >24,000 individuals from across Scotland (mean age 48 years) [71]. Recruitment took place between 2006 and 2011 with a clinical visit where detailed health, cognitive, and lifestyle information was collected along with biological samples (blood, urine, saliva). In GS, there were 9,537 individuals with DNAm and phenotypic information available for this study. The Stratifying Resilience and Depression Longitudinally (STRADL) cohort is a subset of 1,188 individuals from the GS cohort who undertook additional assessments approximately five years after the study baseline [72]. Both DNAm and proteomic data (measured using the SOMAscan platform) were available in 778 individuals from the STRADL cohort.

4.8 DNA methylation in Generation Scotland and STRADL

In the main GS cohort, blood-based DNA methylation has been generated in two separate sets using the Illumina EPIC array. Prior to quality control, Set 1 comprised 5,190 related individuals whereas Set 2 comprised 4,583 individuals, unrelated to each other and to those in Set 1. Quality control details have been reported previously [73,74]. Briefly, probes were removed based on (i) outliers from visual inspection of the log median intensity of the methylated versus unmethylated signal per array, (ii) a bead count <3 in more than 5% of samples, and (iii) ≥0.5% of samples having a detection P value >0.05 in Set 1 and ≥1% of samples having a detection P value >0.01 in Set 2. Samples were removed if there was a mismatch between their predicted sex and recorded sex or if ≥1% of CpGs had a detection P value >0.05 in Set 1 and >0.5% of CpGs had a detection P value >0.01 in Set 2. Ten saliva samples were excluded from Set 1, along with three individuals who had answered "yes" to all self-reported health conditions. One person with suspected XXY genotype and seven genetic outliers were also removed [75]. The quality-controlled dataset comprised 9,537 individuals (n_{Set1}=5,087, n_{Set2}=4,450). The same steps were also applied to process the DNAm measurements in the STRADL subset of GS.

4.9 Proteomics in STRADL

Measurements for 4,236 proteins in 1,065 individuals from the STRADL cohort were recorded using the SOMAscan® technology – with assay details as per section 4.3. There were 793 epitopes which matched between the KORA and STRADL cohorts and were included for training in KORA and testing in STRADL. Of the 160 Olink® proteins present on inflammatory and neurology panels in LBC1936, 56 matched the SOMAscan SOMAmer IDs and were used to test the EpiScores. In the STRADL cohort, there were 778 individuals with proteomics data and DNAm data available for these test set analyses. These individuals had a mean age of 60.1 ± 8.8 years. Protein measurements were transformed by rank-based inverse normalisation and regressed onto age, sex and 20 genetic principal components (derived from multidimensional scaling of genotype data from the Illumina 610-Quadv1 array).

4.10 Electronic health data linkage in Generation Scotland

Over 98% of GS participants consented to allow access to electronic health records via data linkage. This includes GP records (Read 2 codes), prescription data, and hospital records (ICD codes). These data are available both retrospectively and prospectively from the time of initial blood draw, yielding up to ~14 years of follow-up data. For the current analysis, we considered incident disease data for 12 outcomes that are leading causes of mortality and morbidity (**Supplementary Note 1**). For each outcome, prevalent cases (ascertained via retrospective ICD and Read 2 codes or self-report from a baseline questionnaire) were excluded from the analyses. Self-report data was not available for the inflammatory bowel disease (IBD) outcome which meant that prevalent cases (i.e. recorded as having disease at baseline) were only excluded based on data linkage codes. A summary of the included and excluded terms for each of the 12 outcomes can be found in **Supplementary Tables 14-25**. Alzheimer's dementia was limited to those cases/controls with age of event/censoring greater than or equal to 65 years. Breast cancer analyses was restricted to females only. Recurrent and both major and moderate episodes of depression were included in the depression trait, whereas single episodes of depression were excluded. The diabetes trait was comprised of type 2 diabetes and more general diabetes codes such as diabetic retinopathy and diabetes mellitus with renal manifestation. All type 1 and juvenile cases of diabetes were excluded from the diabetes trait.

4.11 Elastic net protein EpiScores in LBC1936 and KORA

Penalised regression models were generated for 160 proteins in LBC1936 and 793 proteins in KORA using Glmnet (Version 4.0-2) [76] in R (Version 4.0) [77]. Protein levels were the outcome and there were 428,489 CpG features per model in the LBC1936 training and 397,630 in the KORA training. An elastic net penalty was specified (alpha=0.5) and cross validation was applied. Both DNAm and protein measurements were scaled to have a mean of zero and variance of one prior to training. In the KORA analyses, 10-fold cross validation was applied. The EpiScores were then tested in STRADL (n=778). Of 480 EpiScores that generated ≥1 CpG features, there were 84 with Pearson $r > 0.1$ and $P < 0.05$ upon projection into Generation Scotland. As test set comparisons were not available for every protein in the LBC1936 analyses, an initial holdout sample was defined to select the most robust EpiScores for training in the full dataset. In the holdout sample, two folds were set aside as the test data and 10-fold cross validation was carried out on the remaining data (n_{train}=576, n_{test}=130 for neurology and n_{train}=725, n_{test}=150 for inflammatory proteins). The optimal predictors, based on

lambda values that minimised the mean cross-validation errors, were applied to the holdout data. We retained EpiScores with Pearson r > 0.1 and P < 0.05 (n=36 EpiScores). New elastic net predictors for these 36 proteins were then generated, using 12-fold cross validation in order to maximise the sample size of the training dataset. The 36 DNAm protein EpiScores from the 12-fold cross validation were then tested externally through correlations with STRADL (n=778) and LBC1921 (n=162, for the neurology panel only) protein measurements. We identified 21 EpiScores with r > 0.1 and P < 0.05 in at least one of the external test sets. Four EpiScores did not have comparisons available and were included based on their LBC1936 holdout performance.

The 109 selected EpiScores were then applied to the DNAm dataset in Generation Scotland (n=9,537). In both the STRADL and GS cohorts DNAm at each CpG site was scaled to have a mean of zero and variance of one prior to the projections. DNAm scaling was performed separately for Set 1 and Set 2 in GS.

4.12 Associations with health linkage phenotypes in Generation Scotland

Cox proportional hazards regression models adjusting for age, sex, and methylation set were used to assess the relationship between the 109 selected DNAm protein EpiScores and 12 leading causes of morbidity and mortality in Generation Scotland. Models were run using the coxme package [78] (Version 2.2-16) in R version 4.0 with a kinship matrix specified to account for relatedness in the Set 1 methylation data. Time to first event was ascertained through data linkage. Disease cases included those who had been diagnosed after baseline and subsequently died, in addition to those who received a diagnosis and remained alive. Control subjects were censored if disease free at time of death or at the end of the data linkage follow-up period. EpiScore levels were rank-base inverse normalised prior to the analyses. Fully-adjusted models included the following additional covariates measured at baseline: alcohol consumption (units consumed in the previous week); deprivation (assessed by the Scottish Index of Multiple Deprivation [79]); body mass index (kg/m^2); educational attainment (an 11-category ordinal variable: How many years altogether did you attend school or study full-time? (0: 0, 1: 1-4,2: 5-9, 3: 10-11, 4: 12-13, 5: 14-15, 6: 16-17, 7: 18-19, 8: 20-21, 9: 22-23, 10: 24+)); and a DNAm-based score for smoking status [80] which was well-correlated with the number of pack years individuals had smoked in the Generation Scotland cohort (r = 0.55, n = 9,311). Covariate phenotypes were prepared according to previous methodology [24]. A false discovery rate multiple testing correction P < 0.05 was applied to the 1306 EpiScore-disease associations (109

EpiScores by 12 incident disease traits, with 2 associations excluded for failing the global proportional hazards assumption).

Proportional hazards assumptions were checked by running the fully-adjusted models and extracting Schoenfeld residuals (global test and a test for the protein- EpiScore variable) using the coxph and cox.zph functions from the survival package [81] (Version 3.2-3). These models did not account for relatedness. For each association failing to meet the assumption (Schoenfeld residuals $P < 0.05$), a sensitivity analysis was run across yearly follow-up intervals.

To understand the influence of immune cells on EpiScore-disease relationships, the fully-adjusted Cox proportional hazards models were run with Houseman-estimated white blood cell proportions as covariates [82]. A further sensitivity analyses ran the fully-adjusted models with the estimated white blood cell proportions and GrimAge acceleration score as covariates, to understand whether the protein-specific EpiScores were contributing an additional signal to that of the GrimAge score in EpiScore-disease relationships. GrimAge acceleration was obtained via Horvath's online calculator [19]. A further analysis then assessed the estimated white blood cell proportion-diseaset relationships independently of EpiScores by running the basic and fully-adjusted Cox models with Monocyte, Bcell, CD4T, CD8T and Natural Killer cell estimates as predictors, for each white cell – disease combination.

The correlation structures of the DNAm EpiScores with DNAm-based estimated white cell proportions and phenotypic information were assessed using Pearson correlations and heatmaps using pheatmap [83] (Version 1.0.12). The ggcorrplot package (Version 0.1.3) [84] was used to generate correlation structures between DNAm EpiScores. The psych package (Version 1.9.12) [85] was used to perform principal components analysis on multiple EpiScores measures which were associated with the same diseases. A network visualisation was produced using the ggraph package (Version 2.0.4) [86]. Figures 1 and 2 were created with BioRender.com.

4.13 Replication of protein – disease associations with EpiScores

Replication testing was performed for EpiScore – diabetes associations using two previous studies [30,31]. These studies identified SOMAscan plasma protein measures that associated with both the prevalence and incidence of diabetes. In both studies, two cohorts were included for replication of biomarkers (Study 1: KORA n= 993, HUNT n= 940 [31], Study 2: AGES-Reykjavik n=5,438 and QMDiab n=356 [30]). Study 1 included the KORA dataset, which the SOMAscan EpiScores we

generated from in our current study. We characterised which of our SOMAscan-based EpiScore – diabetes associations from the fully-adjusted mixed effects Cox proportional hazards models replicated those observed with measured protein levels from the two studies described. We included both basic and fully adjusted results (with either FDR or Bonferroni-corrected $P < 0.05$) wherever available, across each of the four cohorts (**Supplementary Table 12**).

4.14 Data availability

The informed consent given by the KORA study participants does not cover posting of participant level phenotype and genotype data in public databases. However, data are available upon request from KORA-gen (http://epi.helmholtz-muenchen.de/kora-gen). Requests are submitted online and are subject to approval by the KORA board.

Lothian Birth Cohort 1936 data are available on request from the Lothian Birth Cohort Study, University of Edinburgh (simon.cox@ed.ac.uk). Lothian Birth Cohort 1936 data are not publicly available due to them containing information that could compromise participant consent and confidentiality.

According to the terms of consent for GS and GS:STRADL participants, access to data must be reviewed by the GS Access Committee. Applications should be made to access@generationscotland.org.

4.15 Code availability

Code is available with open access at the following Gitlab repository: https://gitlab.com/dannigadd/episcores-for-protein-levels.

4.16 Ethics declarations

All KORA participants have given written informed consent and the study was approved by the Ethics Committee of the Bavarian Medical Association.

All components of GS received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89). GS has also been granted Research Tissue

Bank status by the East of Scotland Research Ethics Service (REC Reference Number: 20/ES/0021), providing generic ethical approval for a wide range of uses within medical research.

Ethical approval for the LBC1921 and LBC1936 studies 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). In both studies, all participants provided written informed consent. These studies were performed in accordance with the Helsinki declaration.

5 Acknowledgements

This research was funded in whole, or in part, by the Wellcome Trust [104036/Z/14/Z, 108890/Z/15/Z, 203771/Z/16/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.**

The LBC1921 was supported by the UK's Biotechnology and Biological Sciences Research Council (BBSRC), a Royal Society–Wolfson Research Merit Award to I.J.D., and the Chief Scientist Office (CSO) of the Scottish Government's Health Directorates. The LBC1936 is supported by Age UK (Disconnected Mind project, which supports S.E.H.), the Medical Research Council (G0701120, G1001245, MR/M013111/1, MR/R024065/1, which supports S.R.C.), and the University of Edinburgh. Genotyping was supported by the Biotechnology and Biological Sciences Research Council (BB/F019394/1). Methylation typing in both the LBC1921 and LBC1936 was supported by 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. Proteomic analyses in LBC1936 and LBC1921 were supported for by the Age UK grant and NIH Grants R01AG054628 and R01AG05462802S1. This work was conducted in the Centre for Cognitive Ageing and Cognitive Epidemiology, which was supported by the Medical Research Council and Biotechnology and Biological Sciences Research Council (MR/K026992/1) and which supports I.J.D. We acknowledge Grant P2CHD042849 for supporting the Population Research Center at the University of Texas. This research was supported by Australian National Health and Medical Research Council (grants 1010374, 1046880 and 1113400) and by the Australian Research Council (DP160102400). P.M.V., and N.R.W. are supported by the NHMRC Fellowship Scheme (1078037, 1078901). A.F.M. is supported by the Australian Research Council Fellowship (FT200100837). P.M.V. was also funded by the Australian Research Council (DP160102400 and

FL180100072). Lothian Birth Cohort 1921 and 1936 proteomic analyses were supported by a National Institutes of Health (NIH) research grant (R01AG054628) which also supports E.M.T-D and S.R.C.

Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates (CZD/16/6) and the Scottish Funding Council (HR03006). Genotyping and DNA methylation profiling of the GS samples was carried out by the Genetics Core Laboratory at the Clinical Research Facility, University of Edinburgh, Scotland and was funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award "STratifying Resilience and Depression Longitudinally" ([STRADL; Reference 104036/Z/14/Z]). Proteomic analyses in STRADL were supported by Dementias Platform UK (DPUK). DPUK funded this work through core grant support from the Medical Research Council [MR/L023784/2]. C.H. is supported by an MRC University Unit Programme Grant MC_UU_00007/10 (QTL in Health and Disease). L.S. is funded by DPUK through MRC (grant no. MR/L023784/2) and the UK Medical Research Council Award to the University of Oxford (grant no. MC_PC_17215). L.S. also received support from the NIHR Biomedical Research Centre at Oxford Health NHS Foundation Trust.

K.S. and S.B.Z. are supported by the Biomedical Research Program at Weill Cornell Medicine in Qatar, a program funded by the Qatar Foundation. K.S. is also supported by QNRF grant NPRP11C-0115-180010.

The KORA study was initiated and financed by the Helmholtz Zentrum München—German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. The KORA study is funded by the Bavarian State Ministry of Health and Care through the research project DigiMed Bayern (www.digimed-bayern.de). We gratefully acknowledge the contribution of all field staff conducting the KORA F4 study.

D.A.G., R.F.H. and A.J.S. are supported by funding from the Wellcome Trust 4-year PhD in Translational Neuroscience–training the next generation of basic neuroscientists to embrace clinical research [D.A.G. and R.F.H.: 108890/Z/15/Z; A.J.S: 203771/Z/16/Z]. D.L.Mc.C. and R.E.M. are supported by Alzheimer's Research UK major project grant ARUK-PG2017B−10.

Finally, we are grateful to all study participants of KORA, LBC1936, LBC1921 and GS for their invaluable contributions to this study.

6 Author contributions

R.E.M., D.A.G., R.F.H., D.L.Mc.C., S. B. Z., and K. S. were responsible for the conception and design of the study. D.A.G., R.F.H., D.L.Mc.C., and S. B. Z. carried out the data analyses. R.E.M. and D.A.G. drafted the article. C.N., and A.C., facilitated data linkage. R.M.W., S.E.H., R.F., L.S., E.M.T.D., A.F.M., I.J.D., C.G., A.P., M.W., D.J.P., C.H., P.M.V., S.R.C., K.L.E., and A.M.M., contributed to data collection and preparation. All authors read and approved the final manuscript.

7 Competing interests

R.E.M has received a speaker fee from Illumina and is an advisor to the Epigenetic Clock Development Foundation. All other authors declare no competing interests.

References

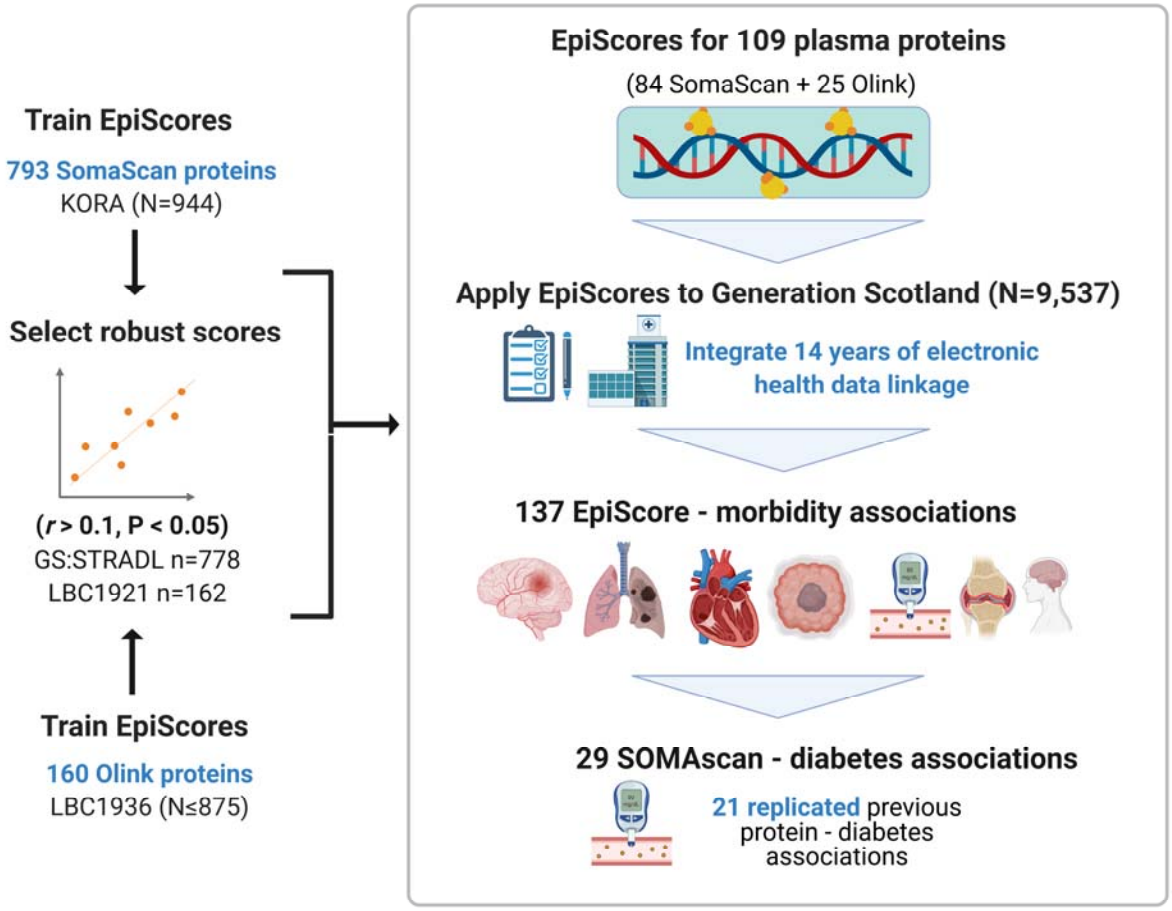
1. NHS England. Improving Outcomes Through Personalised Medicine. https://www.england.nhs.uk/wp-content/uploads/2016/09/improving-outcomes-personalised-medicine.pdf. Last Accessed April 2021. (2016).
2. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nat. Publ. Gr.* **536**, 41–47 (2016).
3. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413.
4. Yao, C. *et al.* Genome wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* **9**, 3268 (2018).
5. Läll, K., Mägi, R., Morris, A., Metspalu, A. & Fischer, K. Personalized risk prediction for type 2 diabetes: The potential of genetic risk scores. *Genet. Med.* **19**, 322–329 (2017).
6. Escott-Price, V., Shoai, M., Pither, R., Williams, J. & Hardy, J. Polygenic score prediction captures nearly all common genetic risk for Alzheimer's disease. *Neurobiol. Aging* **49**, 214.e7–214.e11 (2017).
7. Kachuri, L. *et al.* Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nat. Commun.* **11**, 6084 (2020).
8. Kwak, S. H. & Park, K. S. Recent progress in genetic and epigenetic research on type 2 diabetes. *Exp. Mol. Med.* **48**, e220 (2016).
9. Lord, J. & Cruchaga, C. The epigenetic landscape of Alzheimer's disease. **9**, 1921 (2014).
10. Lea, A. J. *et al.* Genome-wide quantification of the effects of DNA methylation on human gene regulation. *Elife* eLife 2018;7:e37513 (2018) doi:10.7554/eLife.37513.
11. Zhang, Y. *et al.* DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nat. Commun.* **6**, 10192 (2017).
12. McCartney, D. L. *et al.* Epigenetic prediction of complex traits and death. *Genome Biol.* **19**, 136 (2018).
13. Fasanelli, F. *et al.* Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat. Commun.* **6**, 10192 (2015).
14. Zhao, T., Hu, Y., Zang, T. & Wang, Y. Integrate GWAS, eQTL, and mQTL Data to Identify Alzheimer's Disease-Related Genes. *Front. Genet.* **10**, 1021 (2019).
15. Porcu, E. *et al.* Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.* **10**, 3300 (2019).
16. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
17. Suhre, K. *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8**, 14357 (2017).
18. Zaghlool, S. B. *et al.* Epigenetics meets proteomics in an epigenome-wide association study with circulating blood plasma protein traits. *Nat. Commun.* **11**, 15 (2020).
19. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany. NY).* **11**, 303–327 (2019).
20. McCartney, D. L. *et al.* Epigenetic signatures of starting and stopping smoking. *EBioMedicine* **37**, 214–220 (2018).
21. Peters, A., Nawrot, T. S. & Baccarelli, A. A. Hallmarks of environmental insults. *Cell* **184**, 1455–1468 (2021).
22. Stevenson, A. *et al.* Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. *Clin. Epigenetics* **12**, 113 (2020).
23. Stevenson, A. J. *et al.* Creating and validating a DNA methylation-based proxy for interleukin-6. *Gerontol A Biol Sci Med Sci.* 17:glab046 (2021) doi:10.1093/gerona/glab046.

24. Hillary, R. F. *et al.* Epigenetic measures of ageing predict the prevalence and incidence of leading causes of death and disease burden. *Clin. Epigenetics* **12**, 115 (2020).
25. Conole, E. L. S. *et al.* An epigenetic proxy of chronic inflammation outperforms serum levels as a biomarker of brain ageing. medRxiv. (2020) doi:https://doi.org/10.1101/2020.10.08.20205245.
26. Hillary, R. F. *et al.* Multi-method genome- And epigenome-wide studies of inflammatory protein levels in healthy older adults. *Genome Med.* **12**, 60 (2020).
27. Hillary, R. F. *et al.* Genome and epigenome wide studies of neurological protein biomarkers in the Lothian Birth Cohort 1936. *Nat. Commun.* **10**, 3160 (2019).
28. The MRC-IEU catalog of epigenome-wide association studies. Avaialable at: http://www.ewascatalog.org. Last Accessed April 2021.
29. Saffari, A. *et al.* Estimation of a significance threshold for epigenome-wide association studies. **42**, 22–23 (2017).
30. Gudmundsdottir, V. *et al.* Circulating protein signatures and causal candidates for type 2 diabetes. *Diabetes* **69**, 1843–1853 (2020).
31. Elhadad, M. A. *et al.* Deciphering the plasma proteome of type 2 diabetes. *Diabetes* **69**, 2766–2778 (2020).
32. Sugden, K. *et al.* Establishing a generalized polyepigenetic biomarker for tobacco smoking. *Transl. Psychiatry* **9**, 92 (2019).
33. 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).
34. Hall, E. *et al.* Effects of palmitate on genome-wide mRNA expression and DNA methylation patterns in human pancreatic islets. *BMC Med.* **12**, 103 (2014).
35. Pérez-Pérez, R. *et al.* Attenuated metabolism is a hallmark of obesity as revealed by comparative proteomic analysis of human omental adipose tissue. *J. Proteomics* **75**, 783–795 (2012).
36. Kim, S.-H. & Park, M.-J. Effects of growth hormone on glucose metabolism and insulin resistance in human. *Ann. Pediatr. Endocrinol. Metab.* **22**, 145–152 (2017).
37. Lewis, C. M. & Vassos, E. Polygenic risk scores: From research tools to clinical instruments. *Genome Med.* **12**, 44 (2020).
38. Dudbridge, F. & Wray, N. R. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet* **9**, e1003348 (2013).
39. Khera, A. V. *et al.* Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell* **177**, 587-596.e9 (2019).
40. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).
41. Merle, N. S., Noe, R., Halbwachs-Mecarelli, L., Fremeaux-Bacchi, V. & Roumenina, L. T. Complement system part II: Role in immunity. *Front. Immunol.* **6**, 257 (2015).
42. Ma, Y., Liu, Y., Zhang, Z. & Yang, G. Y. Significance of complement system in ischemic stroke: A comprehensive review. *Aging Dis.* **10**, 429–462 (2019).
43. Mantovani, S. *et al.* Elevation of the terminal complement activation products C5a and C5b-9 in ALS patient blood. *J. Neuroimmunol.* **276**, 213–218 (2014).
44. Morgan, B. P. & Harris, C. L. Complement, a target for therapy in inflammatory and degenerative diseases. *Nat. Rev. Drug Discov.* **14**, 857–877 (2015).
45. Hawksworth, O. A., Li, X. X., Coulthard, L. G., Wolvetang, E. J. & Woodruff, T. M. New concepts on the therapeutic control of complement anaphylatoxin receptors. *Mol. Immunol.* **89**, 36–43 (2017).
46. Ort, M., Dingemanse, J., van den Anker, J. & Kaufmann, P. Treatment of Rare Inflammatory Kidney Diseases: Drugs Targeting the Terminal Complement Pathway. *Frontiers in Immunology* vol. 11 (2020).

47. Brandolini, L. *et al.* The Novel C5aR Antagonist DF3016A Protects Neurons Against Ischemic Neuroinflammatory Injury. *Neurotox. Res.* **36**, 163–174 (2019).
48. Hernandez, M. X. *et al.* Prevention of C5aR1 signaling delays microglial inflammatory polarization, favors clearance pathways and suppresses cognitive loss. *Mol. Neurodegener.* **12**, 66 (2017).
49. Alawieh, A., Farris Langley, E. & Tomlinson, S. Targeted complement inhibition salvages stressed neurons and inhibits neuroinflammation after stroke in mice. *Sci. Transl. Med.* **10**, 6459 (2018).
50. Ekdahl, K. N. *et al.* Interpretation of serological complement biomarkers in disease. *Front. Immunol.* **9**, 2237 (2018).
51. Kamil, B., Anna, F., Anna, S., Sławomir, P. & Halina, C. Regulation of sphingomyelin metabolism. *Pharmacol. Reports* **68**, 570–581 (2016).
52. Lee, J. K. *et al.* Acid sphingomyelinase modulates the autophagic process by controlling lysosomal biogenesis in Alzheimer's disease. *J. Exp. Med.* **211**, 1551–1570 (2014).
53. Cataldo, A. M. *et al.* Aβ localization in abnormal endosomes: Association with earliest Aβ elevations in AD and Down syndrome. *Neurobiol. Aging* **25**, 1263–1272 (2004).
54. Park, M. H., Jin, H. K. & Bae, J. sung. Potential therapeutic target for aging and age-related neurodegenerative diseases: the role of acid sphingomyelinase. *Exp. Mol. Med.* **52**, 380–389 (2020).
55. Packer, M. *et al.* Effect of neprilysin inhibition on renal function in patients with type 2 diabetes and chronic heart failure who are receiving target doses of inhibitors of the renin-angiotensin system: a secondary analysis of the PARADIGM-HF trial. *Lancet Diabetes Endocrinol.* **6**, 547–554 (2018).
56. Nougué, H. *et al.* Effects of sacubitril/valsartan on neprilysin targets and the metabolism of natriuretic peptides in chronic heart failure: a mechanistic clinical study. *Eur. J. Heart Fail.* **21**, 598–605 (2019).
57. Jordan, J. *et al.* Improved Insulin Sensitivity With Angiotensin Receptor Neprilysin Inhibition in Individuals With Obesity and Hypertension. *Clin. Pharmacol. Ther.* **101**, 254–263 (2017).
58. Esser, N. & Zraika, S. Neprilysin inhibition: a new therapeutic option for type 2 diabetes? *Diabetologia* **62**, 1113–1122 (2019).
59. Hannon, E., Lunnon, K., Schalkwyk, L. & Mill, J. Interindividual methylomic variation across blood, cortex, and cerebellum: Implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics* **10**, 1024–1032 (2015).
60. Ganz, P. *et al.* Development and validation of a protein-based risk score for cardiovascular outcomes among patients with stable coronary heart disease. *JAMA* **315**, 2532–2541 (2016).
61. Serban, K. A., Pratte, K. A. & Bowler, R. P. Protein biomarkers for COPD outcomes. *Chest* ISSN 0012-3692 (2021) doi:10.1016/j.chest.2021.01.004.
62. Min, J. L. *et al.* Genomic and phenomic insights from an atlas of genetic effects on DNA methylation. *medRxiv* (2020) doi:10.1101/2020.09.01.20180406.
63. McCartney, D. L. *et al.* Genome-wide association studies identify 137 loci for DNA methylation biomarkers of ageing. *bioRxiv* (2020) doi:10.1101/2020.06.29.133702.
64. Petersen, A. K. *et al.* Epigenetics meets metabolomics: An epigenome-wide association study with blood serum metabolic traits. *Hum. Mol. Genet.* **23**, 534–545 (2014).
65. 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).
66. Gold, L. *et al.* Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* **5**, e15004 (2010).
67. 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).
68. Taylor, A. M., Pattie, A. & Deary, I. J. Cohort profile update: The Lothian birth cohorts of

1921 and 1936. *Int. J. Epidemiol.* **47**, 1042r (2018).
69. Shah, S. *et al.* Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Res.* **24**, 1725–1733 (2014).
70. Zhang, Q. *et al.* Genotype effects contribute to variation in longitudinal methylome patterns in older people. *Genome Med.* **10**, 75 (2018).
71. 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).
72. 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).
73. Seeboth, A. *et al.* DNA methylation outlier burden, health, and ageing in Generation Scotland and the Lothian Birth Cohorts of 1921 and 1936. *Clin. Epigenetics* **12**, 49 (2020).
74. McCartney, D. L. *et al.* Investigating the relationship between DNA methylation age acceleration and risk factors for Alzheimer's disease. *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.* **10**, 429–437 (2018).
75. Amador, C. *et al.* Recent genomic heritage in Scotland. *BMC Genomics* **16**, 437 (2015).
76. J, F., T, H. & R, T. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).
77. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. Last Accessed April 2021. (2020).
78. Therneau, T. M. coxme: Mixed Effects Cox Models. R package version 2.2-16. https://CRAN.R-project.org/package=coxme. Last Accessed April 2021. (2020).
79. Scottish Government. The Scottish Index of Multiple Deprivation (SIMD); 1-20. (2016). Available from: http://www.gov.scot/ Resource/0050/00504809.pdf. Last Accessed April 2021.
80. 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).
81. Therneau, T. M. A Package for Survival Analysis in R. R package version 3.2-3, https://CRAN.R-project.org/package=survival. Last Accessed April 2021. (2020).
82. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
83. Kolde, R. Pheatmap□: Pretty Heatmaps. R package version 1.0.12. https://cran.r-project.org/web/packages/pheatmap/index.html. Last Accessed April 2021. (2019).
84. Kassambara, A. ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'. R package (Version 0.1.3). https://cran.r-project.org/web/packages/ggcorrplot/ggcorrplot.pdf. Last Accessed April 2021. (2019).
85. Revelle, W. psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA. (Verison 1.9.12). https://CRAN.R-project.org/package=psych. Last Accessed April 2021. (2019).
86. Pedersen, T. L. ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. R package version 2.0.4. https://CRAN.R-project.org/package=ggraph. Last Accessed April 2021. (2020).

Fig. 1: EpiScores for plasma proteins as tools for biomarker discovery



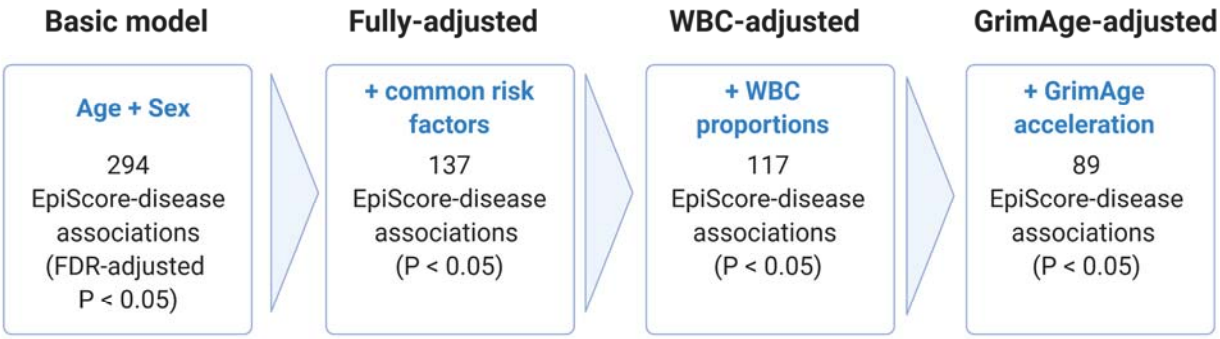
We trained DNA methylation scores specific to 953 circulating plasma protein levels in the KORA and LBC1936 cohorts. We selected the most robust EpiScores based on performance ($r > 0.1$, $P < 0.05$) in independent test sets, or in a holdout test sample for four of the proteins. We then projected the 109 most robust EpiScores into Generation Scotland, a cohort that has extensive data linkage to GP and hospital records. We tested whether levels of each EpiScore at baseline could predict the onset of 12 leading causes of morbidity, over a follow-up period of up to 14 years. 137 EpiScore – disease associations were identified, for 11 morbidities. We then assessed whether EpiScore associations reflected protein associations for diabetes, which is a trait that has been well-characterised using SOMAscan protein measurements. Of the 29 SOMAscan-derived EpiScore – diabetes associations, 21 reflected previous protein – prevalent diabetes associations and 12 replicated previous protein – incident diabetes associations.

Table 1. Incident morbidities in the Generation Scotland cohort

	Basic model			Fully-adjusted model		
Morbidity	**N cases**	**N Controls**	**Years to event (mean, sd)**	**N cases**	**N Controls**	**Years to event (mean, sd)**
Rheumatoid arthritis	65	9281	6.1 (3.5)	54	7736	6.4 (3.3)
Alzheimer's dementia	69	3764	8.3 (2.7)	52	3137	8.2 (2.7)
Bowel cancer	77	9398	6.4 (3.2)	65	7817	6.5 (3.2)
Depression	101	8306	3.9 (3.3)	80	6976	3.7 (3.3)
Breast cancer	129	5355	6 (3.4)	110	4401	5.9 (3.4)
Lung cancer	201	9265	5.2 (3.1)	172	7705	5.1 (3.1)
Inflammatory bowel disease	203	9083	5 (3.5)	163	7567	4.9 (3.5)
Stroke	317	9023	6.5 (3.4)	248	7546	6.4 (3.5)
COPD	346	8939	6.2 (3.4)	273	7459	6.1 (3.4)
Ischaemic heart disease	395	8646	5.8 (3.3)	309	7248	5.9 (3.3)
Diabetes	428	8756	5.7 (3.4)	322	7331	5.7 (3.4)
Pain	1494	5341	5.2 (3.5)	1221	4475	5.3 (3.5)

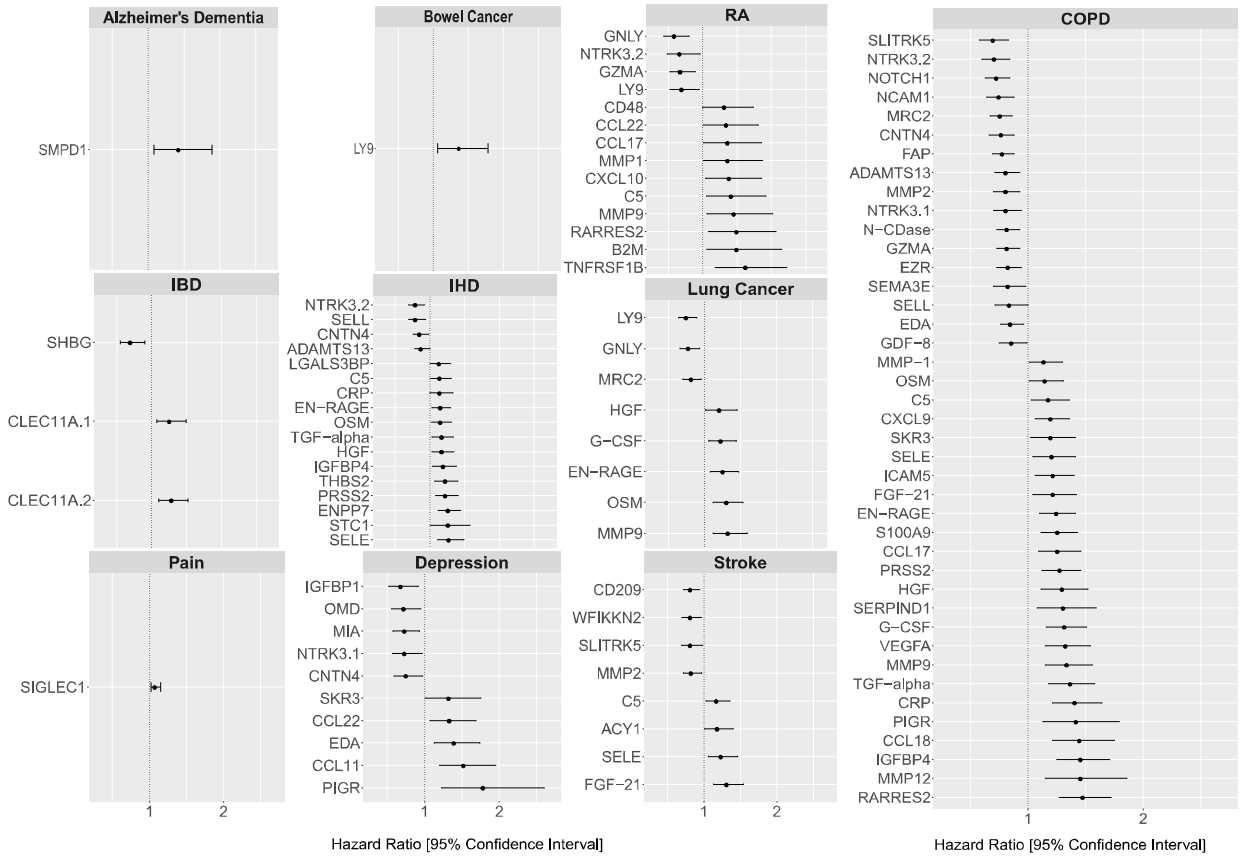
Counts are provided for the number of cases and controls for each incident trait in the basic and fully-adjusted Cox models run in the Generation Scotland cohort (n=9,537). Mean time-to-event is summarised in years for each phenotype. Alzheimer's dementia cases and controls were restricted to those older than 65 years. Breast cancer cases and controls were restricted to females.

Fig. 2: Nested Cox proportional hazards assessment of EpiScore-disease prediction



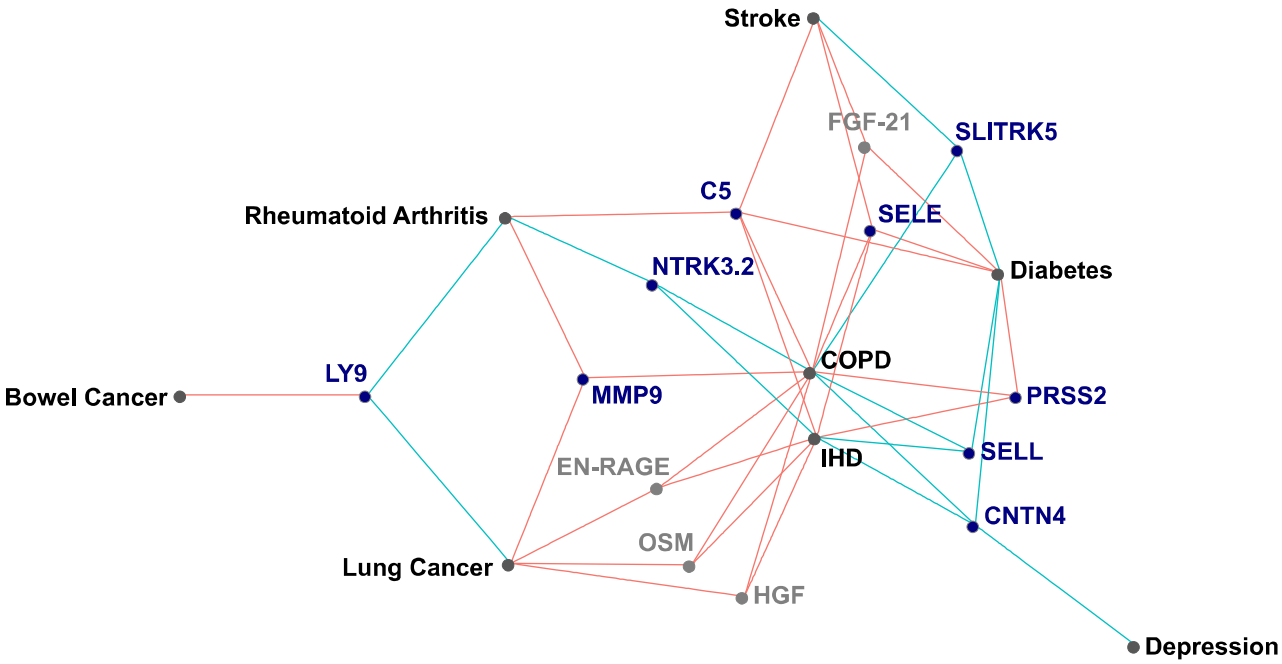
Mixed effects Cox proportional hazards analyses in Generation Scotland (n = 9,537) tested the relationships between each of the 109 selected EpiScores and the incidence of 12 leading causes of morbidity **(Supplementary Tables 5-6)**. The basic model was adjusted for age and sex and yielded 294 associations between EpiScores and disease diagnoses, with FDR-adjusted $P < 0.05$. In the fully-adjusted model, which included common risk factors as additional covariates (smoking, deprivation, educational attainment, BMI and alcohol consumption) 137 of the basic model associations remained significant with $P < 0.05$. In a sensitivity analysis, the addition of estimated White Blood Cells (WBCs) to the fully-adjusted models led to the attenuation of 20 of the 137 associations. In a further sensitivity analysis, 89 associations remained after adjustment for both immune cell proportions and GrimAge acceleration.

Fig. 3: EpiScore associations with incident disease



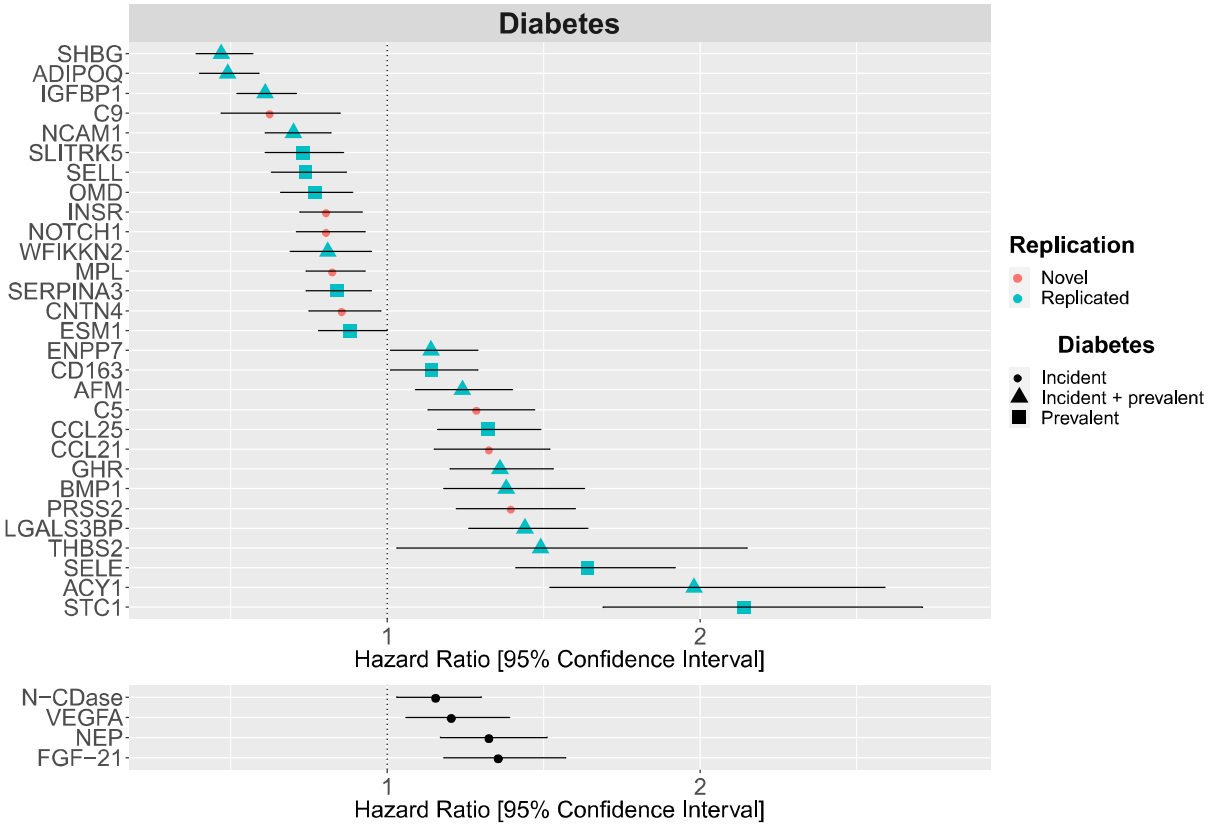
EpiScore-disease associations for ten of the eleven morbidities with associations that achieved $P < 0.05$ in the fully-adjusted mixed effects Cox proportional hazards models in Generation Scotland (n=9,537). Hazard ratios are presented with confidence intervals for 104 of the 137 EpiScore – incident disease associations reported. Models were adjusted for age, sex and common risk factors (smoking, BMI, alcohol consumption, deprivation and educational attainment). IBD: inflammatory bowel disease. IHD: ischaemic heart disease. COPD: chronic obstructive pulmonary disease. For EpiScore - diabetes associations, see **Figure 5**.

Fig. 4: EpiScores that associated with the greatest number of morbidities



EpiScores with a minimum of three relationships with incident morbidities in the fully-adjusted Cox models. The network includes 13 EpiScores as dark blue (SOMAscan) and grey (Olink) nodes, with disease outcomes in black. EpiScore-disease associations with hazard ratios < 1 are shown as blue connections, whereas hazard ratios > 1 are shown in red. COPD: chronic obstructive pulmonary disease. IHD: ischaemic heart disease.

Fig. 5: Replication of known protein-diabetes associations with EpiScores



EpiScore – incident diabetes associations in Generation Scotland (n=9,537). The 29 SOMAscan (top panel) and four Olink (bottom panel) associations shown achieved $P < 0.05$ in fully-adjusted mixed effects Cox proportional hazards models. Of the 29 SOMAscan-derived EpiScores, 21 associations replicated (blue) protein – diabetes associations in one or more of the four comparison cohorts that used SOMAscan protein levels. Eight associations were novel (pink). EpiScore – incident diabetes associations that replicated both prevalent and incident protein – diabetes associations are shown as triangles, whereas associations that replicated prevalent associations are shown as squares.