DNAm scores for serum GDF15 and NT-proBNP levels associate with a range of traits affecting the body and brain

Danni A. Gadd[1], Hannah M. Smith[1], Donncha Mullin[2,3], Ola Chybowska[1], Robert F. Hillary[1], Dorien M. Kimenai[4], Elena Bernabeu[1], Yipeng Cheng[1], Chloe Fawns-Ritchie[1,5], Archie Campbell[1], Danielle Page[6], Adele Taylor[6], Janie Corley[6], Maria Del C. Valdés-Hernández[6,7,8], Susana Muñoz Maniega[6,7,8], Mark E. Bastin[6,7,8], Joanna M. Wardlaw[6,7,8,9], Rosie M. Walker[8,10], Kathryn L. Evans[1], Andrew M. McIntosh[2], Caroline Hayward[1,11], Tom C. Russ[2,3], Sarah E. Harris[6], Paul Welsh[12], Naveed Sattar[12], Simon R. Cox[6], Daniel L. McCartney[1] and Riccardo E. Marioni[1*]

Abstract

Background Plasma growth differentiation factor 15 (GDF15) and N-terminal proB-type natriuretic peptide (NT-proBNP) are cardiovascular biomarkers that associate with a range of diseases. Epigenetic scores (EpiScores) for GDF15 and NT-proBNP may provide new routes for risk stratification.

Results In the Generation Scotland cohort ($N \geq 16,963$), GDF15 levels were associated with incident dementia, ischaemic stroke and type 2 diabetes, whereas NT-proBNP levels were associated with incident ischaemic heart disease, ischaemic stroke and type 2 diabetes (all $P_{FDR} < 0.05$). Bayesian epigenome-wide association studies (EWAS) identified 12 and 4 DNA methylation (DNAm) CpG sites associated (Posterior Inclusion Probability [PIP] > 95%) with levels of GDF15 and NT-proBNP, respectively. EpiScores for GDF15 and NT-proBNP were trained in a subset of the population. The GDF15 EpiScore replicated protein associations with incident dementia, type 2 diabetes and ischaemic stroke in the Generation Scotland test set (hazard ratios (HR) range 1.36–1.41, $P_{FDR} < 0.05$). The EpiScore for NT-proBNP replicated the protein association with type 2 diabetes, but failed to replicate an association with ischaemic stroke. EpiScores explained comparable variance in protein levels across both the Generation Scotland test set and the external LBC1936 test cohort (R^2 range of 5.7–12.2%). In LBC1936, both EpiScores were associated with indicators of poorer brain health. Neither EpiScore was associated with incident dementia in the LBC1936 population.

Conclusions EpiScores for serum levels of GDF15 and Nt-proBNP associate with body and brain health traits. These EpiScores are provided as potential tools for disease risk stratification.

Keywords GDF15, NT-proBNP, Epigenetic, DNA methylation, Dementia, Diabetes, Cardiovascular, Stroke, Risk stratification, Brain

*Correspondence:
Riccardo E. Marioni
riccardo.marioni@ed.ac.uk
Full list of author information is available at the end of the article

Gadd *et al. Clinical Epigenetics*   (2024) 16:124

Page 2 of 15

Background

Delaying or preventing the onset of chronic diseases is a major challenge. Traditional risk factor models provide a foundation to achieve this [1, 2], but can be augmented by molecular-level data. Plasma growth differentiation factor 15 (GDF15) and N-terminal pro-B-type natriuretic peptide (NT-proBNP) are biomarker candidates that have yielded promising results as indicators of a range of morbidities. GDF15 is associated with low-grade inflammation and age-related immunosuppression [3]. Higher levels of GDF15 have been found, through Mendelian randomisation, to causally associate with increased risk of cardiometabolic stroke, atrial fibrillation, coronary artery disease and myocardial infarction [4]. Two recent proteome-wide studies that assessed 1301 [5] and 1468 [6] proteins identified GDF15 as the top marker of multimorbidity. NT-proBNP is a metabolite of pro B-type natriuretic peptide (BNP), which is a natriuretic and diuretic hormone released by heart muscle in response to tension [7]. An inverse relationship between the levels of NT-proBNP in blood and incident diabetes has been reported [8], whereas lower levels of NT-proBNP have been associated with more favourable cardiovascular outcomes in randomised control trials [9–11]. Elevated levels of GDF15 and NT-proBNP in individuals diagnosed with COVID-19 have been linked to more severe outcomes [12, 13]. Both protein markers have also been found to associate with vascular brain injury, poorer neurocognitive performance and incident dementias [14, 15].

DNAm-based epigenetic scores (EpiScores) for blood proteins have been found to serve as markers of incident diseases [16] and augment clinically used risk factors for risk stratification [17, 18]. DNAm can reflect the body's chronic response to low-grade inflammation, environmental and biological exposures [19–21]. A study that directly compared an EpiScore for C-Reactive protein (CRP) to measured CRP found that the EpiScore had greater test–retest reliability over multiple time point measures [22]. This suggests that EpiScores may be more stable indictors than measured proteins in some instances. In instances where DNAm but not protein data are available, it may be possible to approximate the latter using EpiScores [23]. In studies where both DNAm and proteins are assessed, synergistic effects may be observed [21]. EpiScores may therefore lead to improved disease prediction and risk stratification [24–26]. An EpiScore for GDF15 levels based on changes to DNAm at CpG sites across the genome is one of seven protein EpiScores that contribute to GrimAge, a leading epigenetic predictor of biological age acceleration, healthspan and lifespan [27, 28]. However, the performance of protein EpiScores against within-sample protein measurements in relation to incident diseases has not been comprehensively investigated. Additionally, EpiScores have typically been trained in samples of restricted size, with training sets typically ranging from 775 to 2356 individuals [23, 27].

Here, we assess the viability of EpiScores for serum GDF15 and NT-proBNP as markers of disease outcomes and brain health (Fig. 1). Using GDF15 and NT-proBNP measures available in Generation Scotland ($N \geq 16{,}963$), we first profile associations between GDF15 and NT-proBNP and four incident diseases (type 2 diabetes, ischaemic heart disease, ischaemic stroke and dementia), in addition to COVID-19 outcome severity. These diseases were chosen for the study as they have been linked to GDF15 and NT-proBNP and were available through electronic health linkage. We next map the epigenetic architectures of the two proteins, before training and testing protein EpiScores for them in subsets of Generation Scotland. In the test set, direct biomarker comparisons between measured proteins and the EpiScore equivalents are performed in relation to the four incident diseases assessed in the full Generation Scotland sample initially. EpiScores are then retrained in the full sample available and tested externally in the Lothian Birth Cohort 1936 (LBC1936), where associations with brain health traits are also profiled cross-sectionally and longitudinally.

Results

Sample populations

There were 18,413 Generation Scotland participants (59% female) that had DNAm measurements, with a mean age of 48 years (SD 15), a minimum age of 17 years and maximum age of 98 years (Supplementary Table S1) [29, 30]. Of these, 17,489 had GDF15 measurements and 16,963 had NT-proBNP measurements. Subsets of this sample that were unrelated to one another were used to initially train and test EpiScores for GDF15 ($N_{train} = 8{,}207$, $N_{test} = 2954$) and NT-proBNP ($N_{train} = 8002$, $N_{test} = 2808$) (Supplementary Table S1). Measurements of serum GDF15 and NT-proBNP levels were available at Wave 4 (mean age 79 years, with 0.6 SD) of the LBC1936 study. These samples were used as external test sets for EpiScores trained on the full Generation Scotland sample. Of 507 individuals at Wave 4, 322 had GDF15 measures (48% female) and 500 had NT-proBNP measures (49% female). LBC1936 has successive Waves of measurements (Waves 1–5, collected at mean ages of 70, 73, 76, 79 and 82 years old, with SD < 1 at each Wave) [31, 32]. EpiScores were projected into Wave 1 (895 individuals with DNAm, but no protein measures) and evaluated in relation to cross-sectional and longitudinal brain health traits. As consent to dementia linkage was available from Wave 2,
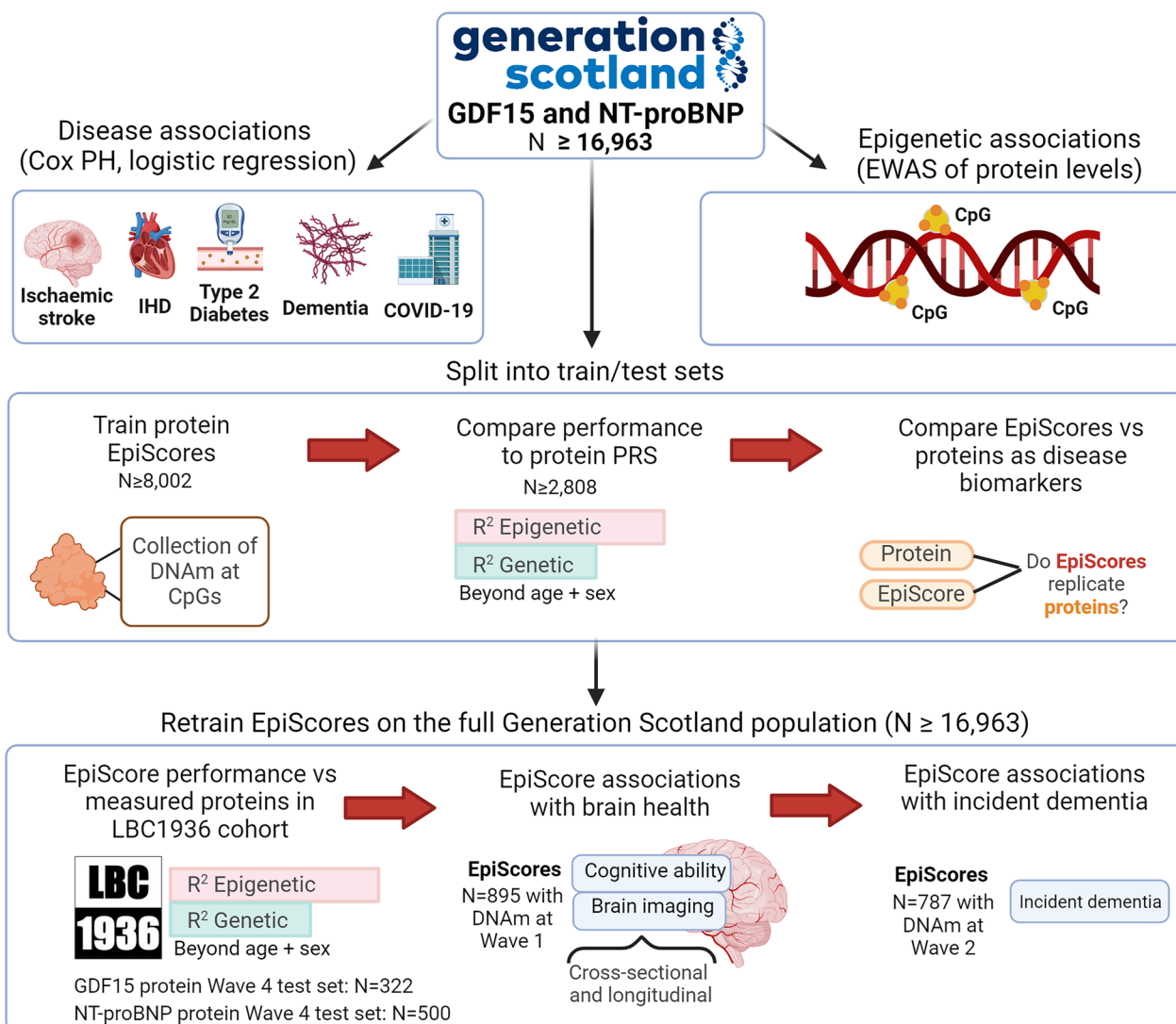
Fig. 1 Study design for this assessment of GDF15 and NT-proBNP EpiScores as biomarkers. Disease associations and epigenome-wide association studies (EWAS) for each protein were first characterised in the full Generation Scotland sample. EpiScores for each protein were initially trained and tested in subsets of the population. This allowed EpiScores to be compared with measured proteins in associations with the four incident diseases profiled in the test set. EpiScores were then retrained on the full sample and tested externally in the LBC1936 Wave 4 population, which had measures of both proteins and DNAm available. EpiScores were projected into the larger LBC1936 Wave 1 population (that has DNAm but no protein measures) and profiled in associations with brain health traits, cross-sectionally and longitudinally. Consent for dementia linkage was available from Wave 2 of the LBC1936; therefore, we also tested whether EpiScores were associated with time-to-dementia. EpiScores were modelled with polygenic risk scores (PRS) for the proteins. CpG: cytosine-phosphate-guanine. IHD: ischaemic heart disease

associations between EpiScores and time-to-dementia were also tested in LBC1936.

GDF15 and NT-proBNP disease associations

Six associations (Fig. 2) were identified in Cox proportional hazards (PH) mixed effects models between protein levels and incident diseases in Generation Scotland ($N \geq 16,963$). These associations had false discovery rate (FDR) $P < 0.05$ in basic (age and sex adjusted) models and

$P < 0.05$ in fully adjusted models (that further adjusted for smoking, alcohol intake, body mass index (BMI), social deprivation and years of education) (Supplementary Table S2). Counts for cases, controls and mean time-to-onset for cases are provided in Supplementary Table S2. In basic logistic regression models, GDF15 was associated with subsequent hospitalisation due to COVID-19 (odds ratio (OR) per SD = 2.0, 95% confidence interval (CI) = [1.2, 3.4], FDR $P = 0.037$), as opposed to having
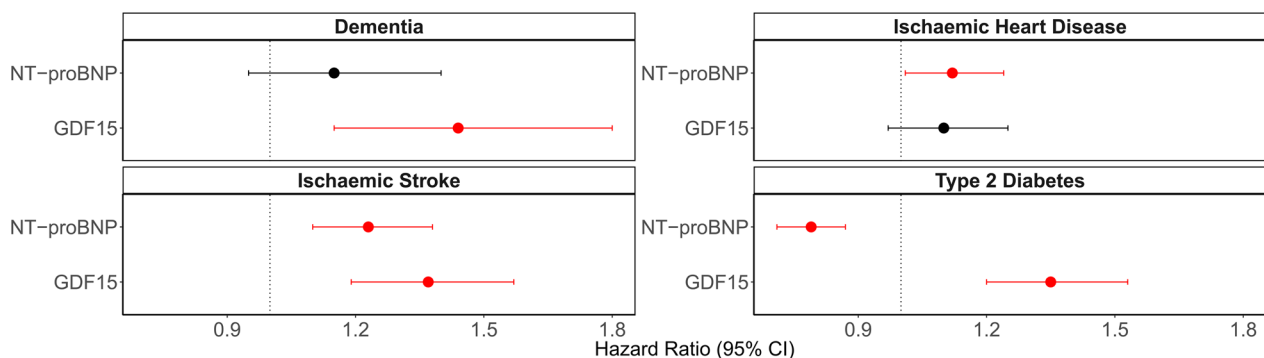
Fig. 2 Disease associations for GDF15 and NT-proBNP in Generation Scotland ($N \geq 16,963$). Fully adjusted hazard ratios from Cox PH mixed effects regression models between protein levels and incident diseases are plotted with 95% confidence intervals. The six associations in red had FDR $P < 0.05$ in basic and $P < 0.05$ in fully adjusted models, whereas associations that had $P > 0.05$ are shown in black. Hazard ratios are plotted per 1 SD increase in the rank-base inverse normalised levels of each marker. Fully adjusted models controlled for age, sex, relatedness and common health and lifestyle factors (smoking, alcohol intake, BMI, social deprivation and years of education)

COVID-19 without hospitalisation. An inverse association was identified between a one SD increase in NT-proBNP levels and COVID-19 hospitalisation (OR = 0.59, 95% confidence interval (CI) = [0.38, 0.93], FDR $P = 0.046$). No associations in relation to long-COVID as a binary outcome had FDR $P < 0.05$. Full summary statistics are provided in Supplementary Table S3.

GDF15 and NT-proBNP epigenetic associations

In variance components analysis of GDF15 and NT-proBNP in Generation Scotland ($N \geq 16,963$), epigenome-wide DNAm explained 36% of the variance in GDF15 levels (lower and upper credible intervals [CIs] = 32% to 39%) and 32% of the variance in NT-proBNP levels (lower and upper CIs = 27% to 36%). In the EWAS, there were 12 and 4 associations (Bayesian Posterior Inclusion Probability [PIP] > 95%) between differential DNAm at 14 unique CpG sites and the levels of GDF15 and NT-proBNP, respectively. The CpG sites cg03546163 (*FKBP5*) and cg13108341 (*DNAH9*) were associated with both GDF15 and NT-proBNP. Table 1 summarises the CpG sites, the biomarkers they associated with, the genes the CpGs are annotated to and a selection of traits that DNAm at these CpGs have previously been associated with in EWAS studies. The full index of MRC IEU EWAS Catalogue associations (available as of August 2023) for these 14 CpG sites is available in Supplementary Table S4.

EpiScores for GDF15 and NT-proBNP within Generation Scotland

EpiScores for GDF15 and NT-proBNP were initially trained and tested in subsets of Generation Scotland that were unrelated to one another. Predictor weights for EpiScores are available in Supplementary Table S5.

Performance in the test set was modelled through the incremental variance (R^2) in protein levels that scores could explain beyond a null linear regression model that adjusted for age and sex. The EpiScore for GDF15 trained using the full set of EPIC array probes had an R^2 of 12.2%, whereas the NT-proBNP EpiScore had an R^2 of 5.7%. Similar performance was observed when comparing with EpiScores trained using sites available on the 450 k array subset (Supplementary Fig. S1). When modelling EpiScores and polygenic risk scores (PRS) derived for each protein (see Methods), additive effects beyond the null model were observed for GDF15 (R^2 of 15.5%) and NT-proBNP (R^2 of 6.9%). A full summary of the results is provided in Supplementary Table S6.

EpiScore replication of protein biomarker associations within Generation Scotland

The same Cox PH model structure (as shown in Fig. 2) was used to directly compare protein levels and EpiScores in the Generation Scotland test set. All protein-disease associations—except the association between NT-proBNP and ischaemic stroke—were replicated by EpiScores in fully adjusted models in the test set (Fig. 3, Supplementary Table S7). Mean time-to-onset, counts for cases and controls and full summary statistics are available in Supplementary Table S7. Mean attenuation in the absolute log of the HR due to the additional adjustment for lifestyle factors beyond age and sex was 6% for protein associations and 12% for EpiScore associations. Of the four protein EpiScore associations identified in fully adjusted models, three withstood further adjustment for estimated immune cell proportions (attenuation in the absolute log of the HR ranging from 0 to 9%). The association between the GDF15 EpiScore and dementia had $P = 0.064$, with 5% attenuation in the absolute log

Gadd *et al. Clinical Epigenetics*   (2024) 16:124

Page 5 of 15

Table 1 EWAS of GDF15 and NT-proBNP levels in Generation Scotland ($N \geq 16{,}963$)

CpG	PIP	Biomarker	CpG gene	CpG trait associations (MRC-IEU EWAS catalogue)
cg03546163*	0.98, 0.96	GDF15, NT-proBNP	*FKBP5*	Chronic kidney disease, foetal vs adult liver, body mass index, waist circumference, mortality, age, neurodegenerative disorders
cg13108341*	1.00, 0.98	GDF15, NT-proBNP	*DNAH9*	Cancer treatment
cg00757033	1.00	NT-proBNP	*WDR51B*	Crohn's disease, inflammatory bowel disease, age
cg05412028	0.99	NT-proBNP	*ABCC4*	Age, ageing, primary Sjogren's syndrome
cg19693031	1.00	GDF15	*TXNIP*	Foetal vs adult liver, triglycerides, sex, hbA1c, alcohol consumption, blood pressures, hepatic fat, waist circumference, cholesterol measures, age
cg02650017	1.00	GDF15	*PHOSPHO1*	Type 2 diabetes, primary Sjogren's syndrome, C-reactive protein, body mass index, serum high-density lipoprotein cholesterol, Crohn's disease, body mass index, coagulation factor VIII, eosinophilia, age
cg06918740	1.00	GDF15	*N/A*	N/A
cg08900409	1.00	GDF15	*PGPEP1*	Age
cg25460262	1.00	GDF15	*GDF15*	Foetal vs adult liver
cg21088460	1.00	GDF15	*GDF15*	N/A
cg05575921	1.00	GDF15	*AHRR*	Extensive set of smoking-associated traits, lung function/cancer traits, body mass index, serum cotinine, C-reactive protein, IgG glycosylation measures, educational attainment, cognitive ability, statin use, urinary cadmium, mortality, post-traumatic stress disorder, age, acute myocardial infarction
cg25410121	1.00	GDF15	*N/A*	N/A
cg15058033	0.97	GDF15	*PLXNB2*	N/A
cg16993186	0.97	GDF15	*CELF2*	N/A

Posterior Inclusion Probabilities (PIPs) are provided for all CpG-protein associations (PIP > 0.95) in the BayesR EWAS. *Two CpGs were associated with both GDF15 and NT-proBNP. A selection of traits implicated in associations ($P < 3.8 \times 10^{-6}$, with $n > 100$) with the CpGs from the MRC-IEU EWAS Catalogue (as of August 2023) is shown. HbA1c: glycated haemoglobin. IgG: immunoglobulin G

of the HR, Supplementary Table S7. Expected effects of covariates were observed in the fully adjusted models, which were specific to each disease. For example, in type 2 diabetes models BMI and smoking were the two lifestyle covariates that had $p < 0.05$, with BMI representing the most significant lifestyle contributor. As only five instances of COVID-19 hospitalisation and nine instances of long-COVID were reported in the test population, we did not conduct protein EpiScore and protein comparisons for these outcomes.

EpiScore application to the LBC1936 external cohort
Demographic, covariate, cognitive test and brain imaging metrics available in the external LBC1936 population are summarised in Supplementary Tables S8–S10. EpiScores for each protein were then retrained in the entire Generation Scotland sample ($N_{GDF15} = 17{,}489$ and $N_{Nt\text{-}proBNP} = 16{,}963$). Although we make predictor weights for EpiScores trained on the full EPIC array probes and the subset of probes on the older 450 k array available (Supplementary Table S11), the LBC1936 external test cohort in this study measured DNAm using the 450 k array. Thus, the EpiScores trained on the 450 k probe subset were projected into this population for external validation.

In the LBC1936 test sample ($N_{GDF15} = 322$ and $N_{Nt\text{-}proBNP} = 500$), incremental R^2 values of 8.9% for GDF15 and 8.1% for NT-proBNP EpiScores were observed, beyond age and sex-adjusted linear regression models (Fig. 4a). When a PRS for each protein was modelled together with the EpiScores, the incremental variance explained rose to 13.7% and 9.1% for GDF15 and NT-proBNP, respectively (Supplementary Table S12). Finally, the GDF15 EpiScore generated previously by Lu et al. as part of the GrimAge biological age acceleration predictor [27] was projected into the Wave 4 GDF15 test set (322 individuals) and evaluated. It explained 5.6% of the variance in GDF15 beyond age and sex, as compared to the 8.9% observed modelling our updated GDF15 score. The two GDF15 EpiScores had a Pearson correlation $r = 0.32$ in the test sample. When modelling the GrimAge GDF15 score with our GDF15 score, 11.3% of the variance in GDF15 protein measures was explained. Adding the PRS for GDF15 into this model increased the variance explained in GDF15 protein measures to 15%.

We also compared the EpiScores derived from the larger ($N \sim 16{,}000$) and smaller ($N \sim 8000$) subsets of Generation Scotland in the external Lothian Birth Cohort 1936. The GDF15 and NT-proBNP EpiScores trained on the smaller subset explained 7.7% and 4.7% of the variance in their corresponding proteins, compared
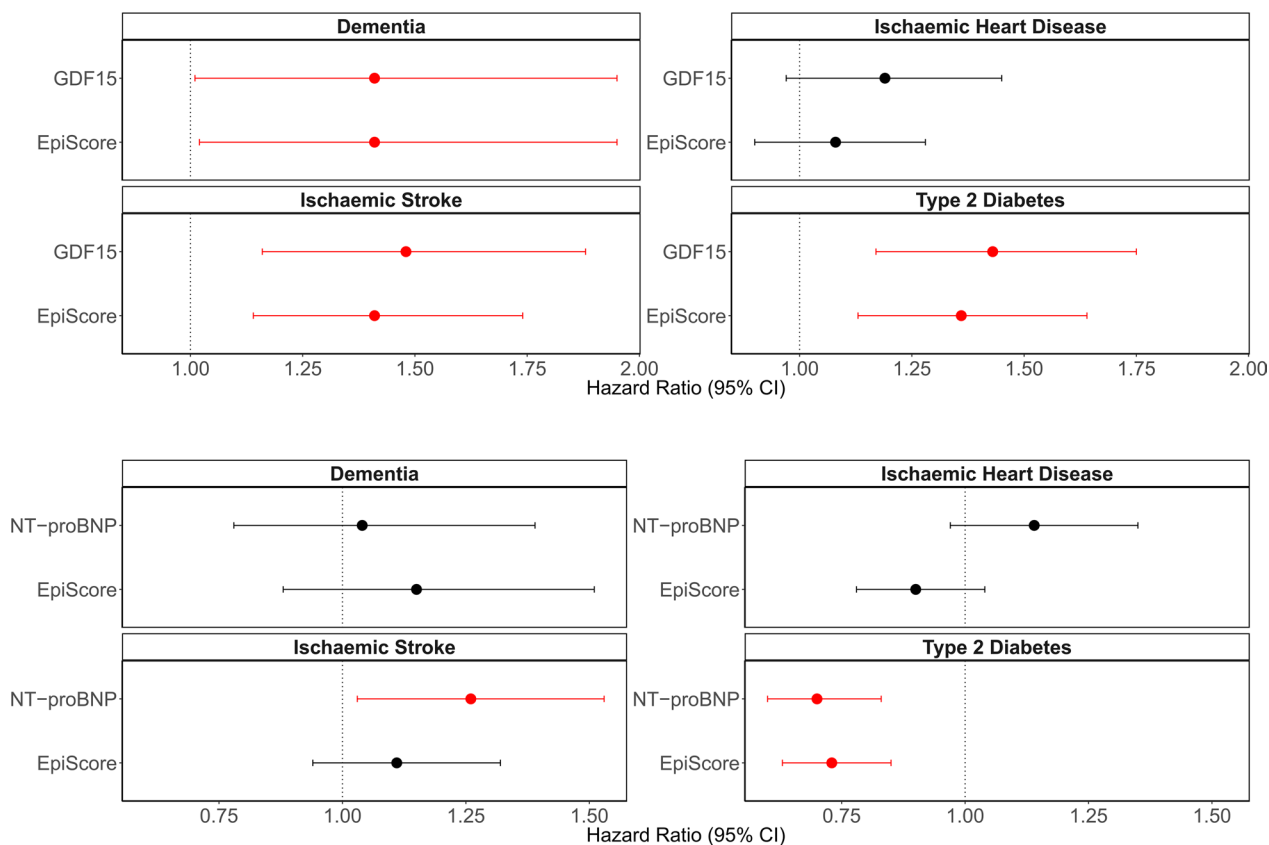
Fig. 3 Comparison of EpiScores versus measured protein equivalents in fully adjusted associations with incident diseases in the Generation Scotland test sample ($N \geq 2808$). For each disease, the protein-disease association is plotted, with the equivalent protein EpiScore-disease association shown directly beneath it for comparison. Hazard ratios are plotted per 1 SD increase in the rank-based inverse normalised levels of each marker. Nine associations (red) had FDR $P < 0.05$ in basic and $P < 0.05$ in fully adjusted Cox proportional hazards mixed effects models in the test samples. Fully adjusted models adjusted for age, sex, relatedness and common lifestyle risk factors (smoking, alcohol intake, BMI, social deprivation and years of education). Associations that were non-significant ($P > 0.05$ in fully adjusted models) are shown in black

to 8.9% and 8.1%, respectively, when using the larger training sample. These findings highlight the importance of sample size when training an EpiScore, and we strongly recommend that interested users apply the weights from Supplementary Table S11 in any applications.

EpiScore associations with brain health traits in LBC1936
The EpiScores that were validated against protein measures in the LBC1936 Wave 4 external test set were then projected into methylation measured at Wave 1 (a time point nine years prior), which has a larger DNAm sample available but no protein measures. Structural equation models were then run to characterise associations between the protein EpiScores and five brain health traits (cognitive ability and four structural brain imaging measures). This allowed for EpiScore relationships with both cross-sectional brain health (Wave 1, $N=895$ individuals with EpiScores, total model $N=1091$) and longitudinal change in brain health (Waves 1–5 for cognitive change and Waves 2–5 for

brain imaging changes) to be tested (five brain health traits x two EpiScores x cross-sectional and longitudinal associations = 20 hypothesis tests).

Seven of the twenty basic model associations tested had FDR $P < 0.05$ (Supplementary Table S13). All seven associations involved cross-sectional brain health phenotypes and had negative effect estimates (standardised betas ranging from -0.05 to -0.19). Higher GDF15 and NT-proBNP EpiScores were associated with lower general cognitive ability and lower brain volumes. None of the ten slope associations assessing relationships between the EpiScores and longitudinal change in the five brain health phenotypes were significant at FDR $P < 0.05$. In models that further adjusted for additional health and lifestyle factors, five associations had $P < 0.05$ (Fig. 4b, Supplementary Table S13). A one standard deviation increase in GDF15 EpiScore levels was associated with lower normal appearing white matter volume (Beta$=-0.07$, SE$=0.02$, $P=2.2\times10^{-3}$), poorer general cognitive ability (Beta$=-0.09$, SE$=0.04$, $P=9.1\times10^{-3}$)
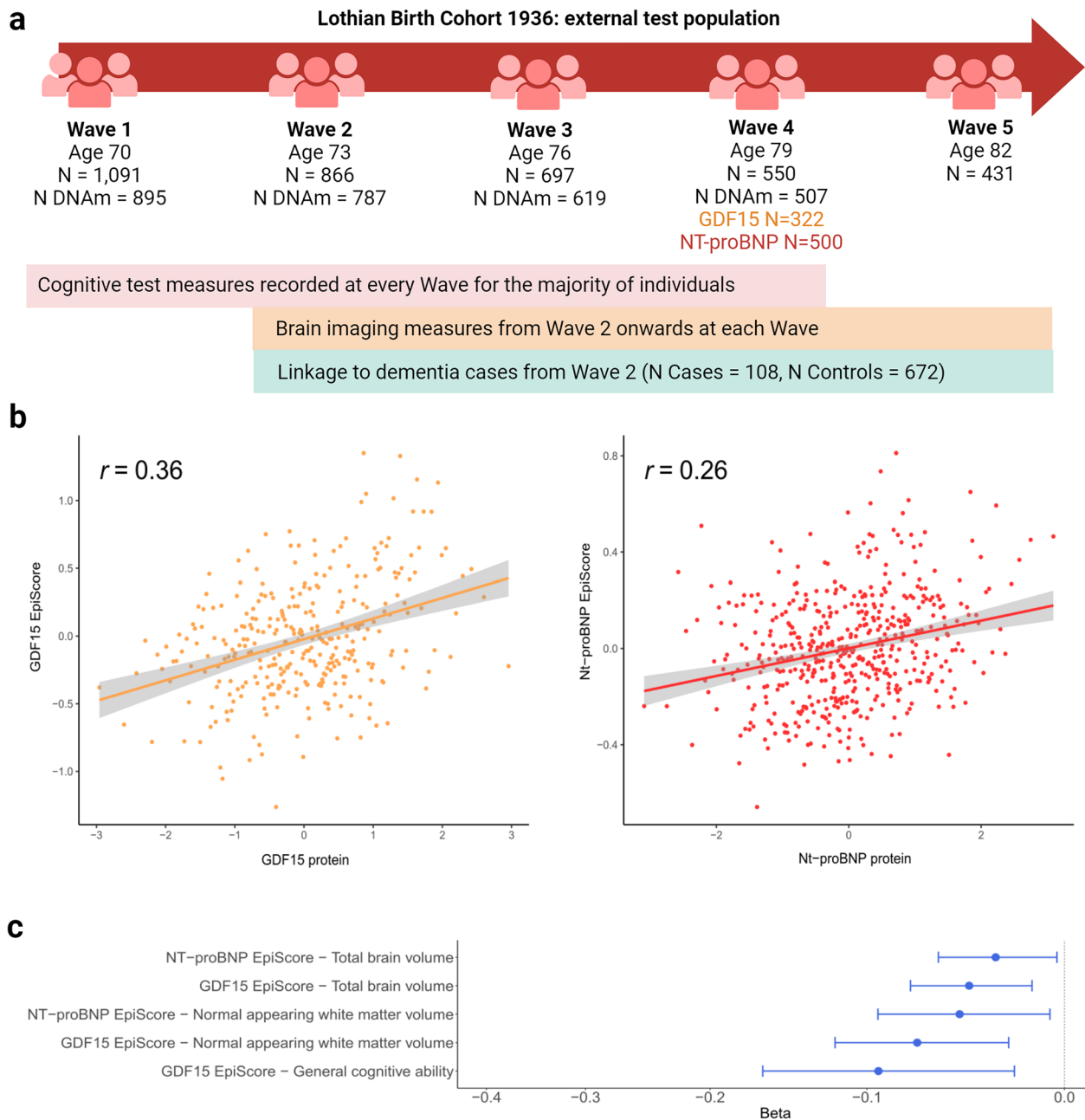
Fig. 4 External assessment of the GDF15 and NT-proBNP EpiScores in LBC1936. **a** Measurements available across the Waves of the LBC1936 external population. **b** Correlation plots between measured protein levels and GDF15 (orange) and NT-proBNP (red) EpiScores in the LBC1936 Wave 4 external test set ($N_{GDF} = 322$, $N_{NT\text{-}proBNP} = 500$). Pearson correlation coefficients are annotated in each instance. **c,** Standardised beta coefficients derived from structural equation models (SEMs) between EpiScore levels at LBC1936 Wave 1 ($N = 895$ with DNAm, $N = 1091$ total) and cross-sectional measures of brain health traits that had FDR $P < 0.05$ in basic (age and sex adjusted) models and $P < 0.05$ after adjustment for further lifestyle covariates. All associations had a negative beta coefficient (blue). Twenty EpiScore-trait associations were tested in total: 10 cross-sectionally and 10 assessing longitudinal change in brain traits

and lower total brain volume (Beta = −0.05, SE = 0.02, $P = 3.5 \times 10^{-3}$). A one standard deviation increase in NT-proBNP EpiScore levels was associated with lower normal appearing white matter volume (Beta = −0.05, SE = 0.02, $P = 0.02$) and lower total brain volume (Beta = −0.03, SE = 0.01, $P = 0.03$). In a sensitivity analysis that further adjusted for immune cell proportions, the two NT-proBNP associations were attenuated ($P > 0.07$)

Gadd *et al. Clinical Epigenetics*   (2024) 16:124

Page 8 of 15

and the association between NT-proBNP and lower cognitive ability was found to be significant (Beta = −0.10, SE = 0.04, $P = 6.0 \times 10^{-3}$). In the sensitivity analysis, the three GDF15 associations remained significant ($P < 0.05$), with a mean attenuation of 11% in Beta effect magnitude (Supplementary Table S13).

Individuals consented to share disease information from electronic health records from Wave 2 of the study onwards. In Cox regression models that utilised Wave 2 as the baseline and modelled incident dementia as the outcome ($N_{cases} = 108$, $N_{controls} = 672$, mean time-to-event for cases = 8.6 years [SD 3.42] and maximum follow-up of 14.3 years), no associations were identified for either EpiScore (Supplementary Table S14).

Discussion

Here, biomarker-disease associations for GDF15 and NT-proBNP were first observed in Generation Scotland, prior to developing EpiScores for these proteins. EpiScores replicated protein associations with incident diseases in the Generation Scotland test sample. In the LBC1936 external test population, the GDF15 and NT-proBNP EpiScores explained 9% and 8% of the variance in the protein levels, respectively, with higher levels of the EpiScores associated with poorer brain health cross-sectionally. EWAS of the proteins highlighted 14 CpGs with differential DNAm.

This study provides EpiScores for GDF15 and NT-proBNP trained in the largest samples to date as tools for health stratification. Despite the LBC1936 test set being older than the Generation Scotland cohort (mean age of 79 *versus* 48 years), the EpiScores had R^2 values comparable to those observed in the Generation Scotland test set. In the external LBC1936 test set, the GDF15 EpiScore had improved performance (an additional R^2 of 3.3%) when compared to the GDF15 EpiScore derived by Lu et al. in 2019 as part of the GrimAge biological age acceleration predictor [27]. This is likely due to differences in the sample sizes used for training the two GDF15 scores (2,356 individuals as compared to 17,489 individuals in our study). It may also be due to our training and testing populations having more homogeneous ancestry (Scottish) than the populations used to train the original GrimAge GDF15 score (mixed white European ancestry). The cumulative variance explained in GDF15 measures (15%) by our GDF15 EpiScore, the GrimAge GDF15 score and the GDF15 PRS indicates that each score may reflect a proportion of the protein signal that is unique. No other EpiScores for either GDF15 or NT-proBNP exist in the literature to our knowledge; these EpiScores can therefore be utilised as new tools for risk stratification and can be projected into any cohort with Illumina-based DNAm. We provide

EpiScore weights trained on both the 450 k and EPIC arrays for use in future research.

Generation Scotland is one of the world's largest single-cohort resources with DNAm, protein measures and extant data linkage to electronic health records. This allowed for direct comparisons between protein and EpiScore measures in the context of incident disease analyses, which have only recently been possible owing to the expansion of the cohort's epigenetic resource. As DNAm may record chronic exposure to a range of environmental risk factors [33] and biological processes such as inflammation [20, 34], EpiScores may be reflective of a range of biological pathways that occur upstream of disease diagnoses. Given that GDF15 and Nt-proBNP are promising biomarkers for a range of diseases, our EpiScores are well-positioned candidates with many potential use cases. The results of inclusion of the PRS for proteins in incremental variance models suggested that EpiScore signals were largely independent of genetic architectures on the proteins, as additive improvements in incremental variance observed when PRS and EpiScores were modelled together. This is in concordance with previous studies that found additive epi/genetic heritability estimates for plasma protein levels [35, 36]. While we have previously regressed out protein quantitative trait loci (pQTLs) from proteins prior to training EpiScores [23], there is an argument that EpiScores capturing a combination of genetic and epigenetic signatures may enhance the disease-predictive signal available. Both approaches are likely viable for the creation of new biomarkers.

The higher proportion of variance explained by the GDF15 EpiScore as compared to the NT-proBNP EpiScore suggests that GDF15 was better characterised by DNAm differences across the genome. This may be due to its association with chronic inflammation, as we have observed particularly strong DNAm signatures associated with inflammatory proteins in previous work [23, 37]. A stronger DNAm signature was also observed for GDF15 in our EWAS analyses. To our knowledge, this represents the first EWAS of NT-proBNP. The only other EWAS of GDF15 levels was performed by us previously, using GDF15 measures from the SomaLogic assay [37], where we identified no associations for GDF15 passing Bonferroni correction. The improved power to detect associations in the present study (17,489 rather than 774 individuals) may have facilitated identification of associations in the present study. There were two CpG sites associated with both GDF15 and NT-proBNP (cg03546163 in *FKBP5* and cg13108341 in *DNAH9*), which suggests a partially shared DNAm signature across the proteins. FK506-binding protein 5 (FKBP5) is implicated in cellular stress response [38]. One previous

Gadd *et al. Clinical Epigenetics*   (2024) 16:124

Page 9 of 15

study found cg03546163 to be differentially methylated in 107 individuals with type 2 diabetes that went onto develop end-stage renal disease versus 253 controls who did not [39].

The lack of associations with incident ischaemic heart disease in the Generation Scotland test set may be due to limited sample size, as an association between protein NT-proBNP and ischaemic heart disease was observed in the full Generation Scotland sample. Additionally, the GDF15 EpiScore association with incident dementia observed in Generation Scotland did not replicate in LBC1936. This may be due to differences in the way the phenotypes were defined across LBC1936 (consensus committee) versus Generation Scotland (Read and ICD codes only), or different cohort sampling frames and recruitment strategies.

Our findings support previous work identifying associations between GDF15 and Nt-proBNP protein levels and severe COVID-19 outcomes in hospitalised individuals [12, 13]. Although very few hospitalisation cases were available ($n = 28$), both proteins (sampled a mean of 11 years prior to COVID-19 diagnoses) associated with subsequent hospitalisation due to COVID-19. GDF15 is likely to be elevated in individuals with multiple morbidities that may contribute towards greater risk of hospitalisation due to viral illnesses. Diabetes has been associated with increased risk of hospitalisation and adverse outcomes in COVID-19 [40, 41]. Both proteins (and equivalent EpiScores) should be investigated in populations that have DNAm quantified nearer to, or at COVID-19 diagnoses to further resolve these signals.

This study has several limitations. First, the EpiScores were trained and tested in two Scottish ancestry cohorts. Future studies should explore if the EpiScores generalise across diverse populations. One recent study from our group showed that a Generation Scotland-trained EpiScore for the inflammatory marker, C-reactive protein, generalised to populations of different ages and ancestries [42]. Second, emerging evidence has quantified differences in genetic associations with the measurements of the same proteins across panels that use antibody-based versus aptamer-based quantification methods [43]. A particular example of interest from this study was GDF15 levels, which was highlighted as a protein that may have different conformational shapes (isoforms) that are targeted by the two assay methods [43]. While it is likely that increased training sample size led to improved performance of our GDF15 score versus the GrimAge GDF15 score in the LBC1936 test set, it is possible that technical or biological variability across protein assays may also underlie differences in performance of scores. EpiScores trained on protein measurements from different panels should therefore be compared further when data become available. Similarly, differences in the protein assay method across the previous EWAS of GDF15 (aptamer-based) that we ran and the present study (immunochemiluminescence) may also introduce variability and EWAS across multiple protein panels should be compared when samples are available. Finally, the two proteins studied here were generated as part of previous Generation Scotland studies. It would be interesting to extend our current approach to wider protein panels should they become available in Generation Scotland or other cohorts.

Conclusions

In conclusion, EpiScores for blood-based GDF15 and NT-proBNP levels are generated in this study and have been found to be useful indictors of disease risk stratification, with disease-specific use cases. The EpiScores can be derived in any population with Illumina-based DNAm measurements and may be integrated into epigenetic screening panels in research studies to better-identify high-risk individuals. The clinical use case and generalisability of the EpiScores require further research.

Methods

Generation Scotland

Generation Scotland is a population-based cohort study that includes ~ 8,000 families from across Scotland [29, 30]. Study recruitment of 23,960 participants occurred between 2006 and 2011, while participants were aged between 18 and 99 years. In addition to completing health and lifestyle questionnaires, participants donated blood samples for biomarker and omics measurement. Details on DNAm quality control in Generation Scotland are provided in Supplementary Information. The quality-controlled DNAm dataset comprised a total of 18,413 individuals with 760,838 CpG sites available on the EPIC array. GDF15 and NT-proBNP measurement details are provided in Supplementary Information. There were 17,489 individuals that had DNAm and GDF15 measures (Supplementary Table S1) (mean 1038.7 pg/mL [SD 928]). There were 16,963 individuals that had DNAm and NT-proBNP measures (mean 94.6 pg/mL [SD 211.2]). Electronic health records via data linkage to GP records (Read 2 codes) and hospital records (10th revision of the International Classification of Diseases codes [ICD-10 codes]) were assessed prospectively from the time of blood draw. Incident data for all-cause dementia, type 2 diabetes, ischaemic stroke and ischaemic heart disease were considered with censoring date October 2020. Dementia cases were defined as per a previous review of dementia linkage codes [44], whereas code lists for all other diseases are available

Gadd *et al. Clinical Epigenetics* (2024) 16:124

Page 10 of 15

in Supplementary Tables S15–S17. Prevalent cases (ascertained via retrospective linkage or self-report from a baseline questionnaire) were excluded from each disease trait, leaving only incident diagnoses. Dementia analyses were limited to cases/controls with age of event/censoring ≥ 65 years. Type 1 and juvenile diabetes cases were treated as control observations in the type 2 diabetes analyses. Death was treated as a censoring event.

Lothian Birth Cohort 1936

The Lothian Birth Cohort 1936 (LBC1936) is a longitudinal study of ageing of people residing in Edinburgh and surrounding areas in Scotland ($N = 1091$) [31, 32]. Individuals were born in 1936 and completed an intelligence test when they were 11 years old. They were later recruited to the cohort at a mean age of 70 years old and have been followed up triennially for a series of cognitive, clinical, social and physical measurements in five Waves (mean ages 70, 73, 76, 79 and 82 – all with SD below 1 for measures at each Wave). Blood samples were taken and used to derive protein, epigenetic and genetic measurements. Sample measurement details for the DNAm measures available in LBC1936 are provided in Supplementary Information. DNAm is available at the four successive waves of the study ($N = 895$, 787, 619 and 507 in Waves 1, 2, 3 and 4, respectively). Both GDF15 ($N = 322$) and NT-proBNP ($N = 500$) serum levels were measured at Wave 4 of the study (mean age 79 years, SD 0.6) and were used to externally test EpiScore performance. From Wave 2 of the LBC1936, individuals consented for linkage to health records for research. Dementia cases were defined by a consensus committee that completed decisions in August 2022 [45]. Potential cases were identified through a combination of electronic health record linkage, death certificate data and clinician visits to individuals that were suspected of having cognitive impairments or dementia. Of the 865 individuals who had provided linkage consent at Wave 2, 118 were confirmed as having dementia.

Epigenome-wide association studies in Generation Scotland

The BayesR software implements penalised Bayesian regression on complex traits and facilitates derivation of epigenome-wide variance explained in traits [33]. The BayesR method has been found to outperform linear and mixed model approaches and implicitly adjusts for probe correlations, data structure (such as relatedness) and unmeasured confounders [33, 35]. Prior mixture variances for the methylation data (760,838 CpG sites) were set to 0.001, 0.01 and 0.1, and epigenome-wide associations studies (EWAS) were run for GDF15 ($N = 17,489$) and NT-proBNP ($N = 16,963$) levels in

Generation Scotland. These prior mixtures correspond to CpGs that have varying effect sizes—explaining 0.1%, 1% or 10% of the variance in the trait outcome. We have previously shown that these variances perform well for protein EWASs, giving similar output to leading frequentist approaches [35]. Protein measurements were transformed by rank-based inverse normalisation, regressed onto age, sex and 20 genetic principal components and scaled to have a mean of 0 and variance of 1. DNAm measurements in beta format were regressed onto age, sex and processing batch and scaled to have a mean of 0 and variance of 1. Houseman immune cell estimates were included as fixed effect covariates [46]. Effect size estimates were obtained through Gibbs sampling over the posterior distribution, conditional on input data. The Gibbs protocol had 10,000 samples, with 5,000 samples of burn-in followed by a thinning of 5 samples to reduce autocorrelation. Methylation probes that had a Posterior Inclusion Probability of ≥ 95% were deemed to be significant for each protein.

EpiScore development

Elastic net penalised regression was used to train EpiScores for GDF15 and Nt-proBNP levels. As Generation Scotland has extensive phenotyping and extant linkage to primary care and hospital records, EpiScores were first trained and tested in subsets of the full sample that were unrelated to one another to facilitate direct comparisons between EpiScore and protein levels in associations with incident diseases. EpiScores were then retrained on the full Generation Scotland sample and tested in LBC1936—an external cohort. For both analyses, DNAm beta values were considered with missing CpG measurements mean imputed. To generate alternative versions of the EpiScores that can be applied to existing cohort studies with older Illumina array data (450 k array), CpGs were filtered to the intersection of the 450 k and EPIC array sites. A total of 760,838 EPIC array probes and 390,461 450 k probes were available. CpG measurements were scaled to have a mean of 0 and variance of 1, prior to training. Protein measurements in training samples were transformed by rank-based inverse normalisation, regressed onto age, sex and 20 genetic principal components and scaled to have a mean of 0 and variance of 1. Penalised regression models were performed using Big Lasso (Version 1.5.1) in R (Version 4.0) [47]. GDF15 and NT-proBNP protein levels were the outcomes. An elastic net penalty was specified (alpha = 0.5). In the within-Generation Scotland analyses tenfold cross validation was applied to select the lambda value that minimised the mean prediction error, whereas 20-fold cross validation was applied when training EpiScores in the full Generation Scotland sample.

A summary of the individuals with protein measurements available that were used to train and test EpiScores in the initial, within-Generation Scotland analyses is provided in Supplementary Fig. S2. Briefly, individuals that were part of the same family as disease cases in the test sample were excluded from the training sample. In the test subset of Generation Scotland, control individuals that were related to those in the training sample were excluded. A total of 8,207 individuals with GDF15 and 8,002 individuals with NT-proBNP measurements were therefore used to train EpiScores, while 2,954 individuals with GDF15 and 2,808 individuals with NT-proBNP measurements comprised the test samples. When retraining the EpiScores on the full Generation Scotland sample, there were 17,489 and 16,963 individuals available for the GDF15 and NT-proBNP EpiScores, respectively. Supplementary Figure S3 summarises the training and testing samples used, which included 500 individuals with NT-proBNP and 322 individuals with GDF15 measures in the external LBC1936 test set.

EpiScore testing

To test EpiScores, the additional variance in protein levels that the EpiScores explained over a null model was quantified by running the following models:

$$\text{Model 1}: \text{protein} \sim \text{age} + \text{sex}$$

$$\text{Model 2}: \text{protein} \sim \text{age} + \text{sex} + \text{protein EpiScore}$$

The incremental variance (R^2) in protein levels explained due to the protein EpiScore was calculated by subtracting the R^2 derived from model 1 from that in model 2. In these models, scaled, rank-based inverse normalised protein levels were used for testing. Pearson correlation coefficients were also calculated between GDF15 and NT-proBNP levels and their respective EpiScores in the test set and plotted. Protein EpiScores were tested using the described approach in both the Generation Scotland test subset ($N \geq 2808$) and the individuals in Wave 4 of the LBC1936 external cohort that had measures of the proteins available ($N_{GDF} = 322$, $N_{NT\text{-proBNP}} = 500$). Three incremental models were run with increasingly complex covariates: (1) basic model with age and sex as covariates, (2) fully adjusted model with lifestyle and health covariates and 3) a sensitivity analyses performed only for associations involving EpiScores, whereby DNAm-derived immune cell estimates [46] were further adjusted for.

To assess the incremental variance that could be attributed to genetic architectures of the proteins, polygenic risk scores (PRS) for the proteins were calculated using genome-wide association study (GWAS) summary statistics generated in the Generation Scotland population via BOLT-LMM [48] (see Supplementary Information). A summary of sentinel protein quantitative trait loci (pQTLs) identified by conditional and joint analyses (COJO) via Genome-wide Complex Trait Analysis (GCTA) software [49] for the GWAS results is available in Supplementary Table S18. PRS were derived using PRSice software [50]. The PRS utilised pQTLs that had $P < 5 \times 10^{-8}$, with clumping (parameters: $R^2 = 0.25$, distance $= 250$ kb, $p1 = 1$). PRS were modelled in incremental variance assessments singularly and additively with the EpiScores in the test sets in relation to measured proteins.

Cox proportional hazards analyses in Generation Scotland

Cox proportional hazards mixed effects regression models were used to assess the relationship between measured levels of GDF15 ($N = 17,489$) and NT-proBNP ($N = 16,693$) levels in the baseline Generation Scotland sample and four incident morbidities. The same model structure was also used in the test subset of the Generation Scotland sample where proteins and EpiScores were available for direct comparisons. All models were run using coxme [51] (Version 2.2–16) with a kinship matrix accounting for relatedness. Cases included those diagnosed after baseline who had died, in addition to those who received a diagnosis and remained alive. Controls were censored if disease free at time of death or at the end of the follow-up period. Date of censoring was set to October 2020, which was the latest date of the GP data linkage information. Protein levels were rank-based inverse normalised and scaled to have a mean of 0 and variance of 1 prior to analyses. Basic models were run adjusting for age and sex. Fully adjusted models further controlled for alcohol consumption (units consumed in the previous week); social deprivation (assessed by the Scottish Index of Multiple Deprivation [52]); body mass index (kilograms/height in metres squared); educational attainment (an 11-category ordinal variable); and a DNAm-based score for smoking status [53]. Each of these covariates was sampled at baseline.

An FDR multiple testing correction $P < 0.05$ was applied to basic model associations across all diseases tested. Basic associations were considered to be significant if they had FDR $P < 0.05$. Associations in fully adjusted models were considered to be significant if they had unadjusted $P < 0.05$. Proportional hazards assumptions were checked through Schoenfeld residuals (global test and a test for the protein

Gadd *et al. Clinical Epigenetics*   (2024) 16:124

Page 12 of 15

variable) using the coxph and cox.zph functions from the survival package [54] (Version 3.2–7). For each association failing to meet the assumption (Schoenfeld residuals $P < 0.05$), a sensitivity analysis was run across yearly follow-up intervals. There were minimal differences in hazard ratios between follow-up periods that did not violate the assumption and those that did. All associations were therefore retained.

COVID-19 analyses in Generation Scotland

Associations between measured levels of GDF15 and NT-proBNP and subsequent long-COVID (derived through CovidLife study survey 3 questionnaire [55]) or COVID-19 hospitalisation (derived through hospital linkage) were tested in the full Generation Scotland population. The preparation of the two binary outcome variables (long-COVID or hospitalisation due to COVID-19) is detailed in Supplementary Information. Logistic regression models with either hospitalisation (28 of 491 possible individuals) or long-COVID status (87 of 269 possible individuals) were run, with standardised (measured) proteins as the independent variables. Controls were defined as individuals that had COVID-19 but did not experience hospitalisation or long-COVID. Sex and age at COVID testing were adjusted for in the models. The latter was defined as the age at positive COVID-19 test or 1st January 2021 if COVID-19 test data were not available.

EpiScore associations with brain health traits in LBC1936

As longitudinal cognitive testing and brain morphology measures are available in LBC1936, structural equation models (SEM) were used to examine the relationship between each EpiScore and brain health traits (cross-sectionally and longitudinally). Outcomes included: general cognitive ability (*g*), total brain volume, normal-appearing white matter volume, global grey matter volume and white matter hyperintensity volume. Cognitive test data were available at all measurement Waves (mean ages 70, 73, 76, 79 and 82), and brain magnetic resonance imaging (MRI) data were available from the second Wave (mean ages 73, 76, 79 and 82). Information on how the SEM analyses were constructed, with information on the number of individuals with cognitive and brain imaging measures at each Wave, is included in Supplementary Information. Basic models were run with adjustment for age and sex, whereas fully adjusted models included further covariates: DNAm-derived immune cell proportion estimates, DNAm-derived smoking score [53], self-reported alcohol consumption, BMI and the Scottish Index of Multiple Deprivation [56]. Intercept (cross-sectional associations) and slope (longitudinal change

coefficients were extracted. A total of 1,091 individuals were modelled as part of the SEM analyses, with 895 individuals that had EpiScore measures available at Wave 1.

Individuals consented to share disease information from electronic health records from Wave 2 of the study onwards. Cox PH models were run to test associations between Wave 2 GDF15 and NT-proBNP EpiScores and incident dementia diagnoses after Wave 2 baseline, with basic adjustments for age and sex. The test population included 780 individuals who had dementia ascertainment and EpiScore information available at Wave 2, with 108 of these individuals having received a dementia diagnosis post-baseline (mean time-to-event 8.6 years [SD 3.4]). For the 108 incident cases, time-to-event was calculated using age at diagnosis. For controls who had died age at death was used for censoring, whereas age at the date of the dementia consensus meeting decision was taken forward for controls that remained alive.

Abbreviations

BMI	Body mass index
CI	Confidence interval
COJO	Conditional and joint analyses
COVID-19	Coronavirus disease 2019
CpG	Cytosine-phosphate-guanine
DNAm	DNA methylation
EpiScore	Epigenetic score
EWAS	Epigenome-wide association study
FDR	False discovery rate
GCTA	Genome-wide Complex Trait Analysis
GDF15	Plasma growth differentiation factor 15
GWAS	Genome-wide association study
HbA1c	Glycated haemoglobin
HR	Hazard ratio
ICD	International Classification of Diseases codes
IHD	Ischaemic heart disease
IgG	Immunoglobulin G
MRI	Magnetic resonance imaging
NT-proBNP	N-terminal pro-B-type natriuretic peptide
LBC1936	The Lothian Birth Cohort of 1936
OR	Odds ratio
PH	Proportional hazards
PIP	Posterior Inclusion Probability
pQTL	Protein quantitative trait loci
PRS	Polygenic risk score
SD	Standard deviation
SE	Standard error
SEM	Structural equation model

Supplementary Information

Additional file 1.

Additional file 2.

Additional file 3.

Gadd *et al. Clinical Epigenetics*   (2024) 16:124

Page 13 of 15

Availability of data and materials
The source datasets from the cohorts that were analysed during the current study are not publicly available due to them containing information that could compromise participant consent and confidentiality. Data can be obtained from the data owners. Instructions for accessing Generation Scotland data can be found here: https://www.ed.ac.uk/generation-scotland/for-resea rchers/access; the 'GS Access Request Form' can be downloaded from this site. Completed request forms must be sent to access@generationscotland.org to be approved by the Generation Scotland Access Committee. 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 access@generationscotland.org. LBC1936 data are available on request from the Lothian Birth Cohort Study, University of Edinburgh https://www.ed.ac.uk/lothian-birth-cohorts/data-access-collaboration. All correspondence and material requests should be sent to Riccardo Marioni at riccardo.marioni@ed.ac.uk. All R code used in analyses is provided at: https://github.com/DanniGadd/EpiScores-GDF15-NT-proBNP. Latent Curve Growth Model (LCGM) LBC cognitive code that was adapted for these analyses is also available at: https://www.ed.ac.uk/lothian-birth-cohorts/summary-data-resou rces.

Declarations

Ethics approval and consent to participate
All components of Generation Scotland received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89). Generation Scotland 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. LBC1936 ethical approval was obtained from the Multicentre Research Ethics Committee for Scotland (Wave 1, MREC/01/0/56), the Lothian Research Ethics Committee (Wave 1, LREC/2003/2/29) and the Scotland A Research Ethics Committee (Waves 2–5, 07/MRE00/58). All participants provided written informed consent.

Consent for publication
Not applicable.

Competing interests
R.E.M is an advisor to the Epigenetic Clock Development Foundation and has received consultant fees from Optima partners. A.M.M has previously received speaker's fees from Illumina and Janssen and research grant funding from The Sackler Trust. R.F.H. has received consultant fees from Illumina and Optima partners. D.A.G. has received consultant fees from and is currently employed in part-time capacity by Optima partners. D.L.M. is currently employed in part-time capacity by Optima partners. P.W. reports grant income from Roche Diagnostics in relation to and outside of the submitted work, as well as grant income from AstraZeneca, Boehringer Ingelheim and Novartis, outside the submitted work and speaker fees from Novo Nordisk, and Raisio outside the submitted work. N.S. has consulted for Afimmune, Amgen, AstraZeneca, Boehringer Ingelheim, Eli Lilly, Hanmi Pharmaceuticals, Merck Sharp & Dohme, Novartis, Novo Nordisk, Pfizer and Sanofi and received grant support paid to his University from AstraZeneca, Boehringer Ingelheim, Novartis and Roche Diagnostics outside the submitted work. All other authors declare no competing interest.

Author details
[1]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK. [2]Division of Psychiatry, Royal Edinburgh Hospital, University of Edinburgh, Edinburgh, UK. [3]Alzheimer Scotland Dementia Research Centre, University of Edinburgh, Edinburgh, UK. [4]BHF Centre for Cardiovascular Science, University of Edinburgh, Edinburgh, UK. [5]Division of Psychology, University of Dundee, Dundee, Scotland, UK. [6]Department of Psychology, Lothian Birth Cohorts, University of Edinburgh, Edinburgh, UK. [7]Scottish Imaging Network, A Platform for Scientific Excellence (SINAPSE) Collaboration, Edinburgh, UK. [8]Centre for Clinical Brain Sciences, Edinburgh Imaging and UK Dementia Research Institute, University of Edinburgh, Edinburgh, UK. [9]Neurovascular Imaging Research Core, UCLA, Los Angeles, CA, USA. [10]School of Psychology, University of Exeter, Perry Road, Exeter, UK. [11]Medical Research Council Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK. [12]School of Cardiovascular and Metabolic Health, BHF Glasgow Cardiovascular Research Centre, University of Glasgow, 126 University Place, Glasgow, UK.

References
1.   Hageman SHJ, McKay AJ, Ueda P, Gunn LH, Jernberg T, Hagström E, et al. Estimation of recurrent atherosclerotic cardiovascular event risk in

Gadd *et al. Clinical Epigenetics*   (2024) 16:124

Page 14 of 15

patients with established cardiovascular disease: the updated SMART2 algorithm. Eur Heart J. 2022;43(18):1715–27.

2. 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(25):2439–54.

3. Pence BD. Growth differentiation factor-15 in immunity and aging. Front Aging. 2022;0:8.

4. Wang Z, Yang F, Ma M, Bao Q, Shen J, Ye F, et al. The impact of growth differentiation factor 15 on the risk of cardiovascular diseases: two-sample Mendelian randomization study. BMC Cardiovasc Disord. 2020;20(1):1–7.

5. Tanaka T, Basisty N, Fantoni G, Candia J, Moore AZ, Biancotto A, et al. Plasma proteomic biomarker signature of age predicts health and life span. eLife. 2020;9:1–24.

6. Gadd DA, Hillary RF, Kuncheva Z, Mangelis T, Cheng Y, Dissanayake M, et al. Blood protein assessment of leading incident diseases and mortality in the UK Biobank. Nature Aging. 2024;4:39–948.

7. Shrivastava A, Haase T, Zeller T, Schulte C. Biomarkers for heart failure prognosis: proteins, genetic scores and non-coding RNAs. Front Cardiovasc Med. 2020;7:252.

8. Birukov A, Eichelmann F, Kuxhaus O, Polemiti E, Fritsche A, Wirth J, et al. Opposing associations of NT-proBNP with risks of diabetes and diabetes-related complications. Diabetes Care. 2020;43(12):2930–7.

9. Noveanu M, Breidthardt T, Potocki M, Reichlin T, Twerenbold R, Uthoff H, et al. Direct comparison of serial B-type natriuretic peptide and NT-proBNP levels for prediction of short- and long-term outcome in acute decompensated heart failure. Crit Care. 2011;15(1):1–15.

10. Myhre PL, Vaduganathan M, Claggett B, Packer M, Desai AS, Rouleau JL, et al. B-type natriuretic peptide during treatment with sacubitril/valsartan: the PARADIGM-HF trial. J Am Coll Cardiol. 2019;73(11):1264–72.

11. Zile MR, Claggett BL, Prescott MF, McMurray JJV, Packer M, Rouleau JL, et al. Prognostic implications of changes in N-terminal pro-B-type natriuretic peptide in patients with heart failure. J Am Coll Cardiol. 2016;68(22):2425–36.

12. Myhre PL, Prebensen C, Strand H, Røysland R, Jonassen CM, Rangberg A, et al. Growth differentiation factor 15 provides prognostic information superior to established cardiovascular and inflammatory biomarkers in unselected patients hospitalized With COVID-19. Circulation. 2020;142(22):2128.

13. Gao L, Jiang D, Wen XS, Cheng XC, Sun M, He B, et al. Prognostic value of NT-proBNP in patients with severe COVID-19. Respir Res. 2020;21(1):1–7.

14. McGrath ER, Himali JJ, Levy D, Conner SC, Decarli C, Pase MP, et al. Growth differentiation factor 15 and NT-proBNP as blood-based markers of vascular brain injury and dementia. J Am Heart Assoc. 2020 Oct 6 [cited 2021 Nov 24];9(19). Available from: https://pubmed.ncbi.nlm.nih.gov/32921207/

15. Walker KA, Chen J, Zhang J, Fornage M, Yang Y, Zhou L, et al. Large-scale plasma proteomic analysis identifies proteins and pathways associated with dementia risk. Nat Aging. 2021;1(5):473–89.

16. Gadd DA, Hillary RF, McCartney DL, Zaghlool SB, Stevenson AJ, Nangle C, et al. Epigenetic scores for the circulating proteome as tools for disease prediction. bioRxiv. 2021;8(5):2020.12.01.404681.

17. Cheng Y, Gadd DA, Gieger C, Monterrubio-Gómez K, Zhang Y, Berta I, et al. DNA Methylation scores augment 10-year risk prediction of diabetes. medRxiv. 2021;2021.11.19.21266469.

18. Chybowska AD, Gadd DA, Cheng Y, Bernabeu E, Campbell A, Walker RM, et al. Augmenting clinical risk prediction of cardiovascular disease through protein and epigenetic biomarkers. medRxiv; 2022 [cited 2023 Jan 5]. p. 2022.10.21.22281355. https://doi.org/10.1101/2022.10.21.22281355v1

19. Stevenson AJ, Gadd DA, Hillary RF, Mccartney DL, Campbell A, Walker RM, et al. Creating and validating a DNA methylation-based proxy for interleukin-6. [cited 2021 Mar 9]; https://doi.org/10.1093/gerona/glab046/6141415

20. Zaghlool SB, Kühnel B, Elhadad MA, Kader S, Halama A, Thareja G, et al. Epigenetics meets proteomics in an epigenome-wide association study with circulating blood plasma protein traits. Nat Commun. 2020;11(1):15.

21. Conole ELS, Stevenson AJ, Maniega SM, Harris SE, Green C, Valdés Hernández MDC, et al. DNA methylation and protein markers of chronic inflammation and their associations with brain and cognitive aging. Neurology. 2021;97(23):e2340–52.

22. Stevenson AJ, McCartney DL, Hillary RF, Campbell A, Morris SW, Bermingham ML, et al. Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. Clin Epigenetics. 2020;27(12):113.

23. Gadd DA, Hillary RF, McCartney DL, Zaghlool SB, Stevenson AJ, Cheng Y, et al. Epigenetic scores for the circulating proteome as tools for disease prediction. eLife. 2022;11.

24. Corley J, Cox SR, Harris SE, Hernandez MV, Maniega SM, Bastin ME, et al. Epigenetic signatures of smoking associate with cognitive function, brain structure, and mental and physical health outcomes in the Lothian Birth Cohort 1936. Transl Psychiatry. 2019;9(1).

25. Bernabeu E, Chybowska AD, Kresovich JK, Suderman M, McCartney DL, Hillary RF, et al. Blood-based DNA methylation study of alcohol consumption. medRxiv; 2024 [cited 2024 Jul 18]. p. 2024.02.26.24303397. https://doi.org/10.1101/2024.02.26.24303397v1

26. Hamilton OKL, Zhang Q, McRae AF, Walker RM, Morris SW, Redmond P, et al. An epigenetic score for BMI based on DNA methylation correlates with poor physical health and major disease in the Lothian Birth Cohort. Int J Obes 2005. 2019;43(9):1795–802.

27. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging. 2019;11(2):303–27.

28. Hillary RF, Stevenson AJ, McCartney DL, Campbell A, Walker RM, Howard DM, et al. Epigenetic clocks predict prevalence and incidence of leading causes of death and disease burden. Clin Epigenet. 2020;12(115).

29. Smith BH, Campbell A, Linksted P, Fitzpatrick B, Jackson C, Kerr SM, 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(3):689–700.

30. Smith BH, Campbell H, Blackwood D, Connell J, Connor M, Deary IJ, et al. Generation Scotland: the Scottish family health study; a new resource for researching genes and heritability. BMC Med Genet. 2006;7:74.

31. Taylor AM, Pattie A, Deary IJ. Cohort profile update: the Lothian birth cohorts of 1921 and 1936. Int J Epidemiol. 2018;47(4):1042r.

32. Deary IJ, Gow AJ, Pattie A, Starr JM. Cohort profile: the lothian birth cohorts of 1921 and 1936. Int J Epidemiol. 2012;41(6):1576–84.

33. Trejo Banos D, McCartney DL, Patxot M, Anchieri L, Battram T, Christiansen C, et al. Bayesian reassessment of the epigenetic architecture of complex traits. Nat Commun. 2020;11.

34. Gadd DA, Hillary RF, McCartney DL, Shi L, Stolicyn A, Robertson N, et al. Integrated methylome and phenome study of the circulating proteome reveals markers pertinent to brain health. Zenodo [Internet]. 2022 Jul 6 [cited 2022 Jul 6]; Available from: https://zenodo.org/record/6801458

35. Hillary RF, Trejo-Banos D, Kousathanas A, McCartney DL, Harris SE, Stevenson AJ, et al. Multi-method genome- and epigenome-wide studies of inflammatory protein levels in healthy older adults. Genome Med. 2020;8(12):60.

36. Hillary RF, McCartney DL, Harris SE, Stevenson AJ, Seeboth A, Zhang Q, et al. Genome and epigenome wide studies of neurological protein biomarkers in the Lothian Birth Cohort 1936. Nat Commun. 2019;10(1):3160.

37. Gadd DA, Hillary RF, McCartney DL, Shi L, Stolicyn A, Robertson NA, et al. Integrated methylome and phenome study of the circulating proteome reveals markers pertinent to brain health. Nat Commun. 2022;13(1):4670.

38. Zannas AS, Wiechmann T, Gassen NC, Binder EB. Gene–Stress–Epigenetic regulation of FKBP5: clinical and translational implications. Neuropsychopharmacology. 2016;41(1):261–74.

39. Smyth LJ, Kilner J, Nair V, Liu H, Brennan E, Kerr K, et al. Assessment of differentially methylated loci in individuals with end-stage kidney disease attributed to diabetic kidney disease: an exploratory study. Clin Epigenetics. 2021;13(1):99.

40. Rawshani A, Kjölhede EA, Rawshani A, Sattar N, Eeg-Olofsson K, Adiels M, et al. Severe COVID-19 in people with type 1 and type 2 diabetes in Sweden: a nationwide retrospective cohort study. Lancet Reg Health – Eur [Internet]. 2021 May 1 [cited 2023 Aug 20];4. Available from: https://www.thelancet.com/journals/lanepe/article/PIIS2666-7762(21)00082-X/fulltext

41. Ortega E, Corcoy R, Gratacòs M, Claramunt FXC, Mata-Cases M, Puig-Treserra R, et al. Risk factors for severe outcomes in people with diabetes hospitalised for COVID-19: a cross-sectional database study. BMJ Open. 2021;11(7):e051237.

42. Hillary RF, Ng HK, McCartney DL, Elliott HR, Walker RM, Campbell A, et al. Blood-based epigenome-wide analyses of chronic low-grade

inflammation across diverse population cohorts. Cell Genomics. 2024;4(5):100544.

43. Pietzner M, Wheeler E, Carrasco-Zanini J, Kerrison ND, Oerton E, Koprulu M, et al. Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. Nat Commun. 2021;12(1):1–13.

44. Wilkinson T, Schnier C, Bush K, Rannikmäe K, Henshall DE, Lerpiniere C, et al. Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data. Eur J Epidemiol. 2019;34(6):557–65.

45. Mullin DS, Stirland LE, Buchanan E, Convery CA, Cox SR, Deary IJ, et al. Identifying dementia using medical data linkage in a longitudinal cohort study: Lothian Birth Cohort 1936. BMC Psychiatry. 2023;1(23):303.

46. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinform. 2012;13:86.

47. R. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. Accessed April 2021. 2020.

48. Loh PR, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed model association for biobank-scale data sets. Nat Genet. 2018;50(7):906–8.

49. Yang J, Ferreira T, Morris AP, Medland SE, Madden PAF, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012;44(4):369–75.

50. Euesden J, Lewis CM, O'Reilly PF. PRSice: polygenic risk score software. Bioinformatics. 2015;31(9):1466–8.

51. Therneau TM. coxme: Mixed Effects Cox Models. R package version 2.2–16. https://CRAN.R-project.org/package=coxme. Accessed April 2021. 2020;

52. GovScot. Scottish Government. The Scottish Index of Multiple Deprivation (SIMD); 1–20. (2016). Available from: http://www.gov.scot/Resource/0050/00504809.pdf. Accessed April 2021. 2016.

53. Bollepalli S, Korhonen T, Kaprio J, Anders S, Ollikainen M. EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data. Epigenomics. 2019;11(13):1469–86.

54. Therneau TM. A Package for Survival Analysis in R. R package version 3.2–7, https://CRAN.R-project.org/package=survival. Accessed April 2021. 2020;

55. Fawns-Ritchie C, Altschul DM, Campbell A, Huggins C, Nangle C, Dawson R, et al. CovidLife: a resource to understand mental health, well-being and behaviour during the COVID-19 pandemic in the UK. Wellcome Open Res 2021 6176. 2021;6:176.

56. GovScot Scottish Government. Scottish Index of Multiple Deprivation 2006: Technical Report. Available at: https://www.gov.scot/publications/scottish-index-multiple-deprivation-2006-technical-report/. Accessed August 2023.

Publisher's Note