The circulating proteome and brain health: Mendelian randomisation and cross-sectional analyses

Rosie M. Walker PhD[a,b,c,*]

Michael Chong PhD[a,*]

Nicolas Perrot PhD[a]

Marie Pigeyre PhD[a,d]

Danni A. Gadd BSc[e]

Aleks Stolicyn PhD[c]

Liu Shi PhD[f,g]

Archie Campbell MA[h]

Xueyi Shen PhD[c]

Heather C. Whalley PhD[c,h]

Alejo Nevado-Holgado PhD[f]

Andrew M. McIntosh PhD[c]

Stefan Heitmeier[i]

Sumathy Rangarajan MSc[a]

Martin O'Donnell[a,j]

Eric E. Smith MD[a,k,l,m]

Salim Yusuf DPhil[a,†]

William N. Whiteley FRCP[a,c,n,†]

Guillaume Paré MD[a,o,p,†]

[a]Population Health Research Institute, Hamilton Health Sciences and McMaster University, Hamilton, ON, Canada

[b]School of Psychology, University of Exeter, Perry Road, Exeter, UK

[c]Centre for Clinical Brain Sciences, The University of Edinburgh, Edinburgh, UK

[d]Department of Medicine, Michael G DeGroote School of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada

[e]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

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

[g]Novo Nordisk Research Centre Oxford (NNRCO), Oxford, UK

[h]Generation Scotland, Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, EH4 2XU, UK

[i]Bayer AG, Pharmaceuticals, R&D, 42113, Wuppertal, Germany

[j]Health Research Board Clinical Research Facility, Geata an Eolais, National University of Ireland, Galway, Ireland

[k]Department of Clinical Neurosciences and Hotchkiss Brain Institute, Cumming School of Medicine, Calgary, Alberta, Canada

[l]University of Calgary, Calgary, Alberta, Canada

[m]Department of Clinical Neurosciences, University of Calgary, Calgary, Alberta, Ontario, Canada

[n]MRC Centre for Population Health, University of Oxford, Oxford, UK

[o]Thrombosis and Atherosclerosis Research Institute, Hamilton Health Sciences and McMaster University, Hamilton, ON, Canada

[p]Department of Pathology and Molecular Medicine, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada

*Joint first authors

[†]Joint final authors

Corresponding authors: pareg@mcmaster.ca and r.walker4@exeter.ac.uk

Abstract

Background: Decline in cognitive function is the most feared aspect of ageing. Poorer midlife cognitive function is associated with increased dementia and stroke risk. The mechanisms underlying variation in cognitive function are uncertain.

Methods: We assessed associations between 1160 proteins' plasma levels and two measures of cognitive function, the digit symbol substitution test (DSST) and the Montreal Cognitive Assessment in 1198 PURE-MIND participants. We assessed key MRI-ascertained structural brain phenotypes as potential mediators of associations between plasma protein levels and cognitive function. Potentially causal effects of protein levels on structural brain phenotypes and neurological outcomes were assessed using Mendelian randomisation (MR) analyses.

Results: We identified five DSST performance-associated proteins (NCAN, BCAN, CA14, MOG, CDCP1), with NCAN and CDCP1 showing replicated association in an independent cohort, GS (N=1053). MRI-assessed structural brain phenotypes partially mediated (8-19%) associations between NCAN, BCAN, and MOG, and DSST performance. MR analyses suggested higher CA14 levels might cause larger hippocampal volume and increased stroke risk, whilst higher CDCP1 levels might increase stroke and intracranial aneurysm risk.

Conclusions: We identify cognition-associated plasma proteins with potentially causal effects on brain structure and risk for neurological diseases. Our findings highlight candidates for further study and the potential for drug repurposing to reduce risk of stroke and cognitive decline.

Key words: cognition, dementia, Alzheimer's disease, stroke, proteins, brain structure, MRI, Mendelian randomisation

1. Background

Decline in cognitive ability and dementia are the most feared aspects of ageing [1], providing a strong rationale for investigating the mechanisms underlying cognitive function. Poorer cognitive function is associated with a greater risk of Alzheimer's dementia and stroke [2, 3]. This may be due to reduced "cognitive reserve", which postulates that lower premorbid cognitive function leads to worse cognitive impairment for a given degree of neuropathology [4]. Better understanding of these mechanisms could inform strategies for the prevention and treatment of dementia and stroke. Recent studies have highlighted the potential for investigating cognition and structural brain phenotypes through the study of plasma proteins; however, they have been limited by the assessment of a small set of proteins, and/or a reliance on observational association analyses [5-8].

Here, we investigate associations between 1160 plasma proteins and cognitive function in the Prospective Urban and Rural Epidemiology (PURE)-MIND [9], and sought replication in the independent imaging subsample of the Generation Scotland cohort (henceforth, referred to as "GS"). Using a simple measure of processing speed, the digit symbol substitution task (DSST), and a cognitive screening tool, the Montreal Cognitive Assessment (MoCA), we carried out a screen for cognition-associated proteins, and then employed mediation analyses to assess the proportion of the protein expression-cognition relationship that could be explained by structural brain phenotypes, including measures of brain volume and white matter hyperintensity (WMH) volume. WMH is a magnetic resonance imaging (MRI) marker of white matter damage, and is one of the manifestations of age-related cerebral small vessel disease. Two sample Mendelian randomisation (MR) analyses were performed to assess potentially causal effects of genetically predicted protein levels on genetically predicted cognitive function, brain structure, stroke subtypes, and Alzheimer's disease (see **Figure 1** for an overview of the study design).

2. Methods

Sample information

This study used data from participants with European (N = 3514), Latin (N = 4309), or Persian (N = 1332) ancestry in the Population Urban Rural Epidemiology (PURE) biomarker sub-study [32]. Participants of non-European ancestry were excluded from the present analyses due to the need to align the PURE genetic data with external genetic datasets, which are predominantly from European participants. Participants of Latin and Persian ancestry were included due to their genetic overlap with European participants [32].

The PURE biomarker study is a nested case-cohort study of the original PURE study [33] with protein biomarkers and genotyping data [32]. Cases were selected if they had experienced at least one of the following adverse health events: myocardial infarction, stroke, heart failure, type II diabetes, or death from any cause. Cohort members were selected by random sampling to obtain a group of participants who were frequency-matched by major country-specific ethnicity to the cases. The PURE biomarker study also included European participants enrolled in PURE-MIND (N = 1198) [9]. Participants from selected countries in PURE [33] were invited to participate in PURE-MIND if they were aged ≥ 39 years, had no history of stroke, dementia, or other neurological disease; had no contraindication to MRI; and could complete cognitive assessments [9].

The European, Latin, and Persian PURE biomarker cohort participants were used to identify protein biomarker protein quantitative trait loci (pQTLs) [32], for use in Mendelian randomisation analyses, whilst data from the European PURE-MIND biomarker participants were used for observational association analyses. We sought replication of our observational findings in GS (max. N = 1053) [34, 35], which was recruited through re-contact of the Generation Scotland: Scottish Family Health Study (GS:SFHS) [36, 37].

Ethical approval

All centres contributing to PURE were required to obtain approval from their respective ethics committees (Institutional Review Boards). Participant data is confidential and only authorized individuals can access study-related documents. The participants' identities are protected in documents transmitted to the Coordinating Office, as well as biomarker and genetic data. Participants provided informed consent to obtain baseline information, and to collect and store genetic and other biological specimens.

The GS:SFHS obtained ethical approval from the NHS Tayside Committee on Medical Research Ethics, on behalf of the National Health Service (reference: 05/S1401/89). All participants provided broad and enduring written informed consent for biomedical research. GS:SFHS has Research Tissue Bank Status (reference: 15/ES/0040), providing generic ethical approval for a wide range of uses within medical research. The imaging subsample of GS:SFHS (referred to as "GS" herein) received ethical approval from the NHS Tayside committee on research ethics (reference 14/SS/0039). All experimental methods were in accordance with the Helsinki declaration.

Brain imaging

PURE-MIND participants enrolled in the PURE biomarker cohort were scanned at four sites in Canada (three at 1.5T (two on General Electric (GE) scanners, one on a Phillips scanner), one at 3T (GE)). The brain imaging phenotypes assessed in this study were total brain volume (excluding ventricles), total white matter volume, hippocampal volume, average cortical thickness, a multi-region composite thickness measure designed to differentiate Alzheimer's disease patients from clinically normal participants [38], silent brain infarcts (SBI), cerebral microbleeds (CMB), and WMH volumes. These will henceforth be referred to as the "structural brain phenotypes". SBI and CMB were defined as described previously [9]. WMH volumes were estimated using Lesion Segmentation Tool (LST) in Statistical Parametric Mapping 12 (SPM12) [39]. For analyses, WMH volumes were natural log-transformed, after adding one to account for values of zero, to reduce positive skew [40]. Brain volume measurements, intracranial volumes (ICV), and average cortical thicknesses were derived from T1-weighted images using FreeSurfer v5.3 (http://surfer.nmr.mgh.harvard.edu/) [41, 42]. GS participants [34] were scanned at two sites in Scotland (both 3T) [34]. Brain volumes and ICVs were derived from T1-weighted images using Freesurfer version 5.3 [41, 42]. As in PURE-MIND, WMH volumes in GS were obtained using LST [8].

Assessment of cognitive function

General cognitive ability was measured in PURE-MIND and GS by trained assessors using the DSST (Wechsler Adult Intelligence Scale, 3rd Edition) [43]. Participants were scored according to the number of correct matches made within two minutes (maximum score: 133). PURE-MIND participants completed the Montreal Cognitive Assessment (MoCA) [10], a questionnaire-based test with scores 0 to 30.

Measurement of plasma protein expression

Proteomic and genetic analysis were conducted in the Clinical Research Laboratory & Biobank – Genetic & Molecular Epidemiology Laboratory (CRLB-GMEL), Hamilton, Canada. In the PURE biomarker cohort, plasma protein levels were measured by proximity extension assay using the Olink Proseek Target 96 reagent kit (Olink, Uppsala, Sweden). Thirteen panels (Cardiometabolic; Cardiovascular Disease II and III; Cell Regulation; Development; Immune Response; Inflammation, Metabolism; Neuro Exploratory; Neurology; Oncology I and III; and Organ Damage) were used to measure a total of 1196 biomarkers in 12066 participants, of which 3735 European, 4695 were Latin, and 1436 were Persian. The analytical performance of these panels has been validated previously and further information can be found elsewhere (https://www.olink.com/products-services/target/). Quality control and pre-processing were performed as described previously for the Cardiovascular Disease II panel [32], with the exception that the data were quantile normalised within three, rather than two, reagent lots. Missing biomarker values were imputed by the mean, separately for each reagent lot, and all values were rank-based inverse normalised by reagent lot, sex and ethnic group. Where multiple biomarker measurements from different proximity extension assays were available for a single protein, the mean value was taken. Following quality control, measurements were available for 1160 biomarkers in between 8369 and 9154 European, Latin, or Persian participants (depending on biomarker-specific missingness).

In GS, plasma protein levels were measured with the SOMAscan assay platform (SomaLogic Inc.), as described previously [44]. Following initial data processing and quality control steps, measures of 4058 proteins were available in 1095 participants. Prior to analysis, protein abundance measurements were log-transformed and rank-based inverse normalised.

Genotyping and imputation of the PURE-MIND discovery sample

PURE participant genotypes (Thermofisher Axiom Precision Medicine Research Array r.3) were called using Axiom Power Tools and in-house scripts. Samples were removed if: they had a low signal-to-noise contrast (Dish Quality Control < 0.82); low quality control rate (QCCR < 0.97); <95% call rate; disagreement between self-reported sex and/or ethnicity and genetically determined sex and/or ethnicity; were duplicated; or had excess heterozygosity. We removed variants with: a call rate <98.5%; Hardy-Weinberg equilibrium $p < 1 \times 10^{-5}$; plate or batch effects; non-Mendelian segregation within families; and/or a minor allele frequency <0.005%. Following quality control, 749,783 variants remained [32].

Imputation was performed on the 749,783 variants following the TOPMed Imputation server pipeline (https://imputation.biodatacatalyst.nhlbi.nih.gov/), using the TOPMed release 2 reference panel [45]. EAGLE v2.4 [46] and Minimac4 programs were applied for phasing and imputation, respectively. Imputed variants with an info score ≥ 0.3 and MAF ≥ 0.01, which did not deviate from Hardy-Weinberg equilibrium ($p \geq 1 \times 10^{-5}$) were retained.

Assessment of the association between protein biomarkers and cognitive function and structural imaging phenotypes

We assessed the association between standardised protein levels and cognitive and structural brain phenotypes using linear (DSST, MoCA, total brain volume, white matter volume, hippocampal volume, WMH volume, cortical thickness) or logistic (CMB, SBI) regression. The cognitive or structural brain phenotype-of-interest was modelled as the dependent variable with the standardised protein expression level, age, age^2, sex, ethnicity, and first ten genetic principal components included as independent variables. A sensitivity analysis was performed for DSST-associated proteins in which we further adjusted for education (a categorical variable with levels: (i) no education; (ii) high school or less; (iii) trade school; and (iv) college or university). We calculated Pearson's correlation coefficient to assess the

pairwise correlations between DSST-associated proteins. Within each analysis, we applied a Bonferroni correction to determine statistical significance, yielding the following significance thresholds: $p < 4.31 \times 10^{-5}$ when assessing associations with 1160 proteins; $p < 2.5 \times 10^{-3}$ when assessing associations with the five DSST-associated proteins across four DSST-associated structural brain phenotypes; and $p < 5 \times 10^{-3}$ when assessing 10 pairwise correlations between proteins.

We performed replication analyses in GS for the significant proteins identified in PURE-MIND. Mixed effects models were fitted using the lmekin function from the R package coxme v.2.2.17 [47] to assess the association of the outcome variable (DSST performance, total brain volume (excluding ventricles), cerebral white matter volume, hippocampal volume, and WMH volume) with standardised protein expression, covarying for age, age^2, sex, study site (Dundee or Aberdeen), the delay between blood sampling and protein extraction, depression (a binary variable representing lifetime depression status), and a kinship matrix. When a brain volume phenotype was the outcome variable, additional covariates were included to account for ICV, the interaction between ICV and study site (to account for a site-associated batch effect on ICV measurement), and whether there was manual intervention using tools within Freesurfer during the quality control process. Replication was defined as a concordant direction of effect, meeting a Bonferroni-corrected threshold of $p < 1.67 \times 10^{-2}$ (accounting for the assessment of three DSST-associated proteins) or $p < 7.14 \times 10^{-3}$ (accounting for the assessment of seven structural brain phenotype-protein combinations).

Assessment of the association of DSST performance with MRI-derived structural brain phenotypes

To identify mediators of the association between protein expression and DSST performance, we first established the structural brain phenotypes that satisfied the requirements of potential mediators (i.e. associated with both DSST performance and at least one DSST-associated protein), and then formally tested the meditation relationship by bootstrap mediation analyses.

We estimated the association between DSST performance and structural brain phenotypes in PURE-MIND using linear models. All brain volume measurements were normalised to ICV and the models included covariates for age, age^2, sex, and the first ten genetic principal components. We defined statistical significance as $p < 0.00625$ (Bonferroni correction for eight phenotypes) and sought replication of significant associations (N = 4) in GS. In GS, brain volumes were residualised for ICV, scanner location, the interaction between ICV and scanner location, and whether there was manual intervention during the quality control process. The resultant residuals were included as the dependent variable in a mixed effects model with DSST score, age, age^2, sex, depression, and a kinship matrix as independent variables. Statistical significance was defined as $p < 0.0125$.

The DSST performance-associated brain MRI phenotypes (N = 4) were assessed as potential mediators of the protein level-DSST associations (N = 3, yielding a total of N = 9 mediations to assess) using bootstrap mediation analysis in PURE-MIND. Analyses were performed using the R package "mediation" [48] with 1000 bootstraps. We corrected for the nine potential mediation relationships assessed using a Bonferroni-corrected threshold of $p < 5.56 \times 10^{-3}$.

Functional and tissue-specific expression enrichment analyses

Proteins associated with DSST performance at $p < 0.05$ in PURE-MIND were included in functional and tissue-specific expression analyses in three groups: (i) all proteins; (ii) positively associated proteins; and (iii) negatively associated proteins. Enrichment was assessed relative to all measured proteins (n=1160). Functional enrichment analyses were performed using WebGestalt (http://www.webgestalt.org/) [49] using default parameter settings for the over-representation analysis method to assess enrichment for: (i) gene ontology categories (biological processes, molecular functions, and cellular compartments); (ii) Reactome pathways; and (iii) disease-associated genes (Disgenet). Tissue-specific enrichment analyses were performed using the "GTEx v8: 54 tissue types" and "GTEx v8: 30 general tissue types" gene expression datasets in FUnctional Mapping and Annotation (FUMA) [50]. For both the functional enrichment and tissue expression analyses, enrichment was assessed using a hypergeometric test and significant enrichment was defined as a Benjamini-Hochberg-adjusted $p < 0.05$, correcting for the number of tests performed within each analysis platform. Analyses were performed using web interfaces accessed on 18/04/2022 (WebGestalt and FUMA) and 14/01/2023 (FUMA).

Two-sample forward MR analyses

We performed two-sample forward MR analyses to identify potentially causal associations between genetically predicted plasma protein levels and: (i) cognitive function; (ii) structural brain phenotypes (total brain volume, cerebral white matter volume, hippocampal volume, WMH volume, and CMB); and (iii) disease outcomes (Alzheimer's disease, all stroke, stroke subtypes (ischaemic, cardioembolic, large artery, and small vessel), and intracranial aneurysm).

Associations between single nucleotide polymorphisms (SNPs) and plasma protein expression levels were calculated in PURE. SNPs located within 200 kilobases up- or downstream of the RefSeq transcript corresponding to a protein-of-interest were assessed as potential pQTLs through separate GWASs of the European (N = 3514), Latin (N = 4309), and Persian (N = 1332) participants, with significant association defined as $p < 5 \times 10^{-6}$. Missense variants and SNPs affecting splice sites were excluded. The GWAS model has been described previously [32]. Effect estimates were then combined by inverse variance-weighted fixed effects meta-analysis using METAL [51], and an independent set of pQTLs obtained by pruning ($r^2 < 0.1$ within the European, Latin, and Persian subgroups from the PURE cohort). Sensitivity analyses were performed in which the pruning threshold was adjusted to $r^2 < 0.01$.

The independent set of pQTLs were assessed for their associations with cognitive function, structural brain phenotypes, and disease outcomes using summary statistics from published studies [52-59].

MR analyses were performed using the R packages MRBase for TwoSample MR v.0.5.6 [60], mr.raps v.0.4.1 [61], and MRPRESSO v.1.0 [62]. We employed several complementary MR approaches: IVW [63], weighted median [64], robust adjusted profile scores (RAPS) [61], MR-Egger [65], and MR-PRESSO [62]. We adopted the IVW approach as our primary methodology and defined statistical significance using a liberal within-outcome variable Bonferroni correction for the two proteins (CA14 and CDCP1) that could be assessed, yielding a significance threshold of $p < 0.025$ (or $p < 0.05$ when an outcome could only be assessed for one protein). The IVW approach has the greatest statistical power but also makes the most assumptions. Hence, we reported IVW findings only where there was: (i) no evidence of pleiotropy; and (ii) corroboration of the direction of effect from at least two other MR approaches. An MR-Egger intercept $p < 0.05$ was deemed to indicate directional pleiotropy. Heterogeneity amongst instrumental variables, suggestive of horizontal pleiotropy, was indicated by a significant Cochran's Q ($p < 0.05$). If Cochran's Q was significant, MR-PRESSO [62] was performed, and, if the MR-PRESSO global test was significant ($p < 0.05$), MR-PRESSO with outlier removal was performed. In addition to the above conditions, we only reported results where there were at least three IVs, there was no evidence of weak instrument bias (*F*-statistic > 10) [66], and when the correction causal direction had been assessed (indicated by the instrumental variables explaining a greater proportion of the variance in the exposure than in the outcome, and a Steiger test $p < 0.05$). For the sensitivity analyses, in which a more stringent r^2 threshold was used to select independent pQTLs, only one or two IVs were available for each analysis. When two IVs were available, results from the IVW approach are reported, and when one IV was available, results from the Wald ratio test were reported.

Pairwise Conditional Analysis and Co-localisation Analysis (PWCoCo)

PWCoCo [30, 67] was performed to assess the existence of a shared causal variant between (i) pQTLs for each of the five proteins-of-interest and (ii) variants associated with the outcomes assessed in the two-sample MR analyses. PWCoCo analyses were performed for all conditionally distinct pQTLs and all conditionally distinct association signals in the outcome data. Analyses were performed using SNP-protein associations calculated in the European PURE-MIND participants (N = 3514). As for the MR analyses, SNPs had to be located within 200 kilobases up- or downstream of the RefSeq transcript corresponding to a protein-of-interest.

Analyses were performed using a C++ implementation of the PWCoCo algorithm, which utilises methods from the GCTA-COJO [68] and coloc [69] R packages. PWCoCo calculates the posterior probabilities (PP) for: the existence of no causal variant(s) for either trait (PP0); the existence of causal variant(s) for trait one or trait two (PP1 and 2,

respectively); both traits being associated with the same region, with different causal variants (PP3); and both traits being associated with the same region, with a shared causal variant(s) (PP4). Failure to find evidence in support of colocalisation can be due to lack of power [69]; therefore, we limited our analyses to those where PP3 + PP4 ≥ 0.8. Colocalisation was defined as PP4/PP3 > 5, as suggested previously [70].

Software

Statistical analyses and plot generation were performed in R (versions 3.6.0, 4.1.1, 4.1.2, 4.2.0, 4.2.1)

3. Results

Participants in PURE MIND (N=1198) and GS (N=1053) (**Table 1**) were similar in age, sex distribution, and clinical characteristics.

Identification of protein biomarkers of cognitive function and enrichment analyses

Five proteins were associated with DSST performance in PURE-MIND (**Figure 2**; **Figure 3**; **Supplementary Table 1**). Higher plasma levels of neurocan (NCAN; β = 2.03 (indicating a 2.03 higher DSST score per a standard deviation higher NCAN level), $p = 9.11 \times 10^{-8}$), brevican (BCAN; β = 1.91, $p = 5.56 \times 10^{-7}$), carbonic anhydrase 14 (CA14; β = 1.90, $p = 5.90 \times 10^{-7}$), and myelin-oligodendrocyte glycoprotein (MOG; β = 1.82, $p = 2.29 \times 10^{-6}$), and lower levels of CUB domain-containing protein 1 (CDCP1; β = -1.57, $p = 3.97 \times 10^{-5}$) were associated with better DSST performance, below the Bonferroni significance threshold. Adjustment for educational attainment modestly attenuated the effect estimate for all five proteins (**Supplementary Table 1**). Levels of NCAN, BCAN and MOG were positively correlated ($0.251 \leq r \geq 0.615$; all $p < 2.20 \times 10^{-16}$), whilst CDCP1 and CA14 expression levels were negatively correlated ($r = -1.01$, $p = 4.82 \times 10^{-4}$; **Supplementary Table 2**). Three proteins (NCAN, BCAN and CDCP1) proteins were also measured in GS, of which two replicated their association with DSST performance: NCAN (β = 1.40, $p = 1.07 \times 10^{-3}$) and CDCP1 (β = -1.99, $p = 9.21 \times 10^{-6}$; **Figure 3**). MoCA performance was not associated with the level of any protein (all $p \geq 2.34 \times 10^{-4}$; **Supplementary Table 3**).

Proteins nominally associated ($p < 0.05$) with DSST performance (N = 184) were enriched for brain expressed proteins, most significantly for proteins with hippocampal expression (FDR-corrected $p = 0.0154$ **Supplementary Table 4**). Better DSST performance was nominally associated with lower levels of 90 proteins. These proteins mapped to the following immune pathways “interleukin-10 signalling”, “glomerulonephritis”, “regulation of granulocyte chemotaxis”, “positive regulation of leukocyte chemotaxis”, “positive regulation of leukocyte migration”, and “inflammation” (FDR-corrected $p \leq 0.0337$; **Supplementary Table 5**).

Structural brain phenotypes as mediators of protein biomarker-DSST performance associations

In PURE-MIND, better DSST performance was associated with greater cerebral white matter volume (β = 0.0615, $p = 4.34 \times 10^{-7}$), greater total brain volume (β = 0.0349, $p = 9.64 \times 10^{-6}$), greater hippocampal volume (β = 2.97, $p = 4.79 \times 10^{-3}$), and lower log-transformed WMH volume (β = -3.20, $p = 1.18 \times 10^{-6}$). These associations replicated in GS (**Supplementary Table 6**).

Assessment of the relationships between protein levels and DSST-associated structural brain phenotypes in PURE-MIND revealed systematic differences between those proteins for which higher levels were associated with better DSST performance (NCAN, BCAN, CA14, and MOG), and CDCP1, which was negatively associated with DSST performance (**Figure 4**). Whilst NCAN, BCAN, CA14, and MOG showed a positive direction of association with total brain, cerebral white matter, and hippocampal volume measurements and a negative association with WMH volume, the converse was true for CDCP1. The associations between NCAN levels and total brain, cerebral white matter, and hippocampal volumes reached statistical significance ($p \leq 2.56 \times 10^{-5}$) and were replicated in GS ($p \leq 6.70 \times 10^{-3}$). BCAN levels were significantly associated with all four brain volumes ($p \leq 4.36 \times 10^{-4}$), with the associations with total brain and cerebral white matter volumes replicating in GS ($p \leq 3.44 \times 10^{-3}$). The associations between MOG levels and total brain and cerebral white matter volumes attained statistical significance ($p \leq 2.63 \times 10^{-9}$), but could not be assessed in GS. We did not identify any significant associations with CA14 or CDCP1 levels after correction for multiple testing.

In PURE-MIND, cerebral white matter volume explained a significant proportion of variance in the relationship between MOG (19%), BCAN (15%), and NCAN (13%) levels and DSST performance (all $p < 2 \times 10^{-16}$) (**Supplementary Table 7**). Log-transformed WMH volume was a significant partial mediator of the association between BCAN levels and DSST performance ($p = 0.002$), mediating 8% of the relationship.

Identification of potentially causal relationships between protein levels and cognitive function structural brain phenotypes, and disease outcomes

Inverse variance weighted (IVW) Mendelian randomisation (MR) analyses were performed to assess the effects of genetically predicted CA14 and CDCP1 levels on cognitive function, structural brain phenotypes, and Alzheimer's disease and stroke. A lack of instrumental variable (IV) SNPs precluded the assessment of BCAN, NCAN and MOG.

A one standard deviation higher level of genetically predicted plasma CA14 was associated with a larger hippocampal volume (β = 0.0990 [95% CI: 0.0272 to 0.171], $p = 6.87 \times 10^{-3}$), and a greater risk of all stroke (odds ratio (OR) = 1.07 [95% CI: 1.01 to 1.14], $p = 0.0153$; **Supplementary Table 8**). A one standard deviation higher level of genetically predicted plasma CDCP1 was associated with an increased risk of all stroke (OR = 1.12[95% CI: 1.03 to 1.22], $p = 0.0116$), ischaemic stroke (OR = 1.13[95% CI: 1.03 to 1.23], $p = 9.65 \times 10^{-3}$), and intracranial aneurysm (OR = 1.28 [95% CI: 1.06 to 1.55], $p = 9.84 \times 10^{-3}$). These associations were corroborated by similar effect estimates from weighted median and MR-RAPS analyses. No evidence of directional or horizontal pleiotropy were observed, and the correct causal direction was assessed. Neither the level of CA14 or CDCP1 was associated with risk of Alzheimer's disease ($p \geq 0.129$).

Sensitivity analyses were performed in which instrumental variables (IVs) were selected using a stricter threshold for independence. For genetically predicted CA14, these analyses supported the association with hippocampal volume (β = 0.144 [95% CI: 0.0435 to 0.244], $p = 4.97 \times 10^{-3}$), and produced a consistent, although non-significant, effect estimate for the association with risk for all stroke (**Supplementary Table 8**). For CDCP1, the sensitivity analyses demonstrated a consistent, although non-significant, effect estimate for the association with risk for intracranial aneurysm. The associations between genetically predicted CDCP1 and risk for (i) all stroke and (ii) ischaemic stroke could not be meaningfully interpreted due to heterogeneity between the two available IVs (Q-test $p \leq 0.0187$).

Pairwise Conditional Analysis and Co-localisation Analyses (PWCoCo) were performed to assess the presence of a shared variant for each of the five proteins-of-interest and the same outcomes as assessed by two-sample MR analyses. We were only adequately powered to assess co-localisation between SNPs associated with one pair of traits: MOG plasma level and cognitive function. We did not observe any evidence in support of co-localisation or conditional co-localisation (posterior probability (PP)4/PP3 ≤ 4.81×10^{-4}).

4. Discussion

In this large-scale analysis of the associations between the plasma levels of 1160 proteins and cognitive function, we identify CA14 and CDCP1 as being associated with processing speed, as measured by the DSST, and having potentially causal effects on hippocampal volume (CA14), and risk of stroke (both) and intracranial aneurysm (CDCP1).

Other proteins (BCAN, NCAN, and MOG) were associated with DSST performance and important structural brain phenotypes, with cerebral white matter volume mediating a significant proportion (13-19%) of the relationship between the levels of all three proteins and DSST performance, and WMH volume mediating 8% of the relationship between BCAN levels and DSST performance. Potentially causal effects of these proteins could not be assessed due to a lack of genetic instruments. Enrichment analyses of proteins that were nominally significantly associated with DSST performance revealed a significant enrichment for brain-expressed proteins.

There were no significant associations between plasma protein levels and performance on the MoCA. This might reflect the fact that the MoCA is a screening tool for mild cognitive impairment [10], meaning its sensitivity to detect variation in cognitive function in non-clinical groups is likely to be limited.

CA14 is one of fifteen isoforms of the carbonic anhydrase family of zinc metalloprotease enzymes, which catalyse the reversible hydration of carbon dioxide [11]. CA14 is expressed by neurons [12] and involved in regulating extracellular pH following synaptic transmission [13, 14]. Consistent with our findings, acute inhibition of CA14 leads to impaired performance on cognitive tasks in mice [15]. Carbonic anhydrase activation may lead to beneficial cognitive effects in rodents [16]. In keeping with our MR results, there are neuroprotective effects of carbonic anhydrase inhibition in models of amyloidosis, Huntington's disease, and ischaemic and haemorrhagic stroke [16]. The mechanisms by which carbonic anhydrase inhibition and activation exert their effects are uncertain [15, 16]. FDA-approved carbonic anhydrase inhibitors, and thus the majority of carbonic anhydrase inhibitors investigated to date, are pan-carbonic anhydrase inhibitors. Of the carbonic anhydrase family members measured in our study (CA1, 2, 3, 4, 5A, 6, 9, 12, 13, and 14), only CA14 levels were significantly associated with DSST performance. Further studies are required to determine the therapeutic potential for carbonic anhydrase modulation in the context of cognitive impairment, Alzheimer's disease, and stroke.

The extracellular matrix (ECM) proteins NCAN and BCAN are brain-specific chondroitin sulfate proteoglycans, which are expressed by neurons and astrocytes (NCAN and BCAN), and oligodendrocytes (BCAN). They contribute to the formation of a specialised structure, the perineuronal net (PNN), which plays a key role in memory and neuronal plasticity, and which is disrupted in Alzheimer's disease [17]. Our findings are consistent with those of Harris et al. (2020) [5], who found plasma levels of NCAN and BCAN were positively associated with brain volume. Plasma levels of BCAN have previously been found to be positively associated with Mini Mental State Examination performance and reduced in patients with Alzheimer's disease or mild cognitive impairment [7]. Mice that are lacking either NCAN or BCAN expression show normal development and memory function but reduced hippocampal long term potentiation [18, 19], whilst quadruple knock-outs, which lack NCAN, BCAN, and two additional ECM proteins (tenascin-C and tenascin-R) show an altered ratio of excitatory to inhibitory synapses and a reduction in the number and complexity of hippocampal PNNs [20]. Genetic variation in the gene encoding A Disintegrin and Metalloproteinase with Thrombospondin Motifs 4 (ADAMTS4), which degrades the four members of the lectican family (including NCAN and BCAN), has been implicated in Alzheimer's disease [21]. Taken together, the evidence suggests NCAN, BCAN and their regulators as molecules-of-interest in Alzheimer's disease.

MOG is an oligodendrocyte-expressed membrane glycoprotein, the exact function of which is unknown [22].

CDCP1 is a widely expressed transmembrane glycoprotein that acts as a ligand for T cell-expressed Cluster of Differentiation 6 (CD6), and is implicated in autoimmune conditions [23]. CDCP1 is amenable to modulation by approved drug treatments: Itolizumab, which is used to treat psoriasis, disrupts CDCP1-CD6 binding and downregulates T-cell-mediated inflammation [24], whilst atomoxetine, a treatment for attention deficit hyperactivity disorder, which is being considered for the treatment of mild cognitive impairment, reduced cerebrospinal fluid (CSF) CDCP1 levels [25]. Intriguingly, findings in mice suggest a functional link between CDCP1 and MOG [6].

Our study has several strengths. We measured 1160 proteins, associated with a wide range of physiological processes, in a large, well-characterised cohort. Replication analyses, where possible, were performed in an independent cohort in which proteins were measured using an independent methodology. The availability of genetic and brain MRI data permitted an exploration of causality and putative causal pathways. The use of MR to identify potentially causal associations will have offered protection against some of the common confounders of observational analyses [26], with the use of multiple MR methods, which generally gave concordant estimates of effect, mitigating against the individual biases of different MR methodologies [27]. Moreover, by requiring instrumental variables to be located in *cis* to their target protein, we limited the chance of pleiotropic effects [28].

There are also several limitations to consider.

First, the 1160 proteins measured represent a small subset of the circulating proteome [29]. Replication analyses were only performed for those proteins for which data were available in the GS cohort, meaning that we did not assess replication of CA14 or MOG.

Second, the availability of suitable IVs mean that our primary MR analyses were only performed for CA14 and CDCP1. Whilst we required a minimum of three IVs for the primary MR analyses, our sensitivity analyses, in which a stricter threshold for independence was applied to the IVs necessitated the use of fewer than three IVs in each analysis. As such, the results of the sensitivity analyses should be interpreted with this caveat in mind.

Third, for all but one pair of traits, we were insufficiently powered to assess co-localisation between genetic variants associated with protein level and cognition, structural brain phenotypes, and disease outcomes. This means that it is possible that significant MR findings might reflect the presence of separate causal variants in linkage disequilibrium (LD) with one another [30]

Fourth, we measured protein levels in the plasma, rather than in the brain or CSF. It is, however, important to note the striking enrichment for brain-expressed proteins amongst the DSST-associated proteins. Previous analyses of the GS cohort, in which replication was sought in the present study, have identified the levels of several plasma proteins as being associated with multiple markers of brain health [8]. These findings support the use of the plasma to assess brain-related phenotypes and emphasise the need for additional research to explain the mechanisms controlling the efflux of brain-expressed proteins into the bloodstream in non-clinical populations. Moreover, the use of *cis* pQTLs, which are likely to be shared across tissues [31], as IVs in our MR analyses, supports the possibility that the MR-identified associations reflect the actions of the proteins-of-interest in the brain.

5. Conclusions

We identified protein biomarkers of cognitive function that may causally affect brain structure and risk for stroke and intracranial aneurysm. Notwithstanding the need for replication, our findings prompt several hypotheses that should be assessed by future studies. Our apparently paradoxical findings of higher CA14 levels being associated with both better cognitive function and increased stroke risk suggest that molecular findings can inform a more nuanced understanding of the relationship between premorbid cognitive function and neurological disease risk. It is possible that improved risk stratification may be achieved through the combination of cognitive assessment and biomarker measurement. The availability of approved drugs targeting our identified proteins raises the possibility of drug repurposing for novel therapeutic interventions to prevent cognitive decline, stroke, and intracranial aneurysm.

Abbreviations: Average causal mediation effect (ACME); brevican (BCAN); carbonic anhydrase 14 (CA14); cluster of differentiation 6 (CD6); CUB-domain containing protein 1 (CDCP1); confidence interval (CI); cerebral microbleed (CMB); cerebrospinal fluid (CSF); digit symbol substitution test (DSST); extracellular matrix (ECM); false discovery rate (FDR); Generation Scotland imaging subsample (GS); Generation Scotland: Scottish Family Health Study (GS:SFHS); instrumental variable (IV); inverse variance weighted (IVW); myelin oligodendrocyte glycoprotein (MOG); Montreal Cognitive Assessment (MoCA); Mendelian randomisation (MR); magnetic resonance imaging (MRI); neurocan (NCAN); perineuronal net (PNN); odds ratio (OR); posterior probability (PP); protein quantitative trait loci (pQTL); pairwise conditional analysis and co-localisation analyses (PWCoCo); Prospective Urban and Rural Epidemiology (PURE); robust adjusted profile score (RAPS); silent brain infarct (SBI); standard deviation (SD); small vessel disease (SVD); white matter hyperintensity (WMH)

Declarations

Ethics approval and consent to participate

All centres contributing to PURE were required to obtain approval from their respective ethics committees (Institutional Review Boards). Participant data is confidential and only authorized individuals can access study-related documents. The participants' identities are protected in documents transmitted to the Coordinating Office, as well as biomarker and genetic data. Participants provided informed consent to obtain baseline information, and to collect and store genetic and other biological specimens.

All components of Generation Scotland received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89). All participants provided broad and enduring written informed consent for biomedical research. Generation Scotland has also been granted Research Tissue Bank status by the East of Scotland Research Ethics Service (REC Reference Number: 15/0040/ES), providing generic ethical approval for a wide range of uses within medical research. . The imaging subsample of Generation Scotland received ethical approval from the NHS Tayside committee on research ethics (reference 14/SS/0039). This study was performed in accordance with the Helsinki declaration.

Consent for publication

Not applicable

Availability of data and materials

The PURE datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

According to the terms of consent for GS participants, access to individual-level data (omics and phenotypes) must be reviewed by the GS Access Committee. Applications should be made to access@generationscotland.org.

Competing interests

MC is supported by a Canadian Institute of Health Research doctoral award and has received consulting fees from Bayer AG. MP is supported by the EJ Moran Campbell Internal Career Research Award from McMaster University. DAG is a part-time employee of Optima partners, a health data consultancy based at the Bayes centre, The University of Edinburgh. SH is an employee of Bayer AG. AMM has previously received speaker's fees from Illumina and Janssen and research grant funding from The Sackler Trust. SY is supported by the Heart and Stroke Foundation/Marion W Burke Chair in Cardiovascular Disease. GP is supported by the CISCO Professorship in Integrated Health Systems. The other authors declare no competing interests.

Funding

PURE

The PURE study is an investigator-initiated study that is funded by the Population Health Research Institute, the Canadian Institutes of Health Research (CIHR), Heart and Stroke Foundation of Ontario, support from CIHR's Strategy for Patient Oriented Research, through the Ontario SPOR Support Unit, as well as the Ontario Ministry of Health and Long-Term Care and through unrestricted grants from several pharmaceutical companies (with major contributions from AstraZeneca [Canada], Sanofi-Aventis [France and Canada], Boehringer Ingelheim [Germany and Canada], Servier, and GlaxoSmithKline), and additional contributions from Novartis and King Pharma and from various national or local organisations in participating countries as follows: Argentina—Fundacion ECLA; Bangladesh—Independent University, Bangladesh and Mitra and Associates; Brazil—Unilever Health Institute, Brazil; Canada—Public Health Agency of Canada and Champlain Cardiovascular Disease Prevention Network; Chile—Universidad de la Frontera; Colombia—Colciencias (grant number 6566–04–18062); South Africa—The North-West University, SANPAD (SA and Netherlands Programme for Alternative Development), National Research Foundation, Medical Research Council of South Africa, The South Africa Sugar Association, Faculty of Community and Health Sciences; Sweden—grants from the Swedish State under the Agreement concerning research and education of doctors, the Swedish Heart and Lung

Foundation, the Swedish Research Council, the Swedish Council for Health, Working Life and Welfare, King Gustaf V's and Queen Victoria Freemasons Foundation, AFA Insurance, Swedish Council for Working Life and Social Research, Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning, grant from the Swedish State under (LäkarUtbildningsAvtalet) Agreement, and grant from the Västra Götaland Region; and United Arab Emirates—Sheikh Hamdan Bin Rashid Al Maktoum Award for Medical Sciences, Dubai Health Authority, Dubai. The PURE biomarker project was supported by Bayer and the CIHR. The biomarker project was led by PURE investigators at the Population Health Research Institute (Hamilton, Canada) in collaboration with Bayer scientists. Bayer directly compensated the Population Health Research Institute for measurement of the biomarker panels, scientific, methodological, and statistical work. Genetic analyses were supported by CIHR (G-18–0022359) and Heart and Stroke Foundation of Canada (application number 399497) in the form of funding to GP.

GS

This work was supported by the Wellcome Trust [104036/Z/14/Z, 220857/Z/20/Z, and 216767/Z/19/Z] and an MRC Mental Health Data Pathfinder Grant [MC_PC_17209] to AMM. DAG is funded by the Wellcome Trust Translational Neuroscience PhD Programme at the University of Edinburgh [108890/Z/15/Z]. LS and ANH are supported by Medical Research Council [MR/L023784/2]: Dementias Platform UK. LS is also supported by a Medical Research Council Award to the University of Oxford [MC_PC_17215]. AS is supported through the Wellcome-University of Edinburgh Institutional Strategic Support Fund (Reference 204804/Z/16/Z), and indirectly through the Lister Institute of Preventive Medicine award with reference 173096. 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 of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the Clinical Research Facility, Edinburgh, Scotland, and was funded by the UK's Medical Research Council and the Wellcome Trust [104036/Z/14/Z].

Authors' contributions

Conception and design: RMW, WNW, GP; data analysis: RMW, MC, NP; drafting the article: RMW, WNW, GP; data preparation: MC, NP, MP, DAG, AC, HCW, AK, LS, XS, EE, MOD; data collection: ANH, AMM, EE, SH, SR, MOD, SY, WNW, GP; revision of the article: RMW, MC, NP, MP; WNW, GP; all authors read and approved the final manuscript.

Acknowledgements

We are grateful to all the families who took part in Generation Scotland, the general practitioners and the Scottish School of Primary Care for their help in recruiting them, and the whole Generation Scotland team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, healthcare assistants and nurses.

We thank Dr Alison Offer for assistance in producing the forest plots.

References

1. Martin, G.M., *Defeating Dementia.* Nature, 2004. **431**: p. 247-248.
2. Rostamian, S., et al., *Cognitive impairment and risk of stroke: a systematic review and meta-analysis of prospective cohort studies.* Stroke, 2014. **45**(5): p. 1342-8.
3. Valenzuela, M.J. and P. Sachdev, *Brain reserve and dementia: a systematic review.* Psychol Med, 2006. **36**(4): p. 441-54.
4. Pettigrew, C. and A. Soldan, *Defining Cognitive Reserve and Implications for Cognitive Aging.* Curr Neurol Neurosci Rep, 2019. **19**(1): p. 1.
5. Harris, S.E., et al., *Neurology-related protein biomarkers are associated with cognitive ability and brain volume in older age.* Nat Commun, 2020. **11**(1): p. 800.
6. Lindbohm, J.V., et al., *Plasma proteins, cognitive decline, and 20-year risk of dementia in the Whitehall II and Atherosclerosis Risk in Communities studies.* Alzheimers Dement, 2022. **18**(4): p. 612-624.
7. Whelan, C.D., et al., *Multiplex proteomics identifies novel CSF and plasma biomarkers of early Alzheimer's disease.* Acta Neuropathol Commun, 2019. **7**(1): p. 169.
8. Gadd, D.A., et al., *Integrated methylome and phenome study of the circulating proteome reveals markers pertinent to brain health.* Nat Commun, 2022. **13**(1): p. 4670.
9. Smith, E.E., et al., *Early cerebral small vessel disease and brain volume, cognition, and gait.* Ann Neurol, 2015. **77**(2): p. 251-61.
10. Nasreddine, Z.S., et al., *The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment.* J Am Geriatr Soc, 2005. **53**(4): p. 695-9.
11. Lindskog, S., *Structure and mechanism of carbonic anhydrase.* Pharmacol Ther, 1997. **74**(1): p. 1-20.
12. Parkkila, S., et al., *Expression of membrane-associated carbonic anhydrase XIV on neurons and axons in mouse and human brain.* Proc Natl Acad Sci U S A, 2001. **98**(4): p. 1918-23.
13. Chen, J.C. and M. Chesler, *pH transients evoked by excitatory synaptic transmission are increased by inhibition of extracellular carbonic anhydrase.* Proc Natl Acad Sci U S A, 1992. **89**(16): p. 7786-90.
14. Shah, G.N., et al., *Carbonic anhydrase IV and XIV knockout mice: roles of the respective carbonic anhydrases in buffering the extracellular space in brain.* Proc Natl Acad Sci U S A, 2005. **102**(46): p. 16771-6.
15. Provensi, G., et al., *A New Kid on the Block? Carbonic Anhydrases as Possible New Targets in Alzheimer's Disease.* Int J Mol Sci, 2019. **20**(19).
16. Lemon, N., et al., *Carbonic Anhydrases as Potential Targets Against Neurovascular Unit Dysfunction in Alzheimer's Disease and Stroke.* Front Aging Neurosci, 2021. **13**: p. 772278.
17. Sorg, B.A., et al., *Casting a Wide Net: Role of Perineuronal Nets in Neural Plasticity.* J Neurosci, 2016. **36**(45): p. 11459-11468.
18. Brakebusch, C., et al., *Brevican-deficient mice display impaired hippocampal CA1 long-term potentiation but show no obvious deficits in learning and memory.* Mol Cell Biol, 2002. **22**(21): p. 7417-27.
19. Zhou, X.H., et al., *Neurocan is dispensable for brain development.* Mol Cell Biol, 2001. **21**(17): p. 5970-8.
20. Gottschling, C., et al., *Elimination of the four extracellular matrix molecules tenascin-C, tenascin-R, brevican and neurocan alters the ratio of excitatory and inhibitory synapses.* Sci Rep, 2019. **9**(1): p. 13939.
21. Marioni, R.E., et al., *GWAS on family history of Alzheimer's disease.* Transl Psychiatry, 2018. **8**(1): p. 99.
22. Peschl, P., et al., *Myelin Oligodendrocyte Glycoprotein: Deciphering a Target in Inflammatory Demyelinating Diseases.* Front Immunol, 2017. **8**: p. 529.
23. Enyindah-Asonye, G., et al., *CD318 is a ligand for CD6.* Proc Natl Acad Sci U S A, 2017. **114**(33): p. E6912-E6921.
24. Dogra, S., S. Uprety, and S.H. Suresh, *Itolizumab, a novel anti-CD6 monoclonal antibody: a safe and efficacious biologic agent for management of psoriasis.* Expert Opin Biol Ther, 2017. **17**(3): p. 395-402.
25. Levey, A.I., et al., *A phase II study repurposing atomoxetine for neuroprotection in mild cognitive impairment.* Brain, 2022. **145**(6): p. 1924-1938.
26. Lawlor, D.A., et al., *Mendelian randomization: using genes as instruments for making causal inferences in epidemiology.* Stat Med, 2008. **27**(8): p. 1133-63.
27. Slob, E.A.W. and S. Burgess, *A comparison of robust Mendelian randomization methods using summary data.* Genet Epidemiol, 2020. **44**(4): p. 313-329.
28. Swerdlow, D.I., et al., *Selecting instruments for Mendelian randomization in the wake of genome-wide association studies.* Int J Epidemiol, 2016. **45**(5): p. 1600-1616.
29. Omenn, G.S., et al., *Progress on Identifying and Characterizing the Human Proteome: 2018 Metrics from the HUPO Human Proteome Project.* J Proteome Res, 2018. **17**(12): p. 4031-4041.

30. Zheng, J., et al., *Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases.* Nat Genet, 2020. **52**(10): p. 1122-1131.
31. Yang, C., et al., *Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders.* Nat Neurosci, 2021. **24**(9): p. 1302-1312.
32. Narula, S., et al., *Plasma ACE2 and risk of death or cardiometabolic diseases: a case-cohort analysis.* Lancet, 2020. **396**(10256): p. 968-976.
33. Teo, K., et al., *The Prospective Urban Rural Epidemiology (PURE) study: examining the impact of societal influences on chronic noncommunicable diseases in low-, middle-, and high-income countries.* Am Heart J, 2009. **158**(1): p. 1-7 e1.
34. Habota, T., et al., *Cohort profile for the STratifying Resilience and Depression Longitudinally (STRADL) study: A depression-focused investigation of Generation Scotland, using detailed clinical, cognitive, and neuroimaging assessments.* Wellcome Open Res, 2021. **4**: p. 185.
35. 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, 2018. **47**(1): p. 13-14g.
36. Smith, B.H., et al., *Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness.* Int J Epidemiol, 2013. **42**(3): p. 689-700.
37. Smith, B.H., et al., *Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability.* BMC Med Genet, 2006. **7**: p. 74.
38. Schwarz, C.G., et al., *A large-scale comparison of cortical thickness and volume methods for measuring Alzheimer's disease severity.* Neuroimage Clin, 2016. **11**: p. 802-812.
39. Schmidt, P., et al., *An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis.* Neuroimage, 2012. **59**(4): p. 3774-83.
40. Cox, S.R., et al., *Associations between vascular risk factors and brain MRI indices in UK Biobank.* Eur Heart J, 2019. **40**(28): p. 2290-2300.
41. Desikan, R.S., et al., *An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest.* Neuroimage, 2006. **31**(3): p. 968-80.
42. Fischl, B., *FreeSurfer.* Neuroimage, 2012. **62**(2): p. 774-81.
43. Wechsler, D., *Wechsler Adult Intelligence Scale-Third Edition (WAIS-III)*. 1997, San Antonio: Harcourt Assessment Inc.
44. Shi, L., et al., *Plasma Proteomic Biomarkers Relating to Alzheimer's Disease: A Meta-Analysis Based on Our Own Studies.* Front Aging Neurosci, 2021. **13**: p. 712545.
45. Das, S., et al., *Next-generation genotype imputation service and methods.* Nat Genet, 2016. **48**(10): p. 1284-1287.
46. Loh, P.R., et al., *Reference-based phasing using the Haplotype Reference Consortium panel.* Nat Genet, 2016. **48**(11): p. 1443-1448.
47. Therneau, T.M., *coxme: mixed effects Cox models.* 2012. p. R Package.
48. Tingley, D., et al., *mediation: R Package for Causal Mediation Analysis.* Journal of Statistical Software, 2014. **59**(5).
49. Liao, Y., et al., *WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs.* Nucleic Acids Res, 2019. **47**(W1): p. W199-W205.
50. Watanabe, K., et al., *Functional mapping and annotation of genetic associations with FUMA.* Nat Commun, 2017. **8**(1): p. 1826.
51. Willer, C.J., Y. Li, and G.R. Abecasis, *METAL: fast and efficient meta-analysis of genomewide association scans.* Bioinformatics, 2010. **26**(17): p. 2190-1.
52. Savage, J.E., et al., *Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence.* Nat Genet, 2018. **50**(7): p. 912-919.
53. Smith, S.M., et al., *An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank.* Nat Neurosci, 2021. **24**(5): p. 737-745.
54. Hibar, D.P., et al., *Novel genetic loci associated with hippocampal volume.* Nat Commun, 2017. **8**: p. 13624.
55. Persyn, E., et al., *Genome-wide association study of MRI markers of cerebral small vessel disease in 42,310 participants.* Nat Commun, 2020. **11**(1): p. 2175.
56. Knol, M.J., et al., *Association of common genetic variants with brain microbleeds: A genome-wide association study.* Neurology, 2020. **95**(24): p. e3331-e3343.

57. Bellenguez, C., et al., *New insights into the genetic etiology of Alzheimer's disease and related dementias.* Nat Genet, 2022. **54**(4): p. 412-436.
58. Mishra, A., et al., *Stroke genetics informs drug discovery and risk prediction across ancestries.* Nature, 2022. **611**(7934): p. 115-123.
59. Bakker, M.K., et al., *Genome-wide association study of intracranial aneurysms identifies 17 risk loci and genetic overlap with clinical risk factors.* Nat Genet, 2020. **52**(12): p. 1303-1313.
60. Hemani, G., et al., *The MR-Base platform supports systematic causal inference across the human phenome.* Elife, 2018. **7**.
61. Zhao, Q., et al., *Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score.* The Annals of Statistics, 2020. **48**(3): p. 1742-1769, 28.
62. Verbanck, M., et al., *Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases.* Nat Genet, 2018. **50**(5): p. 693-698.
63. Burgess, S., A. Butterworth, and S.G. Thompson, *Mendelian randomization analysis with multiple genetic variants using summarized data.* Genet Epidemiol, 2013. **37**(7): p. 658-65.
64. Bowden, J., et al., *Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator.* Genet Epidemiol, 2016. **40**(4): p. 304-14.
65. Burgess, S. and S.G. Thompson, *Interpreting findings from Mendelian randomization using the MR-Egger method.* Eur J Epidemiol, 2017. **32**(5): p. 377-389.
66. Staiger, D. and J.H. Stock, *Instrumental Variables Regression with Weak Instruments.* Econometrica, 1997. **65**(3): p. 557-586.
67. Robinson, J.W., et al., *An efficient and robust tool for colocalisation: Pair-wise Conditional and Colocalisation (PWCoCo).* bioRxiv, 2022: p. 2022.08.08.503158.
68. Yang, J., et al., *GCTA: a tool for genome-wide complex trait analysis.* Am J Hum Genet, 2011. **88**(1): p. 76-82.
69. Giambartolomei, C., et al., *Bayesian test for colocalisation between pairs of genetic association studies using summary statistics.* PLoS Genet, 2014. **10**(5): p. e1004383.
70. Guo, H., et al., *Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases.* Hum Mol Genet, 2015. **24**(12): p. 3305-13.

Tables and figures

Table 1. Demographic information for the discovery sample (PURE-MIND) and the replication sample (GS). Basic sample demographic information are presented together with information on relevant clinical characteristics, cognitive performance, and structural brain imaging measures. The number of participants for whom information was available for each variable are indicated in the "N" columns. In GS, there were participants with missing information for smoking (N = 31), hypertension (N = 2), diabetes (N = 1), stroke (N = 1), and depression (N = 2). *PURE-MIND: college or university; GS: university

	PURE-MIND		GS	
	N (%)	Mean (SD)	N (%)	Mean (SD)
Age (years)	1198	54.5 (8.05)	1053	59.9 (9.59)
Sex				
Female	709 (59.2%)		627 (59.5%)	
Male	489 (40.8%)		426 (40.5%)	
Education (highest level achieved)				
High school or less	346 (28.9%)		316 (29.9%)	
Trade school	83 (6.93%)		328 (31.1%)	
College/university*	768 (64.1%)		295 (28.0%)	
Other	NA		21 (1.99%)	
Unknown	1 (0.0835%)		94 (8.93%)	
Clinical characteristics				
Current/former smoker	583 (48.7%)		466 (45.6%)	
Non-smoker	615 (51.3%)		556 (54.4%)	
Body mass index (kg/m^2)	1193	27.1 (5.15)	1053	28.2 (5.72)
Waist hip ratio	1192	0.870 (0.0965)		
Systolic blood pressure	1197	130 (18.8)	1051	141 (18.9)
Diastolic blood pressure	1197	80.9 (11.3)	1051	82.4 (10.8)
Hypertension	270 (22.5%)			
Total cholesterol (mmol/L)	1197	5.52 (1.11)		
High density lipoprotein cholesterol (mmol/L)	1197	1.47 (0.399)		
Low density lipoprotein cholesterol (mmol/L)	1187	3.35 (0.882)		
Diabetes	61 (5.09%)		73 (6.93%)	
Stroke	3 (0.250%)		27 (2.56%)	
Cardiovascular disease	36 (3.00%)		103 (9.78%)	
Depression	NA		336 (31.9%)	
Cognitive tests				
Digit Symbol Substitution Test (no. correct in two minutes)	1198	69.5 (15.3)	1053	68.1 (15.2)
Montreal Cognitive Assessment (no. items correct)	1198	26.5 (2.39)		
Brain Imaging Measures				
Cortical thickness (mm)	1198	2.35 (0.0891)		
Total brain volume without ventricles (cm^3)	1198	1058 (110)	943	1069 (109)
Total cerebral white matter volume (cm^3)	1198	447 (60.5)	939	455 (56.9)
Total hippocampal volume (cm^3)	1198	3.92 (0.433)	941	4.18 (0.437)
WMH volume (log-transformed cm^3)	1198	0.751 (0.693)	934	0.777 (0.789)
Estimated intracranial volume (cm^3)	1198	1491 (159)	944	1400 (225)
Silent brain infarct	95 (7.93%)			
Cerebral microbleeds	86 (7.18%)			

Abbreviations: GS: Generation Scotland; PURE: Prospective Urban and Rural Epidemiology study; SD: standard deviation; WMH: white matter hyperintensity

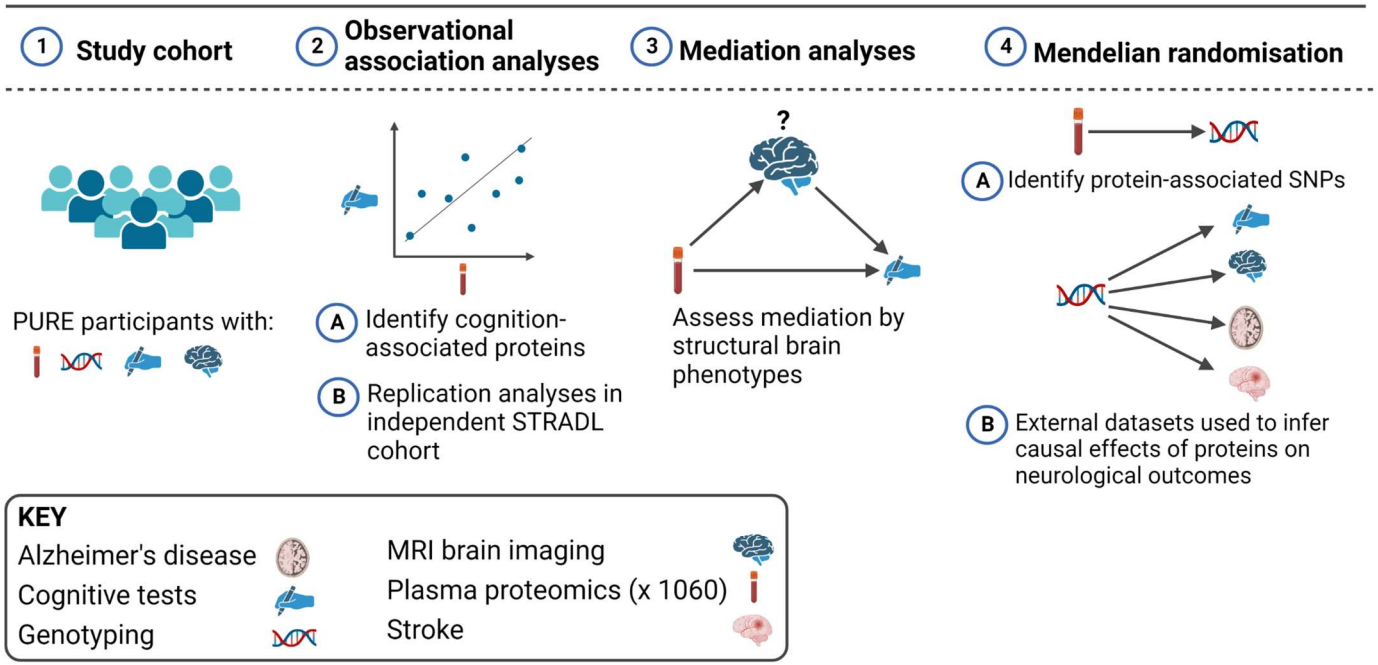

Figure 1. Overview of the study design. This study involved European (N = 3514), Latin (N = 4309), and Persian (N = 1332) PURE participants for whom genetic and plasma proteomic data were available. Observational analyses to detect plasma biomarkers of cognitive function were performed in the subset of these participants who were enrolled in the PURE-MIND sub-study (N = 1198), for whom plasma protein (N = 1060 proteins) and MRI measurements were available. Mediation analyses were performed to assess whether any observed associations between protein levels and cognitive function were mediated by structural brain phenotypes ascertained by MRI. Finally, two-sample Mendelian randomisation analyses were performed to assess potentially causal effects of genetically-predicted cognition-associated protein levels on genetically-predicted neurological outcomes. For these analyses, genetic instrumental variables for protein levels were identified in the European, Latin, and Persian PURE participants, and associations with neurological outcomes were assessed using external (non-PURE) datasets. Created with BioRender.com.

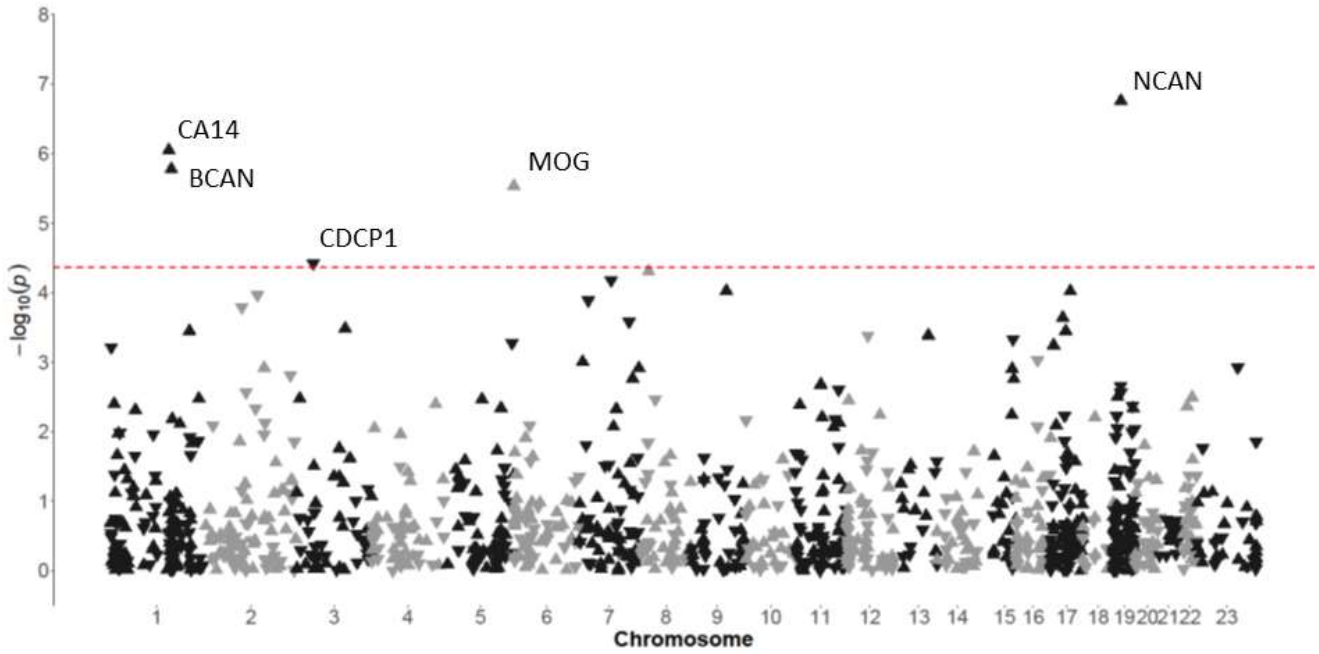

Figure 2. Manhattan plot indicating associations between the levels of plasma proteins and performance on the DSST in participants from the PURE-MIND cohort (N = 1198). Each protein is represented by a triangle with upwards-facing triangles indicating a positive association with DSST performance and downwards-facing triangles indicating a negative association with DSST performance. The position of each protein on the x-axis is determined by the genomic location of its corresponding gene and the position on the y-axis is determined by the $-\log_{10}$ *p*-value. The dashed horizontal line indicates the Bonferroni-corrected significance threshold ($p = 4.31 \times 10^{-5}$) required to maintain a 5% type I error rate.

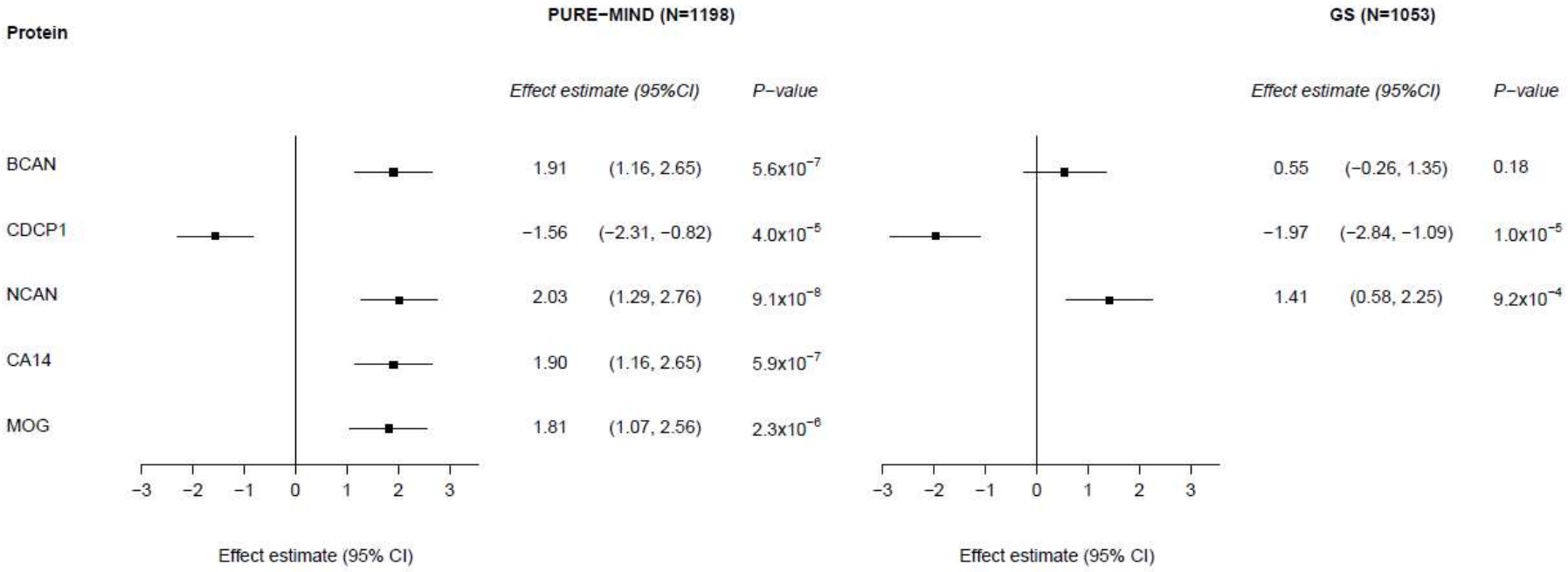

Figure 3. Forest plot indicating the association between protein levels and DSST performance for significantly associated proteins. For each protein, the difference in DSST score associated with a standard deviation higher level of protein is shown, together with the 95% confidence interval. Abbreviations: BCAN: brevican; CA14: carbonic anhydrase 14; CDCP1: CUB-domain containing protein 1; CI: confidence interval; GS: Generation Scotland imaging subsample; MOG: myelin oligodendrocyte glycoprotein; NCAN: neurocan; PURE: Prospective Urban and Rural Epidemiology study.

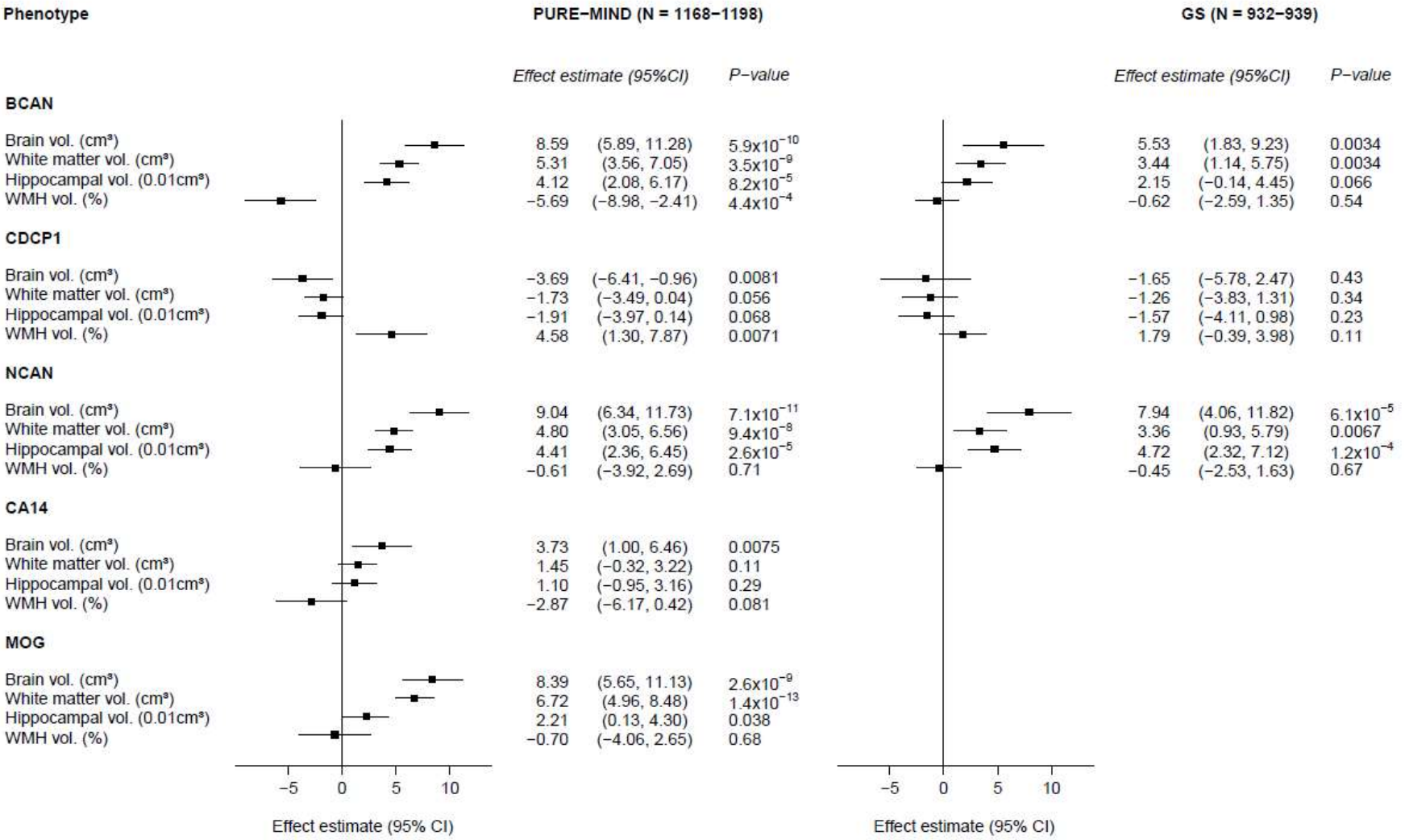

Figure 4. Forest plots indicating the association between the levels of DSST-associated proteins and DSST-associated structural brain phenotype. For each protein, the effect estimate (change in brain volume (cm^3 or %) per standard deviation increase in protein expression) is shown, together with the 95% confidence interval. Abbreviations: BCAN: brevican; CA14: carbonic anhydrase 14; CDCP1: CUB-domain containing protein 1; CI: confidence interval; GS: Generation Scotland imaging subsample; MOG: myelin oligodendrocyte glycoprotein; NCAN: neurocan; PURE: Prospective Urban and Rural Epidemiology study; WMH: white matter hyperintensity