

# Unraveling the role of proteins in dementia: insights from two UK cohorts with causal evidence

Jessica Gong<sup>1,2</sup>, Dylan M. Williams<sup>3</sup>, Shaun Scholes<sup>1</sup>, Sarah Assaad<sup>1</sup>, Feifei Bu<sup>4</sup>, Shabina Hayat<sup>1</sup>, Paola Zaninotto<sup>1</sup>, Andrew Steptoe<sup>4</sup>

## Author affiliations:

<sup>1</sup> Department of Epidemiology and Public Health, University College London, WC1E 7HB, United Kingdom

<sup>2</sup> George Institute for Global Health, Imperial College London, W12 7RZ, United Kingdom

<sup>3</sup> MRC Unit for Lifelong Health & Ageing, University College London, WC1E 7HB, United Kingdom

<sup>4</sup> Department of Behavioural Science and Health, University College London, WC1E 7HB, United Kingdom

## Correspondence to: Dr Jessica Gong

Full address: Department of Epidemiology and Public Health, University College London, 1-19 Torrington Pl, London, WC1E 7HB, United Kingdom

Email: [jessica.gong@ucl.ac.uk](mailto:jessica.gong@ucl.ac.uk)

## Abstract

Population-based proteomics offers a groundbreaking avenue to predict future disease risks, enhance our understanding of disease mechanisms, and discover novel therapeutic targets and biomarkers. The role of plasma proteins in dementia, however, requires further exploration. This study investigated 276 protein-dementia associations in 229 incident all-cause dementia, 89 Alzheimer's disease, and 41 vascular dementia among 3,249 participants (55% women, 97.2% white ethnicity) from the English Longitudinal Study of Ageing (ELSA) over a median 9.8-year follow-up. We used Cox proportional hazard regression for the analysis. Receiver operating characteristic analyses were conducted to assess the precision of the identified proteins from the fully adjusted Cox regression models in predicting incident all-cause dementia, both individually and in combination with demographic predictors, *APOE* genotype, and memory score, to estimate the area under the curve. Additionally, the eXtreme Gradient Boosting machine learning algorithm was used to identify the most important features predictive of future all-cause dementia onset. These associations were then validated in 1,506 incident all-cause dementia, 732 Alzheimer's disease, 281 vascular dementia, and 111 frontotemporal dementia cases among 52,745 individuals (53.9% women, 93.3% white ethnicity) from the UK Biobank over a median 13.7-year follow-up. Two-sample bi-directional Mendelian randomization and drug target Mendelian randomization were further employed to determine the causal direction between protein concentration and dementia. NEFL (Hazard ratio [HR] [95% confidence intervals (CI)]: 1.54 [1.29, 1.84]) and RPS6KB1 (HR [95% CI]: 1.33 [1.16, 1.52]) were robustly associated with incident all-cause dementia; MMP12 (HR [95% CI]: 2.06 [1.41, 2.99]) was associated with vascular dementia in ELSA, after correcting for multiple testing. Additional markers EDA2R and KIM1 were identified from subgroup and sensitivity analyses. Combining NEFL and RPS6KB1 with other predictors yielded high predictive accuracy (area under the curve=0.871) for incident all-cause dementia.

1  
2  
3 The eXtreme Gradient Boosting machine learning algorithm also identified RPS6KB1, NEFL,  
4 and KIM1 as the most important protein features for predicting future all-cause dementia. Sex  
5  
6 difference was evident for the association between RPS6KB1 and all-cause dementia, with  
7  
8 stronger association in men ( $p$  for interaction=0.037). Replication in the UK Biobank  
9  
10 confirmed the associations between the identified proteins and various dementia subtypes. The  
11  
12 results from Mendelian randomization in the reverse direction indicated that several proteins  
13  
14 serve as early markers for dementia, rather than being direct causes of the disease.  
15  
16 These findings provide insights into putative mechanisms for dementia. Future studies are  
17  
18 needed to validate the findings on RPS6KB1 in relation to dementia risk.  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28

29 **Running title:** Unraveling plasma proteins in dementia  
30  
31  
32

33 **Keywords:** dementia, proteomics, mendelian randomization, ELSA, UK Biobank.  
34  
35  
36  
37  
38  
39  
40

## 41 **Introduction**

42  
43 The understanding of Alzheimer's disease and related dementia (ADRD) is increasingly  
44  
45 shifting towards a systemic and multifactorial perspective.<sup>1,2</sup> Circulating proteins, as pivotal  
46  
47 agents in biological processes, offer direct insights into disease mechanisms and can serve as  
48  
49 early indicators, regulators, and effectors in disease pathways. This renders their studies  
50  
51 indispensable in both drug discovery and diagnostics development.<sup>3,4</sup>  
52  
53

54  
55 Mounting evidence supports the significance of proteomics in exploring pathways involved in  
56  
57 ADRD.<sup>3,5-7</sup> At the molecular level, deviations in protein function or expression play a role in  
58  
59 the pathogenesis of prodromal dementia,<sup>3,8,9</sup> while protein biomarkers can forecast disease  
60

1  
2  
3 onset several years before symptoms manifest.<sup>3,8-10</sup> Remarkably, approximately 96% of  
4 currently approved drugs target proteins,<sup>4,11</sup> underscoring the substantial value of proteomics  
5 in ADRD drug discovery.  
6  
7

8  
9  
10 Integrating large-scale proteomics data into population studies represents a recent  
11 development,<sup>3,12</sup> enabling cost-effective simultaneous measurement of multiple proteins on  
12 many samples.<sup>3,13,14</sup> This has led to the identification of distinct protein signatures relevant to  
13 ADRD susceptibility.<sup>4,9,10,15-18</sup> Longitudinal analyses in the UK Whitehall II study spanning  
14 across two decades, demonstrated associations between 15 non-amyloid/non-tau-related  
15 proteins and cognitive decline and dementia.<sup>4</sup> Similarly, the Atherosclerosis Risk in  
16 Communities (ARIC) study in the US identified significant protein signatures for dementia,  
17 including immune and proteostasis/autophagy pathways.<sup>10</sup> Intriguingly, some of these  
18 associations were independent of known Alzheimer's disease (AD) risk factors, suggesting  
19 novel potential targets for intervention.<sup>4,9,10</sup> Recent analyses from the UK Biobank (UKB)  
20 identified few known proteins associated with incident all-cause dementia (ACD), AD, and  
21 vascular dementia (VAD).<sup>16,17</sup> However, previous studies utilized the aptamer-based  
22 SomaScan platform in ARIC and Whitehall II,<sup>4,9,10</sup> and the SomaScan platform is deemed to  
23 have lower specificity compared to the Olink platform, which employs multiplexed antibody-  
24 based immunoassays proximity extension assay (PEA) technology.<sup>19</sup> Moreover, the UKB  
25 studies faced limitations such as a lack of external validation cohorts and confirming  
26 relationships via causal inference such as Mendelian Randomization (MR).<sup>16,17</sup>  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48

49 In this current study, we employed the large-scale Olink proteomics platform and a robust  
50 dementia algorithm to assess the proteomic signature of dementia risk in over 3,000 older adults  
51 using data from the English Longitudinal Study of Ageing (ELSA) as the discovery cohort. We  
52 validated these findings using Olink proteomics data from over 50,000 participants from the  
53 UKB.<sup>20</sup> Two-sample bi-directional MR and drug target MR were utilized to infer causality  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 between protein concentration and dementia outcomes, leveraging summary statistics from  
4  
5 large genome-wide association study (GWAS) consortia.  
6  
7  
8  
9  
10

## 11 **Materials and Methods**

### 12 **Consent statements**

13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
Ethical consent has been obtained for all waves and components of ELSA, according to the  
ethical approval system in operation at the time and in accordance with the Declaration of  
Helsinki.

UK Biobank has approval from the Northwest Multi-center Research Ethics Committee  
(MREC) as a Research Tissue Bank (RTB) approval.

### **ELSA discovery cohort study population**

ELSA is a nationally representative sample of men and women aged 50 years and over living  
in England. Data collection started in 2002-2003, with participants interviewed every two  
years. Details of study design are available elsewhere.<sup>21</sup> Blood sample collection in ELSA took  
place for the first time in wave 2 nurse visit in 2004-2005 and subsequently in every four-year  
interval. The blood collected from wave 4 nurse visits in 2008-2009 were used for the  
proteomics profiling (thus forming the baseline sample of this study), thereby affording a  
temporal perspective that allows for the exploration of the relationship between protein  
concentrations and AD/ADRD over 10 years. The following exclusion criteria were applied: 1)  
participants who died within 2 years since the wave 4 nurse visit; or 2) participants lost to  
follow-up (missing at  $\geq 2$  waves). A total of 3,305 available plasma samples from wave 4 were  
retrieved for the proteomics profiling.

1  
2  
3 The proteomics dataset in ELSA was curated with a focus on investigating the underlying  
4 biological processes associated with ADRD and cognitive decline. The assays encompassed an  
5 extensive array of cardiovascular and inflammatory markers, in addition to markers integral to  
6 neurological processes such as axon guidance, neurogenesis, and synapse assembly. These  
7 analyses were conducted utilizing the Olink proteomics platform, the antibody based Olink  
8 PEA technology.<sup>22</sup> We thereby used Olink Target 96 Neurology, Cardiovascular II and  
9 Neurology Exploratory arrays in these analyses. Frozen samples were shipped to Olink for  
10 aliquoting, plating, and assays. These assays include a built-in quality control based on four  
11 internal controls that are spiked into all samples, and external controls. Following stringent  
12 data quality control (Supplementary Methods), proteins were measured across three panels  
13 containing 276 proteins. Proteins were presented as NPX values, the arbitrary unit on log<sup>2</sup> scale  
14 from Olink. After excluding those who did not pass quality control, a combined dataset  
15 comprising of 3,262 samples were included.

## 35 **Covariates assessment in ELSA**

36  
37  
38 Baseline (at wave 4, 2008-2009) sociodemographic and socioeconomic covariates included age  
39 (in years), biological sex (male vs female) and ethnicity (white vs other ethnic groups) were  
40 self-reported. The age that participants left formal education was coded as follows: none, age  
41 14 or under, 15, 16, 17, 18, 19 or over. Smoking status was self-reported and was categorized  
42 as never, former, and current smoker. Physician-diagnosed cardiovascular disease (heart attack,  
43 angina, or heart failure) was self-reported. Depression was also self-reported. Body mass index  
44 (BMI) was calculated using participant's height and weight measured during the nurse visit.  
45  
46 Three measurements were taken of systolic blood pressure (SBP) on the respondent's right arm  
47 while they were seated, and the average of the three measurements was used. low-density  
48 lipoprotein (LDL) cholesterol was assayed using the blood sample collected by the nurse. The  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 APOE genotype was derived from the analysis of two specific SNPs, namely rs7412 and  
4 rs429358. To determine these genotypes, two TaqMan assays from Assay-On-Demand, a  
5 product of Applied Biosystems and Gene service Ltd in Cambridge, UK, were employed.  
6  
7 These assays were conducted on a 7900HT analyzer, manufactured by Applied Biosystems,  
8 and the genotypes were determined using the Sequence Detection Software (version 2.0), also  
9  
10 from Applied Biosystems. The quality control of genome-wide genotyping has been described  
11  
12 elsewhere.<sup>23</sup> Episodic memory was assessed at wave 4, evaluated through the immediate and  
13  
14 delayed recall tasks of the Consortium to Establish a Registry for AD (CERAD).<sup>24</sup> Participants  
15  
16 were presented with a ten-word list and tasked with recalling it both immediately and after a  
17  
18 delay. The scores from these tests were aggregated to compute a memory score.  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28

## 29 **Dementia algorithm in ELSA**

30  
31 The standardized algorithm for identifying dementia cases relied on five primary data sources:  
32  
33 1) coded information extracted from interviews across all waves using participant self-reported  
34  
35 physician diagnosis of AD and dementia; 2) caregivers who completed a modified short-form  
36  
37 Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE); 3) medication data  
38  
39 collected during nurse visits (wave 6, 8, and 9); 4) linked data from hospital admissions (NHS  
40  
41 Hospital Episode Statistics) and 5) mortality records (Office for National Statistics Mortality  
42  
43 Statistics). All data sources were integrated into algorithm development (see Supplementary  
44  
45 Table 1A for ICD-10 codes used for ascertaining ACD).  
46  
47  
48

49  
50 Dementia subtypes, specifically AD and VAD, were also derived (Supplementary Table 1B).  
51  
52 The computation of time-to-event for dementia cases depended on the data source where  
53  
54 dementia was first recorded. Further details on the dementia algorithm can be found in the  
55  
56 Supplementary Methods.  
57  
58  
59  
60

1  
2  
3 Based on the algorithm, baseline prevalent dementia was excluded ( $N=13$ ), which yielded a  
4  
5 final analytical sample of 3,249.  
6  
7  
8  
9

## 10 **UKB for validation**

11  
12  
13 UKB is a large population-based cohort from the UK with over half a million participants aged  
14  
15 40 – 69 years, recruited between 2006 and 2010.<sup>20</sup> Proteomics profiling was conducted in  
16  
17 54,219 participants at study baseline, with 2,923 unique proteins assayed using the antibody  
18  
19 based Olink Explore 3072 PEA, across eight Olink panels (Cardiometabolic I, Cardiometabolic  
20  
21 II, Inflammation I, Inflammation II, Neurology I, Neurology II, and Oncology I, Oncology II).  
22  
23 Consortium members opted for samples enriched in specific diseases of interest, while the  
24  
25 remaining population was randomly sampled using a stratified approach based on age, sex, and  
26  
27 recruitment center.<sup>20</sup> The current analysis excluded those with baseline dementia, yielding a  
28  
29 total sample of 52,745 individuals. The significant proteins identified ELSA were subsequently  
30  
31 validated using UKB, if assayed.  
32  
33  
34  
35

36 We have attempted, where possible, to derive similar variables for ELSA and the UKB cohorts  
37  
38 with consideration of the level of missingness, to maximize comparability. Participant's age  
39  
40 was derived based on date of birth and date of attending an initial assessment center.  
41  
42 Participant's biological sex was acquired from central registry at recruitment and contains a  
43  
44 mixture of the sex recorded by the NHS and self-reported sex. Ethnicity was self-reported and  
45  
46 categorized into White, Mixed, Asian, or Asian British, and Black or Black British. Highest  
47  
48 qualification was determined by the answers provided to the question: "Which of the following  
49  
50 qualifications do you have?", with options included: College or University degree; NVQ  
51  
52 (National Vocational Qualification) or HND (Higher National Diploma) or HNC (Higher  
53  
54 National Certificate) or equivalent; other professional qualifications e.g.: nursing; A levels/AS  
55  
56 levels; O levels/GCSEs (General Certificate of Secondary Education) or equivalent; CSEs  
57  
58  
59  
60



(General Certificate of Secondary Education) or equivalent; or none of the above. Smoking status was self-reported and categorized as never, former, and current smoker. Self-reported medical conditions were solicited through the touchscreen questionnaire as well as during verbal interview conducted by a trained nurse, and the presence of cardiovascular disease (heart attack, angina, or heart failure) were defined if the participant reported any of these conditions. Depression was affirmatory if the participant confirmed to have any of probable recurrent major depression (severe), probable recurrent major depression (moderate), or single probable major depression episode, if reported on the questionnaire or nurse-administered verbal interview. BMI was constructed from height and weight measured during the initial assessment center visit using an Omron device. Two automatic readings of blood pressure were taken a few moments apart, using an Omron HEM-7015IT digital blood pressure monitor, and the average of the readings were used. LDL cholesterol was measured by enzymatic protective selection analysis on a Beckman Coulter AU5800 from the blood sample collected at recruitment.

Incident ACD, AD, VAD and FTD were defined by the UKB dementia algorithm.<sup>25</sup> The last date of censoring was 31<sup>st</sup> December 2022 (last date of linkage to death and inpatient records). Baseline dementia was removed.

## Statistical analysis

### Analysis 1: Protein-dementia associations in the ELSA discovery cohort

We used Cox proportional hazards regression models to evaluate the associations between each plasma protein NPX value and incident dementia using the ‘survival’ R package.

All proteins had  $\leq 6\%$  missing, with missingness largely due to failure in passing internal quality control. Rank-based inverse normal transformation was first applied to the protein levels given that imputation is done using Euclidean distances and scaled to have a mean of 0

1  
2  
3 and standard deviation of 1 prior to all analyses. Missing protein measurements were imputed  
4 using the K-nearest neighbor (KNN) ( $k=57$ ) imputation using the ‘impute’ R package,<sup>26</sup> which  
5 works by identifying the nearest 57 individuals defined using Euclidean distances and imputing  
6 with their medians, with  $k$  calculated from the square root of the total sample size ( $N=3,249$ ).  
7  
8 Missingness was imputed for proteomics data using KNN, together with clinical data using the  
9 Multiple Imputation by Chained Equations (MICE) procedure (‘mice’ R package).<sup>27</sup> A total of  
10 30 imputed datasets were generated, with the imputation procedure iterating 10 times to ensure  
11 stable and reliable estimates. Additionally, as part of the sensitivity analysis, a Multiple  
12 Random Forest Regression Imputation approach was employed to impute missing data from  
13 the full dataset with scaled protein levels. This was done using 30 imputed datasets and 10  
14 iterations for the main analysis in ELSA, allowing for consideration of non-linear relationships  
15 within the data and minimizing potential biases introduced by missingness.<sup>28</sup>

16  
17 Based on the imputed datasets, we first assessed the protein-dementia associations without any  
18 model adjustments by pooling the estimates from all 30 imputed datasets. The models were  
19 further adjusted for age, sex, and ethnicity for the minimally-adjusted model; and adjusted for  
20 age, sex, ethnicity, education, smoking status, depression, cardiovascular disease, BMI, SBP,  
21 and LDL cholesterol for the fully-adjusted models for each protein NPX value associated with  
22 the risk of incident dementia, with false discovery rate (FDR)-corrected p-value (denoted as  
23  $P_{\text{FDR}}$ ) set at a cut-off of 0.05, this translates to an uncorrected p-value of 0.00018. All p-values  
24 were two-sided.  $P_{\text{FDR}}$  were reported and displayed using volcano plot, accompanying the HR.  
25  
26 To evaluate the potential interactions in the Cox regression models, we incorporated interaction  
27 terms between protein levels and key demographic variables. The following interactions were  
28 explored: 1) protein  $\times$  sex; 2) protein  $\times$  age; 3) protein  $\times$  age squared ( $\text{age}^2$ ); and 4) protein  $\times$   
29  $\text{age}^2 \times$  sex.

30  
31 The Cox proportional hazard regression analyses were repeated for AD and VAD separately.  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 A series of sensitivity analyses were conducted using the same adjustment methods as the main  
4 Cox regression for incident ACD, by excluding the following: 1) dementia cases that occurred  
5 during the first year of follow-up to reduce the possibility of reverse causation bias; 2) other  
6 ethnic groups other than white ethnicity; 3) APOE  $\epsilon$ 4 carriers; and 4) participants < 60 years  
7 of age. We also conducted competing risk of death using Fine-Gray regression models<sup>29</sup> with  
8 the same covariate adjustments, given that death may preclude dementia from occurring. The  
9 competing risk models, which estimated sub-distribution HRs, were conducted for incident  
10 ACD accounting for all-cause mortality as a competing risk, incorporating time-to-event  
11 data.<sup>29,30</sup>

12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Next, receiver operating characteristic (ROC) analyses were conducted to assess the precision of the identified proteins from the fully adjusted Cox regression models in predicting incident ACD. These analyses were performed independently and in combination with additional factors including age, sex, ethnicity, education, APOE  $\epsilon$ 4 status, and memory score. To evaluate the performance of the Cox models, bootstrapping was performed to assess the stability of the AUC estimates. A total of 2000 bootstrap resamples were generated. Bootstrapping involves resampling with replacement from the original dataset to estimate the sampling distribution of a statistic. We used the bootstrapped samples to compute 95% CI for the AUC of each model, utilizing the R packages ‘caret’<sup>31</sup>, ‘boot’<sup>32</sup> and ‘pROC’.<sup>33</sup> The mean AUC value across all bootstrap resamples was calculated to estimate the overall predictive performance of the survival model.

We employed eXtreme Gradient Boosting (XGBoost), a powerful machine learning algorithm based on decision-tree ensembles within a gradient boosting framework, to identify the most important features predictive of future ACD, AD, and VAD onset. To ensure a thorough and reliable analysis, we included all available protein, demographic, and clinical predictors in our models, using imputed data to address missing values. The data was partitioned into two sets:

1  
2  
3 80% of the data was used as the training set to build and tune the XGBoost model, while the  
4  
5 remaining 20% was reserved as the test set to evaluate the model's predictive performance.  
6  
7  
8 Feature importance scores generated by the XGBoost model were used to identify key  
9  
10 predictors of ACD onset. To enhance the interpretability of these feature importance rankings,  
11  
12 we employed SHapley Additive exPlanations (SHAP) values, which provide a consistent and  
13  
14 theoretically grounded approach for explaining the contribution of individual features to the  
15  
16 model's predictions. SHAP values enabled us to gain a clearer understanding of how each  
17  
18 predictor influenced the model's outcomes. In addition, this process of evidence triangulation  
19  
20 using XGBoost and SHAP values was conducted to complement and validate the findings from  
21  
22 our Cox regression models. By comparing results from both the machine learning-based  
23  
24 XGBoost approach and the statistical Cox regression models, we aimed to improve the  
25  
26 robustness and reliability of our predictive models for future dementia onset.  
27  
28  
29

### 30 31 **Analysis 2: Protein-dementia associations in the UKB validation cohort**

32  
33 The significant proteins from the main and sensitivity analyses based on ELSA were then  
34  
35 validated using the UKB proteomics data, if assayed. The reason for choosing ELSA as the  
36  
37 discovery cohort and UKB as the validation cohort is that ELSA's target panels were  
38  
39 specifically designed to investigate proteomic signatures related to cognitive decline and  
40  
41 dementia, resulting in a more focused selection of proteins. In contrast, UKB offers a broader  
42  
43 protein selection, allowing for effective validation.  
44  
45

46  
47 The UKB proteomics samples underwent sample selection, processing, and quality control  
48  
49 procedures.<sup>20</sup> Missing protein measurements for the remaining individuals were imputed using  
50  
51 K-nearest neighbor imputation ( $k=230$ ), the protein levels were normalized and scaled akin to  
52  
53 the analyses in ELSA proteomics data described above.  
54

55  
56 Data was processed and analyzed using the R Studio Workbench on the UKB Research  
57  
58 Analysis Platform, under application No.71702.  
59  
60

### Analysis 3: Two-sample bi-directional MR

We then assessed the potential causal relationships between the circulating protein concentration, in relation to dementia outcomes, using two-sample bi-directional MR. Summary statistics for genetic variants associated with the circulating protein levels, protein quantitative trait locus (pQTL), that is also associated with dementia in GWAS were used to infer causality.

Selection of instruments to proxy for altered protein abundance were derived using pQTL mapping proteins that identifies genetic associations in participants of European ancestry from the UK Biobank based on Olink data (<https://doi.org/10.7303/syn51364943>), and genome-wide association meta-analysis (GWAMA) based on 12 participating Olink cohorts from the SCALLOP Consortium for RPS6KB1 since it was not assayed in the UKB.<sup>20,34</sup> The effects of protein pQTL were standardized to align with the same effect allele.

Methods for the MR analyses were detailed in the Supplementary Methods.

### Analysis 4: Two sample drug target MR (cis-MR)

Next, we used a two-sample MR study design, based predominantly on genetic variants located in or near genes that encode the relevant drug targets, to infer causality from protein concentration → dementia (cis-MR).<sup>35-38</sup> Cis-MR is considered to be less susceptible to pleiotropy, and the potential effect of a drug by analyzing the genomic locus encoding protein targets, which may be informative for drug trial design.<sup>36,38</sup>

Methods for conducting drug target MR were further described in the Supplementary Methods.

## Analysis 5: Enrichment analysis

Enrichment analysis was conducted by searching open-source databases to further characterize the identified proteins from the Cox regressions. We employed the Enrichr,<sup>39</sup> which is a computational method infers knowledge about an input gene set by comparing it to annotated gene sets that represent existing biological knowledge. It determines if the input set of genes significantly overlaps with these annotated gene sets. We used the full set of ELSA proteins as the background gene set, to glean a deeper biological understanding. We searched the following bioinformatics databases: Gene Ontology (GO): GO Molecular Function, GO Biological Process, and GO Cellular Component<sup>40</sup>, Kyoto Encyclopedia of Genes and Genomes (KEGG)<sup>41</sup>, Reactome Pathway Database (REACTOME)<sup>42</sup>, Illuminating the Druggable Genome (IDG)<sup>43</sup>, Proteomics Drug Atlas (PDA)<sup>44</sup>, and Genotype-Tissue Expression (GTEx)<sup>45</sup>. Statistical significance was indicated if  $P_{FDR} < 0.05$ . Furthermore, we utilized the Open Targets platform (<https://www.opentargets.org/>) for the systematic identification of potential therapeutic drug targets among the identified proteins.<sup>46</sup> All analyses were done using R Studio (version 4.4.1).

## Results

### Analysis 1: Protein-dementia associations in the ELSA discovery cohort

The participant selection for the proteomics assay in ELSA is depicted in Supplementary Fig. 1. In 3,262 samples with proteomics assayed, based on the dementia algorithm, prevalent dementia cases were excluded ( $N=13$ ), resulting in a final sample of 3,249 in the analysis.

1  
2  
3 The mean age was 63.4 years ( $SD=9.2$ ), 55% were women, and 97.2% were of white ethnicity  
4 (Supplementary Table 2). A total of 229 incident ACD, 89 AD and 41 VAD cases were  
5 documented over a median follow-up of 9.8 years (min–max: 0.4–10.9 years). Specific details  
6 on the data sources where these cases were extracted from in the first instance are included in  
7 Supplementary Fig. 2. The normalized protein expression (NPX) levels for participants with  
8 no dementia, incident dementia, and prevalent dementia are presented as box plots for each  
9 protein. P-values were calculated using the Kruskal-Wallis method, with adjustments for  
10 multiple comparisons using the False Discovery Rate (FDR). These results are shown in  
11 Supplementary Figures 3-5 for each Olink Target 96 panel.

12  
13 We initially assessed the relationship between the NPX value of 276 plasma proteins and ACD  
14 risk in the ELSA cohort, using Cox Proportional Hazard regression models. Unadjusted  
15 analyses revealed that 95 measured proteins were significantly associated with ACD  
16 (Supplementary Fig. 6). Among these, **NEFL** exhibited the strongest association with ACD  
17 ( $P_{FDR}=8.66\times 10^{-37}$ , hazard ratio (HR) [95% confidence intervals (CI): **3.01 [2.63, 3.44]**),  
18 followed by **EDA2R**, **SCARF2**, **LAYN**, **PGF**, **DCN**, **GFR-alpha-1**, **BNP**, **UNC5C**, **Dkk-4**,  
19 **KIM1** (also known as **HAVCR1**), **TNFRSF12A**, **CADM3**, **TRAIL-R2**, **VWC2**, and  
20 **MMP12**.

21  
22 In the minimally adjusted models (adjusted for age, sex, and ethnicity), **NEFL** ( $P_{FDR}=0.0002$ ;  
23 HR [95% CI]: **1.55 [1.30, 1.85]**), **RPS6KB1** ( $P_{FDR}=0.003$ ; HR [95% CI]:**1.34 [1.18, 1.53]**),  
24 **EDA2R** ( $P_{FDR}=0.046$ ; HR [95% CI]:**1.43 [1.19, 1.72]**) and **KIM1** ( $P_{FDR}=0.049$ ; HR [95%  
25 CI]:**1.31 [1.14, 1.50]**) were significantly associated with ACD (Supplementary Fig. 7).

26  
27 In the fully adjusted models (adjusted for age, sex, ethnicity, education, smoking status,  
28 depression, presence of cardiovascular diseases, BMI, SBP, LDL cholesterol), **NEFL**  
29 ( $P_{FDR}=0.0008$ ; HR [95% CI]:**1.54 [1.29, 1.84]**) and **RPS6KB1** ( $P_{FDR}=0.01$ ; HR [95% CI]:**1.33**  
30 **[1.16, 1.52]**) remained significantly associated with ACD (Fig. 1).

Using multiple random forest regression as an alternative data imputation method produced broadly similar results, with **RPS6KB1** ( $P_{FDR}=0.01$ ; HR [95% CI]:1.28 [**1.14, 1.44**]), **NEFL** ( $P_{FDR}=0.02$ ; HR [95% CI]:**1.31 [1.15, 1.49]**), and **KIM1** ( $P_{FDR}=0.039$ ; HR [95% CI]:**1.28 [1.13, 1.45]**) demonstrating robust associations with ACD (Supplementary Fig. 8).

There was evidence supporting a sex difference in **RPS6KB1** ( $p$  for interaction=0.037), with higher level of RPS6KB1 was associated with a greater risk of dementia in men (HR [95% CI]:**1.54 [1.24, 1.91]**), but the effect was smaller in women (HR [95% CI]:**1.21 [1.02, 1.43]**).

There was no evidence of interaction effects by age, age<sup>2</sup>, and between age<sup>2</sup> and sex.

Sensitivity analyses demonstrated the robustness of the association between **NEFL** and **ACD**, with significance persisting after excluding participants in other ethnic groups ( $P_{FDR}=0.001$ ; HR [95% CI]:**1.53 [1.28, 1.83]**) (Supplementary Fig. 9), APOE  $\epsilon 4$  carriers ( $P_{FDR}=0.009$ ; HR [95% CI]:**1.61 [1.29, 2.01]**) (Supplementary Fig. 10), cases of dementia occurring within the first year of follow-up ( $P_{FDR}=0.005$ ; HR [95% CI]:**1.50 [1.25, 1.80]**) (Supplementary Fig. 11), participants aged <60 years ( $P_{FDR}=0.011$ ; HR [95% CI]:**1.47 [1.22, 1.78]**) (Supplementary Fig. 12), and when death was considered as a competing risk in Fine-Gray regression models ( $P_{FDR}=0.002$ ; HR [95% CI]:**1.50 [1.26, 1.80]**) (Supplementary Fig. 13). Similarly, **RPS6KB1** exhibited a robust association with **ACD**, which remained significant after excluding other ethnic groups ( $P_{FDR}=0.004$ ; HR [95% CI]:**1.34 [1.18, 1.53]**), cases of dementia occurring within the first year of follow-up ( $P_{FDR}=0.025$ ; HR [95% CI]:**1.31 [1.15, 1.50]**), participants aged <60 years ( $P_{FDR}=0.031$ ; HR [95% CI]:**1.31 [1.14, 1.50]**), and it was significantly associated with ACD in the competing risk model ( $P_{FDR}=0.014$ ; HR [95% CI]:**1.33 [1.16, 1.52]**). However, the significance in association between RPS6KB1 and ACD based on  $P_{FDR}$  attenuated after excluding APOE  $\epsilon 4$  carriers ( $P_{FDR}=0.961$ ; HR [95% CI]:**1.28 [1.08, 1.50]**). After excluding other ethnic groups, **KIM1** was additionally associated with **ACD** ( $P_{FDR}=0.027$ ; HR [95% CI]:**1.32 [1.15, 1.53]**). When assessed by dementia subtypes, after full



1  
2  
3 adjustment, no protein was found to be significantly associated with AD indicated by  
4  $P_{FDR} < 0.05$  (Supplementary Fig. 14). **MMP12** was found to be associated with **VAD**  
5 ( $P_{FDR} = 0.046$ ; HR [95% CI]:**2.06 [1.41, 2.99]**) (Supplementary Fig. 15). Albeit being non-  
6 significant after FDR correction, of the proteins significantly associated with ACD, based on  
7 uncorrected nominal statistical significance (denoted as  $P_{uncorrected} < 0.05$ ), **RPS6KB1** was  
8 associated with **AD** ( $P_{uncorrected} = 0.006$ ; HR [95% CI]:**1.29 [1.07, 1.55]**); and **NEFL** was  
9 associated with **VAD** ( $P_{uncorrected} = 0.001$ ; HR [95% CI]:**1.98 [1.31, 2.99]**).

10 For predicting incident ACD, plasma NEFL and RPS6KB1 parsimonious models yielded  
11 modest Area Under the Receiver Operating Characteristic Curve (AUC) values [95% CI] of  
12 0.787 [0.757, 0.815], and 0.609 [0.571, 0.647], respectively (Fig. 2). We also evaluated the  
13 performance of these two proteins in combination with other measures, including demographic  
14 predictors (age, sex, ethnicity, education), APOE  $\epsilon 4$  status, and memory score. When NEFL  
15 was combined with these predictors, the model achieved an accuracy of AUC [95% CI]:0.866  
16 [0.840, 0.888]. Comparatively, when RPS6KB1 was combined with other predictors, the model  
17 achieved a comparable accuracy of AUC [95% CI]:0.866 [0.842, 0.891]. NEFL and RPS6KB1  
18 in combination with all the other predictors yielded AUC [95% CI]:0.871 [0.845, 0.894].

19 For ACD, the XGBoost models revealed that age (mean |SHAP|=0.0508) and memory score  
20 (mean |SHAP|=0.0170) were the most important features contributing to the prediction ACD  
21 onset. Additionally, protein markers **NEFL** (mean |SHAP|=0.010), **RPS6KB1** (mean  
22 |SHAP|=0.0080) and **KIM1** (mean |SHAP|=0.047) emerged as the most prominent protein  
23 marker in predicting **ACD** (Fig. 3). The SHAP plot also illustrated that, as an example,  
24 individuals with elevated levels of RPS6KB1 were more predisposed to developing ACD,  
25 while those with lower levels were more likely to remain ACD-free.

## Analysis 2: Protein-dementia associations in the UKB validation cohort

Based on the main, sensitivity, and subtype analyses results in ELSA, all identified proteins were selected for validation using the UKB cohort. However, RPS6KB1 was not assayed in the UKB.

In the UKB, which included 52,745 participants with proteomics assayed and without dementia at study baseline (53.9% women, 93.3% white ethnicity), the mean age was 56.8 years ( $SD=8.2$ ) (Supplementary Table 2). UKB participants with proteomics assayed were, on average, younger than participants in ELSA. Over a median of 13.7 years (min–max: 0.03–16.8 years) of follow-up, a total of 1,506 incident ACD, 732 AD, 281 VAD, and 111 FTD cases were recorded.

Using the same adjustment strategy for the Cox regression models in ELSA, **NEFL** was replicated in the UKB for **ACD** ( $P_{FDR}=1.02\times 10^{-81}$ ; HR [95% CI]: **1.87 [1.75, 1.99]**), **AD** ( $P_{FDR}=1.89\times 10^{-35}$ ; HR [95% CI]:**1.81 [1.65, 1.99]**), **VAD** ( $P_{FDR}=1.59\times 10^{-17}$ ; HR [95% CI]:**1.90 [1.64, 2.19]**), and **FTD** ( $P_{FDR}=1.10\times 10^{-21}$ ; HR [95% CI]:**2.97 [2.30, 3.70]**) (Fig. 4; Supplementary Table 3). **KIM1** was replicated for **ACD** ( $P_{FDR}=3.15\times 10^{-4}$ ; HR [95% CI]:**1.13 [1.06, 1.20]**), **AD** ( $P_{FDR}=0.077$ ; HR [95% CI]:**1.11 [1.02, 1.21]**), and **VAD** ( $P_{FDR}=1.13\times 10^{-6}$ ; HR [95% CI]:**1.44 [1.25, 1.66]**). **MMP12** was replicated for **VAD** ( $P_{FDR}=6.85\times 10^{-5}$ ; HR [95% CI]:**1.36 [1.18, 1.56]**) and it was also associated with **ACD** ( $P_{FDR}=2.00\times 10^{-6}$ ; HR [95% CI]:**1.17 [1.10, 1.24]**). **EDA2R** was replicated for **ACD** ( $P_{FDR}=3.18\times 10^{-13}$ ; HR [95% CI]:**1.31 [1.22, 1.40]**), **AD** ( $P_{FDR}=6.06\times 10^{-5}$ ; HR [95% CI]:**1.25 [1.13, 1.39]**), and **VAD** ( $P_{FDR}=0.001$ ; HR [95% CI]:**1.34 [1.15, 1.58]**).



#### Analysis 4: Two sample drug target MR (cis-MR)

The instrument selection for drug target MR (cis-MR) relies on single nucleotide polymorphisms (SNPs) within or near the gene encoding region that regulates the protein of interest. However, the encoding region of EDA2R is located within the X chromosome, which precluded the analysis of drug target MR on EDA2R, as the sex chromosomes were excluded from GWAS summary statistics. There were insufficient valid instruments for RPS6KB1. Cis-MR was conducted for NEFL, KIM1, and MMP12.

In drug target MR, there was no causal evidence for any of the protein-dementia relationships (Supplementary Table 5, Supplementary Fig. 24-26).

In sensitivity analysis using a less stringent instrument selection approach, results were largely consistent. There was some evidence indicating a causal relationship between **KIM1** and **AD** (FinnGen 2023) ( $\beta$  [se]:-0.102 [0.041];  $P=0.037$ ) and **ACD** (and  $\beta$  [se]:-0.094 [0.037];  $P=0.036$ ) based on MR-Egger.

#### Analysis 5: Enrichment analysis

In Fig. 5 and Fig.6 (also depicted in Supplementary Table 6), the enrichment analyses revealed several biological pathways potentially implicated for the identified proteins (NEFL, RPS6KB1, KIM1, EDA2R, and MMP12), including the immune system, cancers, and insulin signaling. Tissue expression analysis showed expression in the brain for NEFL and in the kidney for KIM1. Notably, one drug, LY2584702, which is a selective, adenosine triphosphate (ATP)-competitive p70S6K inhibitor, has been investigated in clinical trials for the treatment of renal cell carcinoma, metastases, neoplasm, and neuroendocrine tumors – where RPS6KB1 was shown to be implicated in the mechanisms of action of the drug.

Upon searching the Open Targets platform, we identified ten known small molecule drugs in clinical trials (including LY2584702) that are linked to two proteins (RPS6KB1 and MMP12),



1  
2  
3 inhibition of mTOR may extend lifespan given that the mTOR activity becomes abnormally  
4 high with age.<sup>48</sup> In the nervous system, the mTOR pathway is implicated in the regulation of  
5 synaptic remodeling and long term potentiation.<sup>49-52</sup> Importantly, mTOR plays a crucial role in  
6 autophagy regulation in neurons,<sup>52,53</sup> and the mTOR/p70S6K axis is shown to be essential in  
7 the early phases of plasticity for synaptic modifications and the formation of enduring  
8 memory.<sup>54</sup> Previous analyses of the ARIC study similarly highlighted the importance of  
9 autophagy signaling pathways in the two decades before dementia onset.<sup>10</sup> Interestingly, our  
10 sensitivity analysis which excluded APOE  $\epsilon$ 4 carriers observed an attenuation in association  
11 between RPS6KB1 and dementia after considering multiple testing. Previous literature  
12 highlighted the mechanisms affected by APOE  $\epsilon$ 4,<sup>55</sup> such that the presence of APOE  $\epsilon$ 4 may  
13 be necessary for the overactivation in mTOR pathway which subsequently lead to tau  
14 hyperphosphorylation and reduced A $\beta$  clearance.<sup>56</sup> Nevertheless, genomics studies found that  
15 the combined effect of alleles in the RPS6KB1 gene, along with other genes in the tau kinase  
16 pathway, is linked to an increased risk of late-onset AD in without APOE  $\epsilon$ 4 allele.<sup>57</sup> At a lower  
17 expression level, RPS6KB1 can facilitate the growth of damaged axons resulting from CNS  
18 injury.<sup>58</sup> Transcriptomic exploration has revealed the central role of RPS6KB1, and alterations  
19 in its co-expression occur during the initial stages of AD, which highlights its potential as a  
20 biomarker for the early diagnosis of AD.<sup>57</sup> Several small clinical trials of rapamycin are  
21 underway for investigating age-related diseases including AD,<sup>48</sup> with primary outcomes  
22 assessing the effects on cognitive performance, and biomarkers of aging.<sup>48</sup> From the  
23 Proteomics Drug Atlas, one drug (LY2584702) was found to target RPS6KB1, which is a  
24 highly selective adenosine triphosphate competitive inhibitor against p70S6 Kinase. Further  
25 explorations are therefore needed to decipher the relationship between RPS6KB1 and the  
26 protein's pharmacological properties. Moreover, sex-specific association between RPS6KB1  
27 and ACD was intriguing. Animal studies have shown that the genetic deletion of RPS6KB1  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 inhibits cellular senescence and promotes longevity, but only in female mice.<sup>59</sup> Additionally,  
4 evidence suggests that RPS6KB1 mRNA expression varies based on menopausal status,<sup>60</sup>  
5 pointing to complex molecular interactions with hormones. The increased dementia risk that  
6 RPS6KB1 poses for men, however, warrants further investigation.  
7  
8

9  
10 Elevated NEFL were found to be associated with an increased risk of ACD in ELSA, and with  
11 ACD, AD, VAD, and FTD in the UKB. Consistently, previous studies from the UKB ranked  
12 NEFL as the most important protein associated with future dementia events out of 1463 protein  
13 markers,<sup>16</sup> and was associated AD and VAD.<sup>17</sup> NEFL is a marker of axonal injury,<sup>61,62</sup> and is  
14 implicated in several biological mechanisms related to dementia,<sup>63,64</sup> including  
15 neurodegeneration,<sup>65,66</sup> inflammation,<sup>67</sup> central nervous system (CNS) injury,<sup>68-70</sup> and  
16 atherosclerosis.<sup>71</sup> It is a well-established and non-specific marker of neurodegenerative  
17 diseases. While NEFL was found in our study to be causally linked to dementia based on the  
18 MR findings, the strongest indication was in the backwards direction, which points toward its  
19 role as a manifestation of prodromal dementia and anomalies in the brain, rather than a cause  
20 of dementia. This underscores the value of NEFL as an important diagnostic and early  
21 identification marker, as also demonstrated by the prediction models. It should be noted that  
22 the inconsistencies in MR findings from the current study across various AD GWAS could  
23 stem from the fact that the chosen genetic variant serving as the instrumental variable might  
24 exert a varied impact on the outcome within the represented population.  
25  
26

27  
28 Matrix metalloproteinases (MMPs) belong to a multigenic family of membrane-bound or  
29 secreted zinc-containing endopeptidases, which indirectly modulate the cellular processes  
30 through activation and inactivation of signaling molecules such as trophic factors cytokines,  
31 and receptor.<sup>72,73</sup> MMPs play important roles in cell proliferation and death,  
32 neuroinflammation, neurodegeneration, and glial reactivity,<sup>74</sup> and are linked to their proteolytic  
33 disruption action on the blood-brain barrier.<sup>75</sup> More specifically, MMP2, MMP3, and MMP9  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 were shown to play a crucial role in AD,<sup>76-78</sup> and damage to the white matter associated with  
4 VAD.<sup>79</sup> Based on experimental models, A $\beta$ <sub>40</sub> contribute to the changes in blood-brain barrier  
5 (BBB) permeability, and increased expression of MMPs in transgenic human amyloid  
6 precursor protein (hAPP)-overexpressing mice, in turn compromises BBB integrity.<sup>80</sup> Selective  
7 inhibitors for MMP12<sup>81</sup> was shown to reduce inflammation and delay of atherosclerosis  
8 progression.<sup>82,83</sup> For dementia, the associations between MMP12 and VAD risk in the UKB,  
9 and MMP12 and AD risk in the ARIC cohort was similarly highlighted.<sup>10,16</sup> At an elevated  
10 dosage, this medication can penetrate the BBB and manifest an inhibitory effect on  
11 metalloproteinase activity within the brain,<sup>84</sup> and was shown to decrease some seizure-related  
12 parameters.<sup>84</sup>

13  
14  
15 In our study, although the significant findings for KIM1 and EDA2R attenuated after full  
16 adjustments, MR analyses showed some possible causal links between these proteins and  
17 dementia. However, some of these MR results may be biased by horizontal pleiotropy with the  
18 proteins affecting multiple diseases,<sup>85</sup> possibly via immune, renal, and metabolic disease  
19 pathways,<sup>85-87</sup> which subsequently contribute to the risk of dementia.<sup>85</sup> There was also evidence  
20 from previous studies indicating higher levels of EDA2R were associated a smaller total brain  
21 volume, smaller grey matter volume, and less normal-appearing white matter volume.<sup>86</sup>

22  
23  
24 The current study exhibits robustness through several key strengths. Firstly, it draws upon two  
25 extensive population-based cohorts with prolonged follow-up, employing high-throughput and  
26 reliable proteomics data. The selection of the protein panel in ELSA is noteworthy for its  
27 focused curation on dementia-related markers, enhancing the study's precision in investigating  
28 associations with dementia risk in a nationally representative sample of older adults. The  
29 findings through the inclusion of the UKB encompassing a broader selection of proteins  
30 enhance the validity of our results. Both cohorts are well-characterized longitudinal cohorts,  
31 which enabled adjustment for wide range of factors. Furthermore, our study benefits from  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



1  
2  
3 applying a robust and comprehensive dementia algorithm in ELSA, which integrated  
4 information from various sources, bolstered by details on medication use and informant-  
5 solicited information, which has been reported to correlates better with objective cognitive  
6 performance than self-report alone as well as medication which particularly captured those with  
7 younger onset dementia,<sup>88</sup> allowing for accurate identification of incident dementia cases and  
8 the exclusion of prevalent cases. Another significant strength lies in our approach to assessing  
9 protein-dementia associations through a range of established methods for evidence  
10 triangulation. Additionally, the utilization of Olink antibody-based PEA is recognized for its  
11 superior specificity in proteomics assays.<sup>19</sup> The integration of proteogenomic in MR analyses  
12 was an additional strength.<sup>12</sup>

13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
Some limitations should be acknowledged. First, several circulating protein markers potentially  
relevant to dementia, for example, GFAP and GDF-15, were not assayed in ELSA. There is  
also limited specification, or a lack of protein measurements from the A/T/N classification  
framework,<sup>89</sup> such as beta-amyloid, p-tau217 and p-tau181. Third, another limitation is that the  
algorithm used for dementia ascertainment lacked information from primary care data, and  
pathology for confirming dementia cases, thus uncertainties exist ascertaining dementia  
subtypes. Nevertheless, it is important to note that AD and VAD pathology often co-exist on a  
population level, and many dementia patients exhibit mixed neuropathology.<sup>90</sup> Fourth, it is  
important to note that we lacked external validation cohorts for RPS6KB1. Lastly, there are  
inherent assumptions in MR analyses, and for drug target MR specifically, genetics might not  
directly inform on specific pharmacological aspects of drug exposure.

In conclusion, our proteomics analysis from two large-scale, population-based cohorts in the  
UK highlighted the utility of proteomics in identifying novel targets, enhancing our  
understanding of the mechanisms underlying dementia. MR analyses leveraging extensive  
GWAS data substantiated some of these protein-dementia relationships with causal evidence.

Looking forward, integrating large-scale population-based proteomics with other omics, such as genomics,<sup>12</sup> offers potential for deeper biological insights into diseases. Further research and validation are required to clarify the role of RPS6KB1 in ADRD and to explore potential sex differences in its effects.

## Data availability

The ELSA data is available on the UK Data Service. The proteomics data in ELSA will be deposited on the UK Data Service upon publication. All data from the UK Biobank, including the proteomics data, is available by directly submitting a project application to the UK Biobank.

All GWAS summary statistics are available online: <https://doi.org/10.7303/syn51364943>; [https://www.finngen.fi/en/access\\_results](https://www.finngen.fi/en/access_results); <https://gwas.mrcieu.ac.uk/>.

The codes used for all analyses are available on GitHub repository: [https://github.com/jgong94/ELSA\\_proteomics\\_dementia](https://github.com/jgong94/ELSA_proteomics_dementia).

## Acknowledgements

We would like to thank the participants in ELSA for their contribution to the research. We additionally want to acknowledge the participants and investigators of the FinnGen study. We would like to thank Olink representatives and the Newcastle laboratory for their support for data procurement and data pre-processing.

This research has been conducted using the UK Biobank Resource under application number 71702. This work uses data provided by patients and collected by the NHS as part of their care and support. Copyright © 2023, NHS England. Re-used with the permission of the NHS England and UK Biobank. All rights reserved.

## Funding

The English Longitudinal Study of Ageing is funded by the National Institute on Aging (grant number R01AG17644) and the National Institute for Health and Care Research (198/1074-02).

The National Institute of Aging (NIA) (grant Number [R01AG17644]) funded the proteomics data curation in ELSA. J.G. is supported by the NIA (grant Number [R01AG17644]). D.M.W. is supported by the British Medical Research Council (MRC) (grant Number [MC\_UU\_00019/3]).

The funders had no role in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

## Competing interests

Olink had no part in designing the study or analyzing the data. No conflicts of interest to be declared from any of the authors.

## References

1. Shi Y, Holtzman DM. Interplay between innate immunity and Alzheimer disease: APOE and TREM2 in the spotlight. *Nature Reviews Immunology*. 2018;18(12):759-772.
2. Yousef H, Czupalla CJ, Lee D, et al. Aged blood impairs hippocampal neural precursor activity and activates microglia via brain endothelial cell VCAM1. *Nature medicine*. 2019;25(6):988-1000.
3. Suhre K, McCarthy MI, Schwenk JM. Genetics meets proteomics: perspectives for large population-based studies. *Nature Reviews Genetics*. 2021;22(1):19-37.
4. Lindbohm JV, Mars N, Walker KA, et al. Plasma proteins, cognitive decline, and 20-year risk of dementia in the Whitehall II and Atherosclerosis Risk in Communities studies. *Alzheimer's & dementia*. 2022;18(4):612-624.
5. Gomes B, Ashley EA. Artificial intelligence in molecular medicine. *New England Journal of Medicine*. 2023;388(26):2456-2465.
6. Teunissen CE, Verberk IM, Thijssen EH, et al. Blood-based biomarkers for Alzheimer's disease: towards clinical implementation. *The Lancet Neurology*. 2022;21(1):66-77.
7. Zetterberg H. Biofluid-based biomarkers for Alzheimer's disease-related pathologies: An update and synthesis of the literature. *Alzheimer's & Dementia*. 2022;18(9):1687-1693.

8. Pietzner M, Wheeler E, Carrasco-Zanini J, et al. Mapping the proteo-genomic convergence of human diseases. *Science*. 2021;374(6569):eabj1541.
9. Walker KA, Chen J, Zhang J, et al. Large-scale plasma proteomic analysis identifies proteins and pathways associated with dementia risk. *Nature Aging*. 2021;1(5):473-489.
10. Walker KA, Chen J, Shi L, et al. Proteomics analysis of plasma from middle-aged adults identifies protein markers of dementia risk in later life. *Science translational medicine*. 2023;15(705):eadf5681.
11. Santos R, Ursu O, Gaulton A, et al. A comprehensive map of molecular drug targets. *Nature reviews Drug discovery*. 2017;16(1):19-34.
12. Sun BB, Suhre K, Gibson BW. Promises and challenges of populational proteomics in health and disease. *Molecular & Cellular Proteomics*. 2024:100786.
13. Assarsson E, Lundberg M, Holmquist G, et al. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PloS one*. 2014;9(4):e95192.
14. Gold L, Ayers D, Bertino J, et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. *Nature Precedings*. 2010:1-1.
15. Kivimäki M, Walker KA, Pentti J, et al. Cognitive stimulation in the workplace, plasma proteins, and risk of dementia: three analyses of population cohort studies. *bmj*. 2021;374
16. Guo Y, You J, Zhang Y, et al. Plasma proteomic profiles predict future dementia in healthy adults. *Nature Aging*. 2024/02/12 2024;doi:10.1038/s43587-023-00565-0
17. Gadd DA, Hillary RF, Kuncheva Z, et al. Blood protein assessment of leading incident diseases and mortality in the UK Biobank. *Nature Aging*. 2024:1-10.
18. Jiang Y, Zhou X, Ip FC, et al. Large-scale plasma proteomic profiling identifies a high-performance biomarker panel for Alzheimer's disease screening and staging. *Alzheimer's & Dementia*. 2022;18(1):88-102.
19. Eldjarn GH, Ferkingstad E, Lund SH, et al. Large-scale plasma proteomics comparisons through genetics and disease associations. *Nature*. 2023;622(7982):348-358.
20. Sun BB, Chiou J, Traylor M, et al. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature*. 2023;622(7982):329-338.
21. Steptoe A, Breeze E, Banks J, Nazroo J. Cohort profile: the English longitudinal study of ageing. *International journal of epidemiology*. 2013;42(6):1640-1648.
22. Wik L, Nordberg N, Broberg J, et al. Proximity extension assay in combination with next-generation sequencing for high-throughput proteome-wide analysis. *Molecular & Cellular Proteomics*. 2021;20
23. Ajnakina O, Cadar D, Steptoe A. Interplay between socioeconomic markers and polygenic predisposition on timing of dementia diagnosis. *Journal of the American Geriatrics Society*. 2020;68(7):1529-1536.
24. Mirra SS, Heyman A, McKeel D, et al. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD) Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. *Neurology*. 1991;41(4):479-479.
25. Wilkinson T, Schnier C, Bush K, et al. Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data. *European journal of epidemiology*. 2019;34:557-565.
26. Hastie T, Tibshirani R, Narasimhan B, Chu G. impute: impute: Imputation for microarray data. R package version; 2020.
27. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *Journal of statistical software*. 2011;45:1-67.
28. Chhabra G, Vashisht V, Ranjan J. A comparison of multiple imputation methods for data with missing values. *Indian Journal of Science and Technology*. 2017;10(19):1-7.



51. Hoeffler CA, Klann E. mTOR signaling: at the crossroads of plasticity, memory and disease. *Trends in neurosciences*. 2010;33(2):67-75.
52. Yates SC, Zafar A, Hubbard P, et al. Dysfunction of the mTOR pathway is a risk factor for Alzheimer's disease. *Acta neuropathologica communications*. 2013;1:1-15.
53. Wang Y, Zhang H. Regulation of autophagy by mTOR signaling pathway. *Autophagy: Biology and diseases: Basic science*. 2019:67-83.
54. Magnuson B, Ekim B, Fingar DC. Regulation and function of ribosomal protein S6 kinase (S6K) within mTOR signalling networks. *Biochemical Journal*. 2012;441(1):1-21.
55. Hunsberger HC, Pinky PD, Smith W, Suppiramaniam V, Reed MN. The role of APOE4 in Alzheimer's disease: strategies for future therapeutic interventions. *Neuronal signaling*. 2019;3(2):NS20180203.
56. Jung CH, Ro S-H, Cao J, Otto NM, Kim D-H. mTOR regulation of autophagy. *FEBS letters*. 2010;584(7):1287-1295.
57. Upadhyaya Y, Xie L, Salama P, et al. Differential co-expression analysis reveals early stage transcriptomic decoupling in alzheimer's disease. *BMC Medical Genomics*. 2020;13:1-10.
58. Al-Ali H, Ding Y, Slepak T, et al. The mTOR substrate S6 kinase 1 (S6K1) is a negative regulator of axon regeneration and a potential drug target for central nervous system injury. *Journal of neuroscience*. 2017;37(30):7079-7095.
59. Regitz-Zagrosek V, Kararigas G. Mechanistic pathways of sex differences in cardiovascular disease. *Physiological reviews*. 2017;97(1):1-37.
60. Zhao H, Wang J, Fang D, et al. Adiposity results in metabolic and inflammation differences in premenopausal and postmenopausal women consistent with the difference in breast cancer risk. *Hormones and Cancer*. 2018;9:229-239.
61. Ljungqvist J, Zetterberg H, Mitsis M, Blennow K, Skoglund T. Serum neurofilament light protein as a marker for diffuse axonal injury: results from a case series study. *Journal of neurotrauma*. 2017;34(5):1124-1127.
62. Graham NS, Zimmerman KA, Moro F, et al. Axonal marker neurofilament light predicts long-term outcomes and progressive neurodegeneration after traumatic brain injury. *Science translational medicine*. 2021;13(613):eabg9922.
63. Olsson B, Portelius E, Cullen NC, et al. Association of cerebrospinal fluid neurofilament light protein levels with cognition in patients with dementia, motor neuron disease, and movement disorders. *JAMA neurology*. 2019;76(3):318-325.
64. Ashton NJ, Janelidze S, Al Khleifat A, et al. A multicentre validation study of the diagnostic value of plasma neurofilament light. *Nature communications*. 2021;12(1):3400.
65. Marks JD, Syrjanen JA, Graff-Radford J, et al. Comparison of plasma neurofilament light and total tau as neurodegeneration markers: associations with cognitive and neuroimaging outcomes. *Alzheimer's Research & Therapy*. 2021;13:1-14.
66. Illán-Gala I, Lleo A, Karydas A, et al. Plasma tau and neurofilament light in frontotemporal lobar degeneration and Alzheimer disease. *Neurology*. 2021;96(5):e671-e683.
67. Novakova L, Zetterberg H, Sundström P, et al. Monitoring disease activity in multiple sclerosis using serum neurofilament light protein. *Neurology*. 2017;89(22):2230-2237.
68. Gisslén M, Price RW, Andreasson U, et al. Plasma concentration of the neurofilament light protein (NFL) is a biomarker of CNS injury in HIV infection: a cross-sectional study. *EBioMedicine*. 2016;3:135-140.
69. Zetterberg H, Hietala MA, Jonsson M, et al. Neurochemical aftermath of amateur boxing. *Archives of neurology*. 2006;63(9):1277-1280.
70. Kuhle J, Gaiottino J, Leppert D, et al. Serum neurofilament light chain is a biomarker of human spinal cord injury severity and outcome. *Journal of Neurology, Neurosurgery & Psychiatry*. 2015;86(3):273-279.

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60
71. Wingo AP, Fan W, Duong DM, et al. Shared proteomic effects of cerebral atherosclerosis and Alzheimer's disease on the human brain. *Nature neuroscience*. 2020;23(6):696-700.
72. Sternlicht MD, Werb Z. How matrix metalloproteinases regulate cell behavior. *Annual review of cell and developmental biology*. 2001;17(1):463-516.
73. Chen Q, Jin M, Yang F, Zhu J, Xiao Q, Zhang L. Matrix metalloproteinases: inflammatory regulators of cell behaviors in vascular formation and remodeling. *Mediators of inflammation*. 2013;2013
74. Brkic M, Balusu S, Libert C, Vandenbroucke RE. Friends or foes: matrix metalloproteinases and their multifaceted roles in neurodegenerative diseases. *Mediators of inflammation*. 2015;2015
75. Lakhan SE, Kirchgessner A, Tepper D. Matrix metalloproteinases and blood-brain barrier disruption in acute ischemic stroke. *Frontiers in neurology*. 2013;4:42109.
76. Lim NK-H, Villemagne VL, Soon CP, et al. Investigation of matrix metalloproteinases, MMP-2 and MMP-9, in plasma reveals a decrease of MMP-2 in Alzheimer's disease. *Journal of Alzheimer's Disease*. 2011;26(4):779-786.
77. Lorenzl S, Albers DS, Relkin N, et al. Increased plasma levels of matrix metalloproteinase-9 in patients with Alzheimer's disease. *Neurochemistry international*. 2003;43(3):191-196.
78. Perveen A, Mir RA, Gulfishan M, Hafeez A. MMP-3 And MMP-9 Concentrations In Alzheimer's Disease: A Cerebrospinal Fluid And Serum Analysis. *Educational Administration: Theory and Practice*. 2024;30(4):6248-6254.
79. Rosenberg GA, Sullivan N, Esiri MM. White matter damage is associated with matrix metalloproteinases in vascular dementia. *Stroke*. 2001;32(5):1162-1168.
80. Hartz AM, Bauer B, Soldner EL, et al. Amyloid- $\beta$  contributes to blood-brain barrier leakage in transgenic human amyloid precursor protein mice and in humans with cerebral amyloid angiopathy. *Stroke*. 2012;43(2):514-523.
81. Hunninghake GM, Cho MH, Tesfaigzi Y, et al. MMP12, lung function, and COPD in high-risk populations. *New England Journal of Medicine*. 2009;361(27):2599-2608.
82. Olejarz W, Łacheta D, Kubiak-Tomaszewska G. Matrix metalloproteinases as biomarkers of atherosclerotic plaque instability. *International journal of molecular sciences*. 2020;21(11):3946.
83. Iyer RP, Patterson NL, Zouein FA, et al. Early matrix metalloproteinase-12 inhibition worsens post-myocardial infarction cardiac dysfunction by delaying inflammation resolution. *International journal of cardiology*. 2015;185:198-208.
84. Pijet B, Konopka A, Rejmak E, et al. The matrix metalloproteinase inhibitor marimastat inhibits seizures in a model of kainic acid-induced status epilepticus. *Scientific Reports*. 2020;10(1):21314.
85. You J, Guo Y, Zhang Y, et al. Plasma proteomic profiles predict individual future health risk. *Nature Communications*. 2023;14(1):7817.
86. Harris SE, Cox SR, Bell S, et al. Neurology-related protein biomarkers are associated with cognitive ability and brain volume in older age. *Nature Communications*. 2020;11(1):800.
87. Llaurador-Coll M, Rios S, García-Gavilán JF, Babio N, Vilella E, Salas-Salvadó J. Plasma levels of neurology-related proteins are associated with cognitive performance in an older population with overweight/obesity and metabolic syndrome. *GeroScience*. 2023;45(4):2457-2470.
88. Nichols E, Gross AL, Zhang YS, et al. Considerations for the use of the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) in cross-country comparisons of cognitive aging and dementia. *Alzheimer's & Dementia*. 2024;

89. van der Flier WM, Scheltens P. The ATN framework—moving preclinical Alzheimer disease to clinical relevance. *JAMA neurology*. 2022;79(10):968-970.

90. Langa KM, Foster NL, Larson EB. Mixed dementia: emerging concepts and therapeutic implications. *Jama*. 2004;292(23):2901-2908.

## Figure legends

**Figure 1. Volcano plot shows the Hazard ratio (x-axis) and two-sided P values (y-axis) for the association of protein concentration with incident all-cause dementia using imputed data.**

X-axis displays the hazard ratios from Cox Proportional Hazard Regression models adjusted for age, sex, education, ethnicity, smoking status, depression, cardiovascular disease, body mass index, systolic blood pressure, low-density lipoprotein (LDL) cholesterol, in a sample size of 3249.

Y-axis displays the nominal uncorrected p-value ( $-\log_{10}$ ). Proteins above the horizontal dotted red line were significantly associated with incident all-cause dementia with false discovery rate (FDR)-corrected p-value  $< 0.05$ .

**Figure 2. Predictive accuracy of NEFL and RPS6KB1, alone or in combination with demographic variables, Apolipoprotein E 4 (APOE 4) status, and memory score for all-cause dementia.**

Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curves illustrate the performance of various variable models in predicting the incidence of all-cause dementia in a sample size of 3249. Demographics variables included sex, age, education, and ethnicity. Memory score included a combined test score of immediate recall and delayed recall.

**Figure 3. Protein importance ranking using XGBoost decision tree-based machine learning algorithm and SHAP visualization for selected features on all-cause dementia.**

A) SHapley Additive exPlanations (SHAP) values from eXtreme Gradient Boosting (XGBoost) model displaying the top 20 selected features in a sample size of 3249. The Y-axis indicates the feature names in order of importance ranked from top to bottom. The X-axis represents the SHAP value, which indicates the degree of change in log odds. The width of the range of the horizontal bars showed the extent of the contribution to the prediction of all-cause dementia. The colour of each point on the graph represents the value of the corresponding feature, with orange indicating high values and purple indicating low values. The direction on the x-axis indicates the likelihood of developing all-cause dementia towards the right, and likelihood of free from dementia towards the left. B) Mean absolute SHAP values for the top 20 selected features derived from XGBoost model in a sample size of 3249.



1  
2  
3 **Figure 4. Forest Plots for the associations between identified proteins from ELSA and**  
4 **dementia and dementia subtypes validated in the UK Biobank.**  
5

6 Multiple adjusted hazard ratios and 95% confidence intervals (95% CI) from Cox Proportional  
7 Hazard Regression models for NEFL, KIM1 (HAVCR1), MMP12, EDA2R and the  
8 associations with: A) all-cause dementia; B) Alzheimer's disease; C) vascular dementia; D)  
9 frontotemporal dementia. All models adjusted for age, sex, education, ethnicity, smoking  
10 status, depression, cardiovascular disease, body mass index, systolic blood pressure, low-  
11 density lipoprotein (LDL) cholesterol, in a sample size of 52,745. P values were false discovery  
12 rate (FDR) corrected.  
13  
14  
15  
16  
17

18 **Figure 5. Enrichment analysis of the identified proteins in Genotype-Tissue Expression**  
19 **(GTEx) 2023, Illuminating the Druggable Genome (IDG) drug target 2022, and**  
20 **Proteomics Drug Atlas (PDA) 2023.**  
21

22 Enrichment for Genotype-Tissue Expression (GTEx) 2023, Illuminating the Druggable  
23 Genome (IDG) drug target 2022, and Proteomics Drug Atlas (PDA) 2023. Significant proteins  
24 after false discovery rate (FDR) correction (denoted as  $P_{FDR}$ ) derived from Cox proportional  
25 hazard regressions in minimally- and fully adjusted models were fed into Enrichr  
26 (<https://maayanlab.cloud/enrichr/>) for enrichment analysis. The full list of proteins from ELSA  
27 was used as the background gene set. Terms above the horizontal dotted red line were enriched  
28 after FDR-correction with p-value < 0.05, and the text were highlighted in red.  
29  
30  
31  
32

33 **Figure 6. Enrichment analysis of the functional annotations in identified proteins in Gene**  
34 **Ontology (GO) 2023, Kyoto Encyclopaedia of Genes and Genomes (KEGG) 2021, and**  
35 **Reactome pathways 2022.**  
36

37 Enrichment for Gene Ontology (GO) 2023 (GO\_MF: Gene Ontology Molecular function),  
38 Kyoto Encyclopaedia of Genes and Genomes (KEGG) 2021, and Reactome pathways 2022.  
39 Significant proteins after false discovery rate (FDR) correction (denoted as  $P_{FDR}$ ) derived from  
40 Cox proportional hazard regressions in minimally- and fully adjusted models were fed into  
41 Enrichr (<https://maayanlab.cloud/enrichr/>) for enrichment analysis. The full list of proteins  
42 from ELSA was used as the background gene set. Terms above the horizontal dotted red line  
43 were enriched after FDR-correction with p-value < 0.05, and the text were highlighted in red.  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

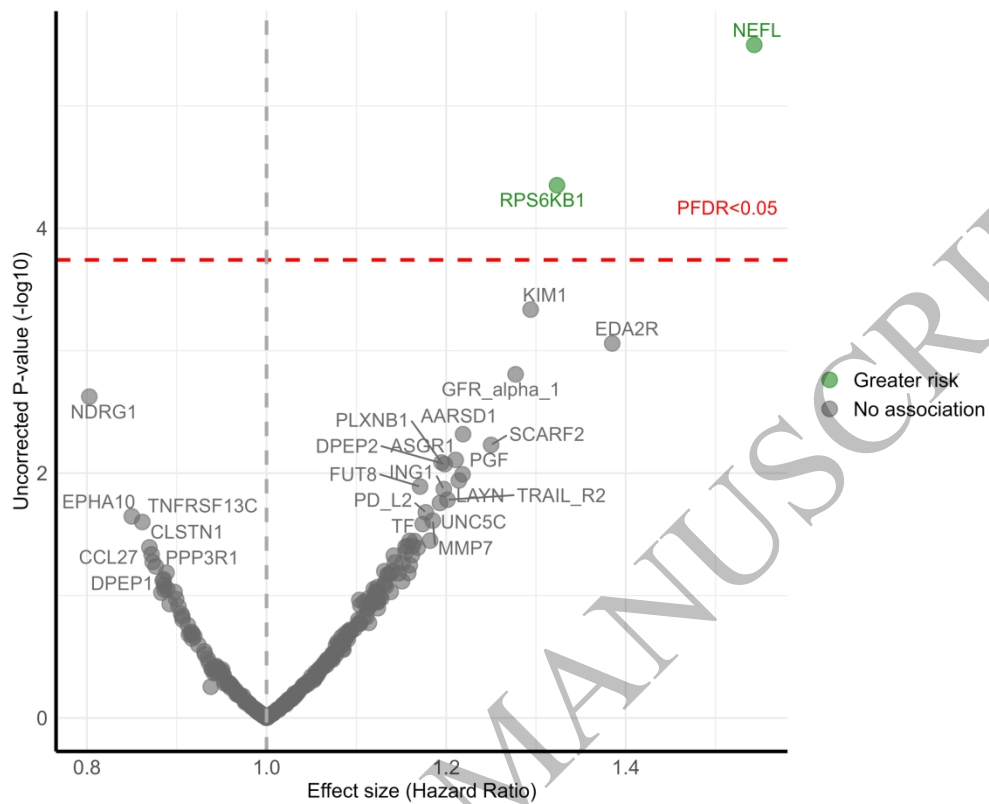


Figure 1

209x169mm (300 x 300 DPI)

ACCEPTED MANUSCRIPT

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

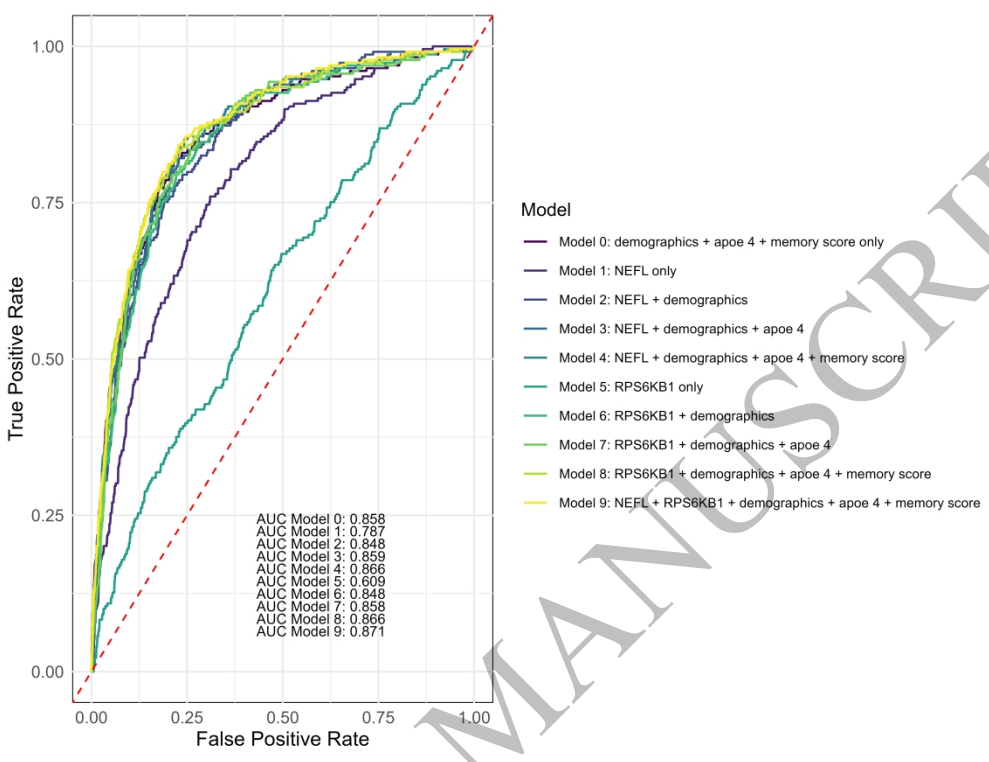


Figure 2

533x431mm (315 x 315 DPI)

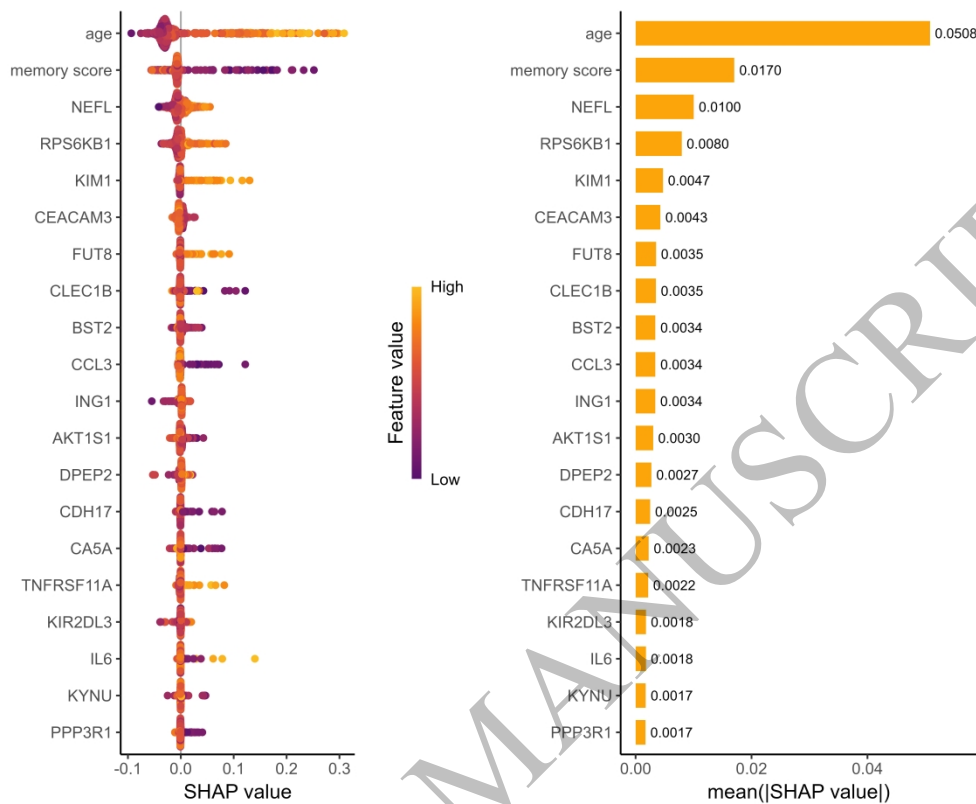


Figure 3

533x431mm (315 x 315 DPI)

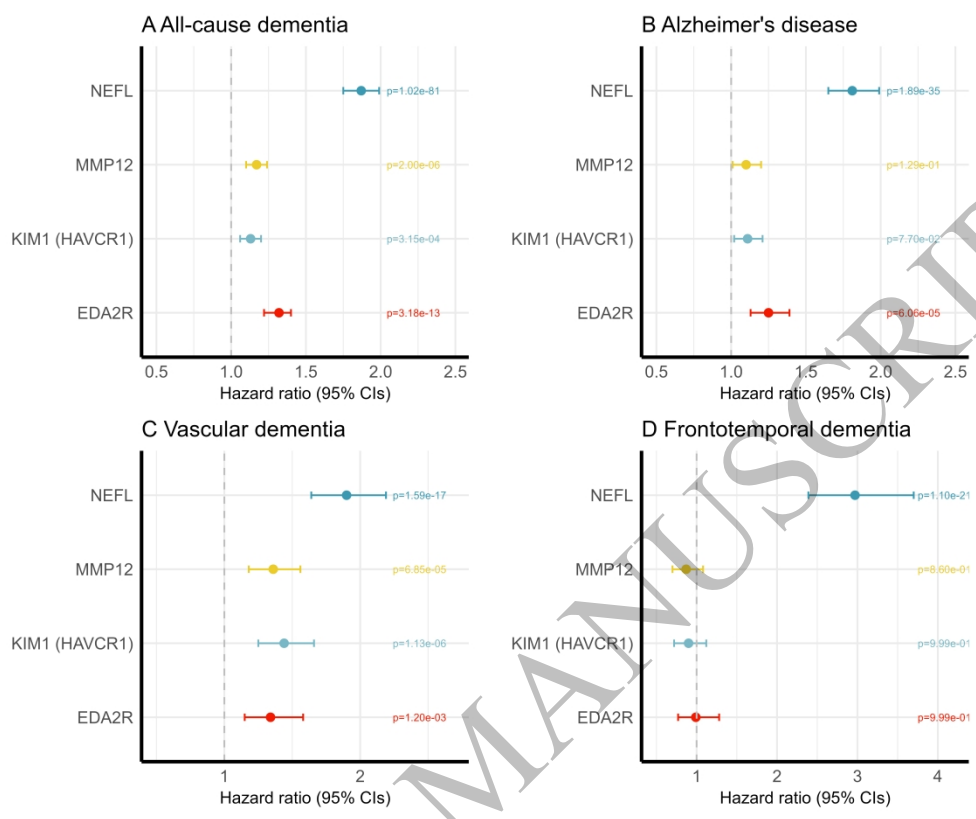


Figure 4

533x431mm (315 x 315 DPI)

ACCEPTED MANUSCRIPT

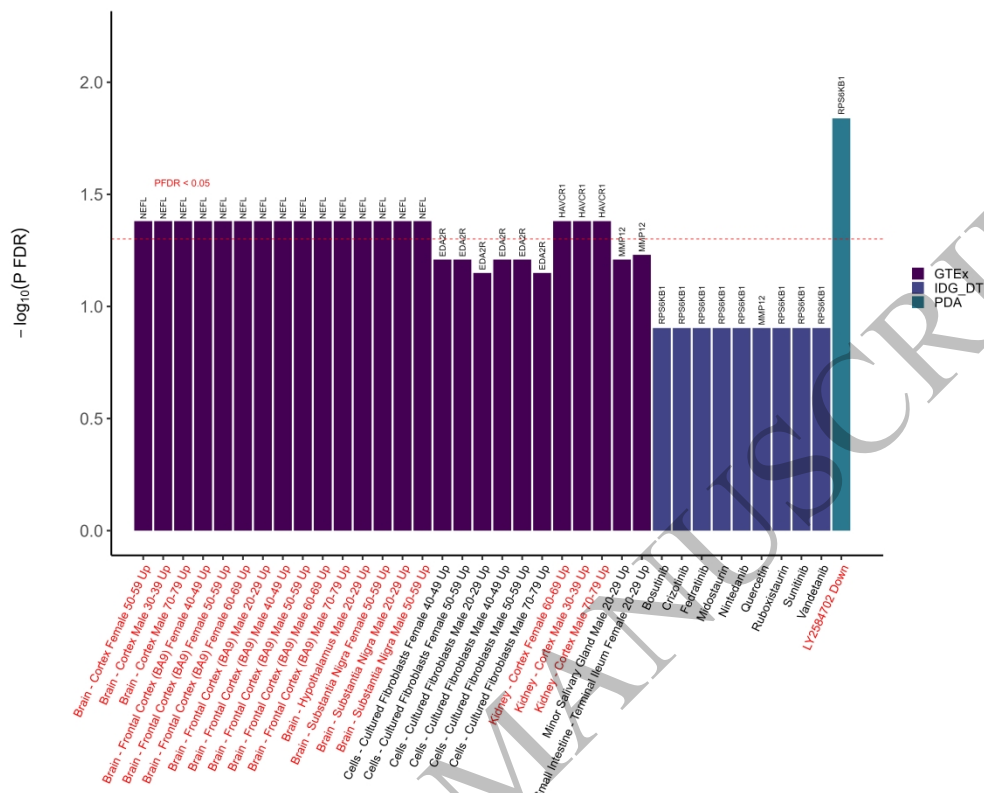


Figure 5

533x431mm (315 x 315 DPI)

ACCEPTED MANUSCRIPT

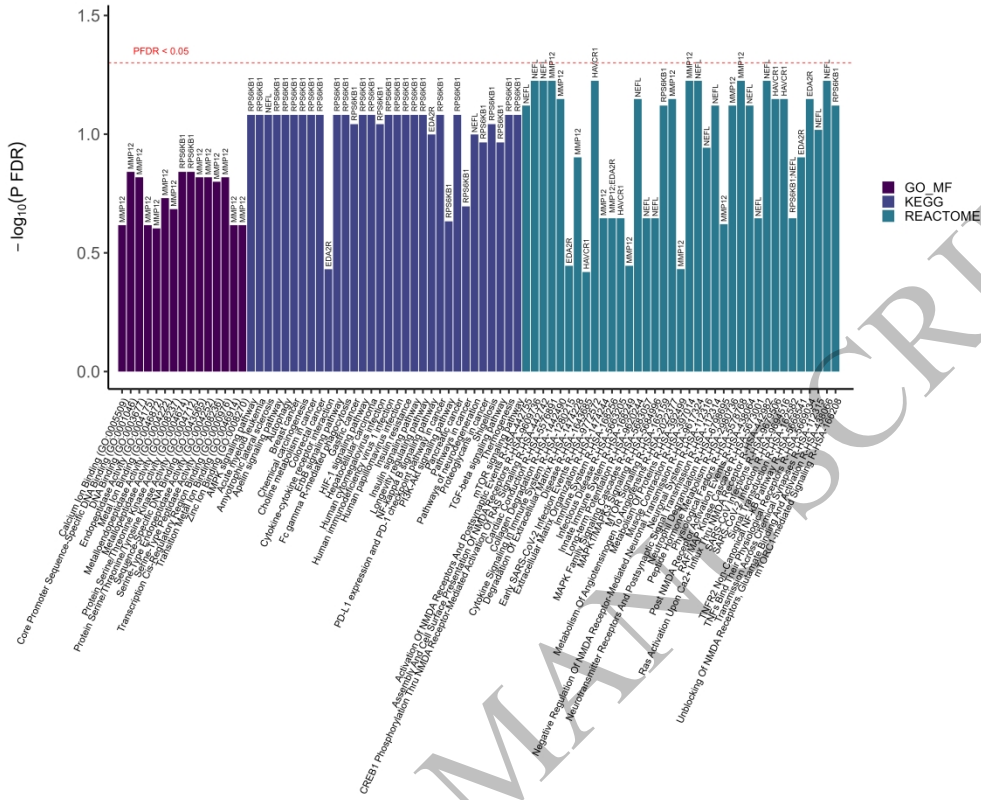
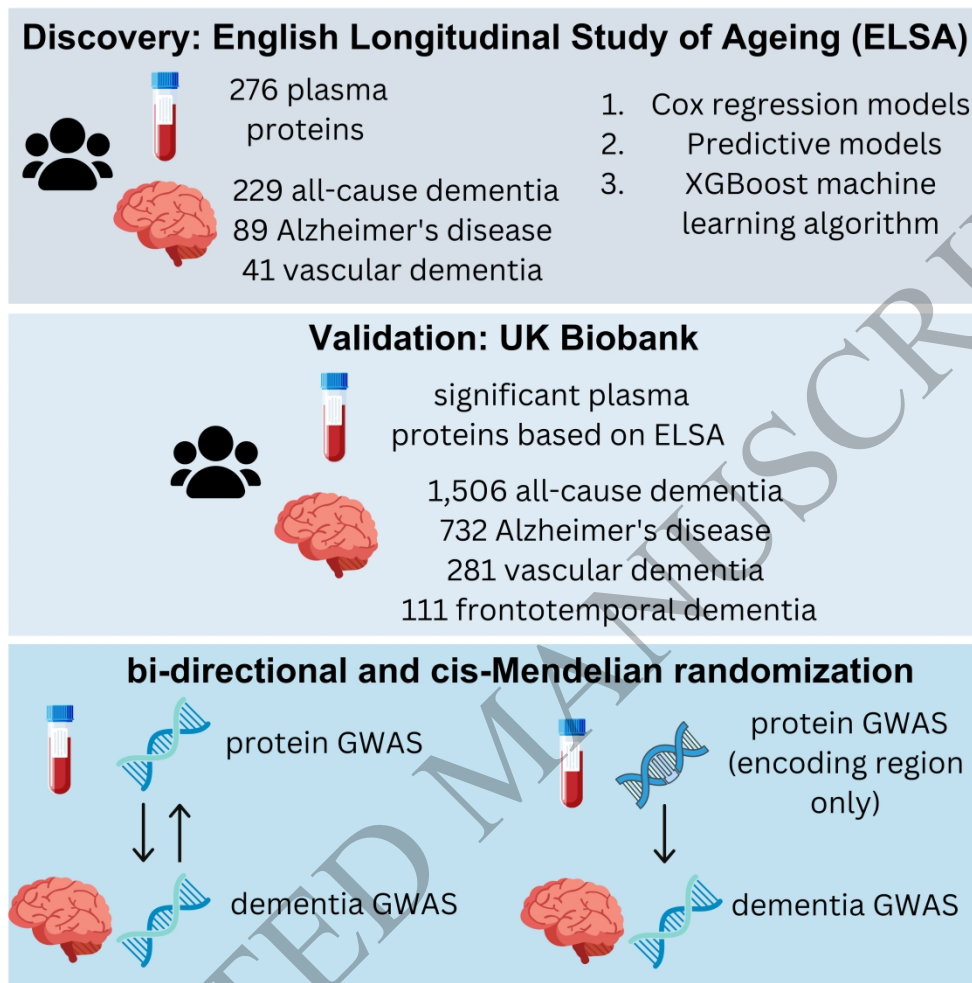


Figure 6

533x431mm (315 x 315 DPI)

ACCEPTED MANUSCRIPT



Graphical abstract

304x304mm (300 x 300 DPI)