nature aging

Article

https://doi.org/10.1038/s43587-023-00565-0

Plasma proteomic profiles predict future dementia in healthy adults

Received: 9 June 2023

Accepted: 22 December 2023

Published online: 12 February 2024

Check for updates

Yu Guo[1,4], Jia You[1,2,4], Yi Zhang[1,4], Wei-Shi Liu[1], Yu-Yuan Huang[1], Ya-Ru Zhang[1], Wei Zhang[2], Qiang Dong[1], Jian-Feng Feng[2,3] ✉, Wei Cheng[1,2,3] ✉ & Jin-Tai Yu[1] ✉

The advent of proteomics offers an unprecedented opportunity to predict dementia onset. We examined this in data from 52,645 adults without dementia in the UK Biobank, with 1,417 incident cases and a follow-up time of 14.1 years. Of 1,463 plasma proteins, GFAP, NEFL, GDF15 and LTBP2 consistently associated most with incident all-cause dementia (ACD), Alzheimer's disease (AD) and vascular dementia (VaD), and ranked high in protein importance ordering. Combining GFAP (or GDF15) with demographics produced desirable predictions for ACD (area under the curve (AUC) = 0.891) and AD (AUC = 0.872) (or VaD (AUC = 0.912)). This was also true when predicting over 10-year ACD, AD and VaD. Individuals with higher GFAP levels were 2.32 times more likely to develop dementia. Notably, GFAP and LTBP2 were highly specific for dementia prediction. GFAP and NEFL began to change at least 10 years before dementia diagnosis. Our findings strongly highlight GFAP as an optimal biomarker for dementia prediction, even more than 10 years before the diagnosis, with implications for screening people at high risk for dementia and for early intervention.

Dementia progresses slowly from the asymptomatic stage to a fully expressed clinical syndrome over many years[1,2]. Because no effective therapy is currently available, correctly determining whether a person will progress to dementia in the near future has become a public health priority[3]. This task is of the utmost importance for the timely referral of at-risk populations and for subsequent early diagnosis and prompt intervention. Nonetheless, it remains a major challenge for clinicians, and it is not known how to best predict the onset of dementia. A possible turning point has recently emerged with the advancement of blood-based biomarkers, which could serve as a preferable tool to facilitate early risk screening in the preclinical phase among the general population[4–6].

Although some blood markers have been proven to be strongly associated with dementia risk[7–14], biomarker discovery efforts have typically focused on one or a small number of proteins because of technical constraints, and lacked the systematic comparison of human proteomics. It has not been established which of the high-performing markers harbors the greatest potential for risk prediction and monitoring. Other investigations have utilized the proteomics strategy to reveal differences in blood proteins between people with and people without dementia[15–17]. However, most of these investigations were cross-sectional and did not take into account the impact of possible reverse causality, nor did they address whether abnormal protein levels were present preceding dementia onset and how long before dementia such abnormalities could be detected. A recent prospective study adopted proteomic analysis to predict incident dementia[18]. Yet, the mixed dementia outcome and the relatively small sample size reduced the power to identify proteins relevant to specific dementia etiologies. Whether and how proteomic patterns differ across the incident dementia subtypes remain unclear. Therefore, large-scale prospective studies

[1]Department of Neurology and National Center for Neurological Disorders, Huashan Hospital, State Key Laboratory of Medical Neurobiology and MOE Frontiers Center for Brain Science, Shanghai Medical College, Fudan University, Shanghai, China. [2]Institute of Science and Technology for Brain-inspired Intelligence, Fudan University, Shanghai, China. [3]Key Laboratory of Computational Neuroscience and Brain-inspired Intelligence, Fudan University, Ministry of Education, Shanghai, China. [4]These authors contributed equally: Yu Guo, Jia You, Yi Zhang. ✉e-mail: jianfeng64@gmail.com; wcheng@fudan.edu.cn; jintai_yu@fudan.edu.cn

with data on blood proteomics and specific dementias (for example, AD) are crucial and necessary.

Furthermore, the predictive ability of the proteins, separately or in combination, in different incidence time groups (for example, 10 years, >10 years) has been neglected to date. However, this is particularly important for the ultra early detection of dementia and for substantially advancing the window for prevention and intervention. Beyond predictive accuracy, optimal dementia prediction biomarkers should be highly specific for the corresponding pathology[11,19]. Disappointingly, blood proteomic biomarkers that can predict future dementia with the required sensitivity and specificity remain largely undetermined[4,6].

We innovatively employed a data-driven proteomic approach in a large prospective cohort with long follow-up to identify the plasma biomarkers best associated with dementia prediction and explore their predictive performance. The recent release of data on 1,463 plasma proteins from more than 50,000 individuals in the UK Biobank (UKB) provides us with an unprecedented opportunity to: (1) comprehensively test their associations with incident ACD, AD and VaD to identify a set of candidate dementia-associated proteins; (2) determine the magnitude of the protein contributions to the prediction of dementia; (3) investigate the predictive accuracy of the top-ranked proteins, individually and in combination, over 5, 10 and many more years; (4) examine the relationships between plasma proteins and the risk of clinical progression, and further evaluate whether such relationships are specific to dementia and not seen in people without dementia; and (5) trace the trajectories of plasma proteins back from the time of dementia diagnosis and assess when each protein begins to deviate from normal control values.

Results

Participants' characteristics

This study included 52,645 adults without dementia at baseline, with a median age of 58 years, of whom 53.9% were female and 93.7% were of white ancestry (Table 1). During a median follow-up of 14.1 years, 1,417 (2.7%) incident dementia cases were identified, of which 219 occurred within 5 years, 833 within 10 years and 584 beyond 10 years. For incident ACD participants, the median age was 66 years, 48.5% were female and 96.5% were of white ethnicity (all $P < 0.001$). There were 691 patients diagnosed with AD, among whom 384 had incidents within 10 years and 307 had incidents over 10 years. In addition, 285 patients were diagnosed with VaD, among whom 148 had incidents within 10 years and 137 had incidents over a decade. The incidence of ACD per 1,000 person-years from age 39 to 70 years was 2.00. In the 60–64-year age group, the incidence rate of ACD was 2.07 per 1,000 person-years, and it was 6.26 per 1,000 person-years in the 65–69-year age group.

Identifying proteins associated with incident dementia

Of the 1,463 proteomic biomarkers tested, after adjusting for age, sex, education and *APOE* ε4 alleles in model 1, we found 184, 16 and 139 proteins were remarkably associated with incident ACD, AD and VaD, respectively (Fig. 1a and Supplementary Table 2). Other proteins, such as MAPT, were positively associated with the risk of ACD ($P = 0.002$), AD ($P = 0.043$) and VaD ($P = 0.038$), but the associations were not significant after Bonferroni corrections. As a sensitivity analysis, we re-ran all analyses under model 2, which additionally adjusted for vascular variables, and found several significant associations could be replicated (Fig. 1a). Importantly, after Bonferroni corrections, GFAP (ACD: hazard ratio (HR) = 1.53, $P = 1.35 \times 10^{-91}$; AD: HR = 1.65, $P = 9.86 \times 10^{-75}$; VaD: HR = 1.61, $P = 2.06 \times 10^{-31}$) and NEFL (ACD: HR = 1.56, $P = 7.27 \times 10^{-76}$; AD: HR = 1.52, $P = 3.25 \times 10^{-27}$; VaD: HR = 1.56, $P = 1.92 \times 10^{-15}$) had the most significant associations with the studied dementia types. Higher levels of GDF15 (ACD: HR = 1.28, $P = 5.91 \times 10^{-16}$; AD: HR = 1.19, $P = 0.038$; VaD: HR = 1.43, $P = 4.99 \times 10^{-8}$) and LTBP2 (ACD: HR = 1.24, $P = 7.97 \times 10^{-9}$; AD: HR = 1.26, $P = 2.05 \times 10^{-4}$; VaD: HR = 1.35, $P = 0.010$) could also increase the risk of incident dementia (model 2). CST5, NPTXR and BCAN were associated with AD incidence. Other proteins linked to VaD risk were EPHA2, GFRA1 and SPON2, among others. The main results we obtained from the total population largely remained in *APOE* ε4 carrier and noncarrier subgroups (Supplementary Table 3). Enrichment analyses implicated several biological pathways related to the significant proteins, such as extracellular matrix organization, immune system and infectious diseases (Fig. 1b and Supplementary Table 4).

Protein importance ranking

For those dementia-associated proteins in both models 1 and 2, we further sorted them based on their importance to the prediction task. Detailed analytic results are presented in Supplementary Table 5. As shown in the bar chart (Fig. 2a and Supplementary Fig. 2), plasma GFAP, NEFL and GDF15 consistently ranked highest in predicting either ACD or its subtypes. In detail, NEFL was the strongest predictor of ACD, followed by GFAP and GDF15. GFAP was the strongest predictor of AD, followed by NEFL and GDF15. GDF15 was the strongest predictor of VaD, followed by NEFL and GFAP. When the top few proteins were included, the predictive power for dementia (AUC on the right axis) escalated steeply and gradually fell into a flat fluctuation as more proteins entered. Using this sequential forward selection scheme, we ultimately chose the top 11 (NEFL, GFAP, GDF15, BCAN, LTBP2, NPTXR, EDA2R, NTproBNP, EGFR, HPGDS and CST5), 7 (GFAP, NEFL, GDF15, LTBP2, BCAN, NPTXR and CST5) and 4 (GDF15, NEFL, GFAP and MMP12) proteins for ACD, AD and VaD prediction, respectively, for subsequent analyses.

To have an intensive survey of the incidence time, we further categorized patients into 5-year, 10-year and over 10-year incidents. Importance ranking procedures were repeated independently for these target populations, and the yielded results were fed into corresponding later analyses. The final selected important proteins were highly overlapped with those chosen for predicting all incident dementia events.

Shapley additive explanations (SHAP) plots were leveraged to intuitively interpret the effect of each selected protein by its value magnitude (coded by a gradient of colors) and tendency direction on the horizontal axis (the likelihood of developing dementia) (Fig. 2b and Supplementary Fig. 2). The protein GFAP, for example, appeared to hold the widest range for all incident AD, suggesting that it had the most considerable predictive power. In addition, participants with higher GFAP levels (colored in red) were more likely to develop AD (right side), whereas those with lower GFAP levels (blue) tended to remain healthy (left). Similar explanations were given for the rest of the proteins.

Predictive accuracy of plasma proteins

Employing the tenfold cross-validation approach, we then examined the predictive accuracy of the above-selected important proteins for future dementia. Relevant results are summarized in Supplementary Tables 6 and 7. As for all incident ACD, plasma NEFL, GFAP or GDF15 alone produced modest AUC values of 0.746, 0.718 and 0.712, respectively, which was similar to or higher than those achieved by other separate proteins. This revealed the potential of these proteins to aid in dementia prediction. To achieve higher predictive accuracy, we explored the performance of plasma proteins in combination with other readily available measures including demographic indicators (age, sex, education and *APOE* ε4 status) and cognitive tests (pairs matching time and reaction time). The most marked increase in accuracy was seen when NEFL (AUC = 0.898) or GFAP (AUC = 0.900) was combined with demographic features and brief cognitive tests (DeLong test $P < 0.001$; Fig. 3a). Notable improvement was also observed when combining NEFL (AUC = 0.890) or GFAP (AUC = 0.891) with demographic indicators, whereas combining NEFL (AUC = 0.779) or GFAP (AUC = 0.768) with cognitive metrics resulted in less prediction improvement.

Adding proteins to demographic features significantly improved the prediction for incident all-time or 10-year ACD and AD (ΔAUC ranged from 0.009 to 0.028, DeLong test $P < 0.017$). Applying the

https://doi.org/10.1038/s43587-023-00565-0

Table 1 | Baseline characteristics of UK Biobank participants included in the study

Participants' characteristics	Overall	Control	Incident dementia		Incident AD		Incident VaD	
	N=52,645	*N*=51,228	*N*=1,417	*P* value	*N*=691	*P* value	*N*=285	*P* value
Age, years	58 [50–64]	58 [50–63]	66 [63–68]	4.33×10^{-289}	67 [63–68]	3.79×10^{-163}	66 [63–68]	1.38×10^{-68}
Sex (female)	28,393 (53.9)	27,706 (54.1)	687 (48.5)	3.39×10^{-5}	379 (54.8)	0.717	111 (38.9)	4.34×10^{-7}
Ethnicity (white)	49,353 (93.7)	47,985 (93.7)	1,368 (96.5)	1.36×10^{-5}	667 (96.5)	0.003	277 (97.2)	0.020
Education, years	11 [10–15]	11 [10–15]	11 [10–12]	7.50×10^{-22}	10 [9–12]	8.46×10^{-18}	11 [11,12]	1.37×10^{-5}
APOE ε4 single-copy carriers	13,610 (25.9)	13,041 (25.5)	569 (40.2)	1.39×10^{-263}	308 (44.6)	8.57×10^{-290}	116 (40.7)	9.35×10^{-46}
APOE ε4 double-copies carriers	1,474 (2.8)	1,244 (2.4)	230 (16.2)		160 (23.2)		40 (14.0%)	
Systolic blood pressure, mmHg	138 [126–152]	138 [125–152]	145 [132–159]	5.34×10^{-36}	145 [132–159]	9.08×10^{-23}	147 [132–159]	2.72×10^{-9}
Hypertension treatment	11,648 (22.1)	11,078 (21.6)	570 (40.2)	6.16×10^{-62}	275 (39.8)	2.84×10^{-30}	131 (46.0)	6.26×10^{-23}
Diabetes	2,979 (5.7)	2,783 (5.4)	196 (13.8)	3.49×10^{-41}	94 (13.6)	2.42×10^{-20}	53 (18.6)	9.14×10^{-22}
Current smoker	5,562 (10.6)	5,424 (10.6)	138 (9.7)	0.326	60 (8.7)	0.120	31 (10.9)	0.951
Atrial fibrillation	1,140 (2.2)	1,072 (2.1)	68 (4.8)	9.65×10^{-12}	26 (3.8)	0.004	17 (6.0)	1.52×10^{-5}
Coronary heart disease	3,244 (6.2)	3,004 (5.9)	240 (16.9)	3.91×10^{-65}	99 (14.3)	2.43×10^{-20}	61 (21.4)	7.97×10^{-28}
Heart failure	481 (0.9)	457 (0.9)	24 (1.7)	0.003	13 (1.9)	0.012	6 (2.1)	0.064
Stroke	936 (1.8)	873 (1.7)	63 (4.4)	2.90×10^{-14}	24 (3.5)	6.78×10^{-4}	23 (8.1)	1.58×10^{-15}
Peripheral artery disease	1,191 (2.3)	1,132 (2.2)	59 (4.2)	1.68×10^{-6}	21 (3.0)	0.180	20 (7.0)	1.34×10^{-7}
Total cholesterol, mmol l^{-1}	5.6 [4.9–6.4]	5.6 [4.9–6.4]	5.4 [4.6–6.2]	4.92×10^{-10}	5.5 [4.6–6.4]	0.006	5.3 [4.4–6.0]	9.80×10^{-7}
HDL cholesterol, mmol l^{-1}	1.4 [1.2–1.6]	1.4 [1.2–1.6]	1.4 [1.2–1.6]	0.003	1.4 [1.2–1.6]	0.856	1.4 [1.1–1.5]	9.44×10^{-4}
Body mass index	26.8 [24.2–29.9]	26.8 [24.2–29.9]	27.2 [24.4–30.2]	0.048	26.9 [24.2–29.8]	0.787	28.2 [25.2–31.3]	4.59×10^{-4}
Pairs matching time, s	189 [149–247]	189 [149–246]	229 [172–326]	2.70×10^{-74}	233 [173–330]	2.09×10^{-48}	233 [179–342]	2.80×10^{-19}
Reaction time, ms	543 [484–617]	540 [481–614]	594 [531–687]	1.17×10^{-81}	594 [529–685]	7.83×10^{-43}	605 [543–696]	3.14×10^{-27}

Continuous data are presented as median [interquartile range] and categorical variables as number (percentage). Differences between incident dementia and healthy control groups were compared using Student's *t*-test for continuous variables and Pearson's chi-squared test for discrete variables.

bootstrap method to compare AUC values produced similar conclusions, and additionally revealed an improvement in discriminative performance with the addition of GDF15 to demographic features in ACD, and NEFL to demographic features in AD.

For assessing the combined power of selected important proteins, we utilized the protein panel or protein risk score (ProRS) instead of the single protein, by which significantly better accuracies were attained. Specifically, the protein panel generated an AUC of 0.841 (Fig. 3a), similar to that achieved by ProRS of 0.837. Adding demographic features further improved accuracies substantially, raising the AUC values to 0.908, whereas adding cognitive features to the protein panel increased the accuracy to only 0.850. The best-performing model was the full model, combining the protein panel with demographic and cognitive information, with an AUC of 0.913. Integrating ProRS with demographic and cognitive data also provided an excellent AUC of 0.903.

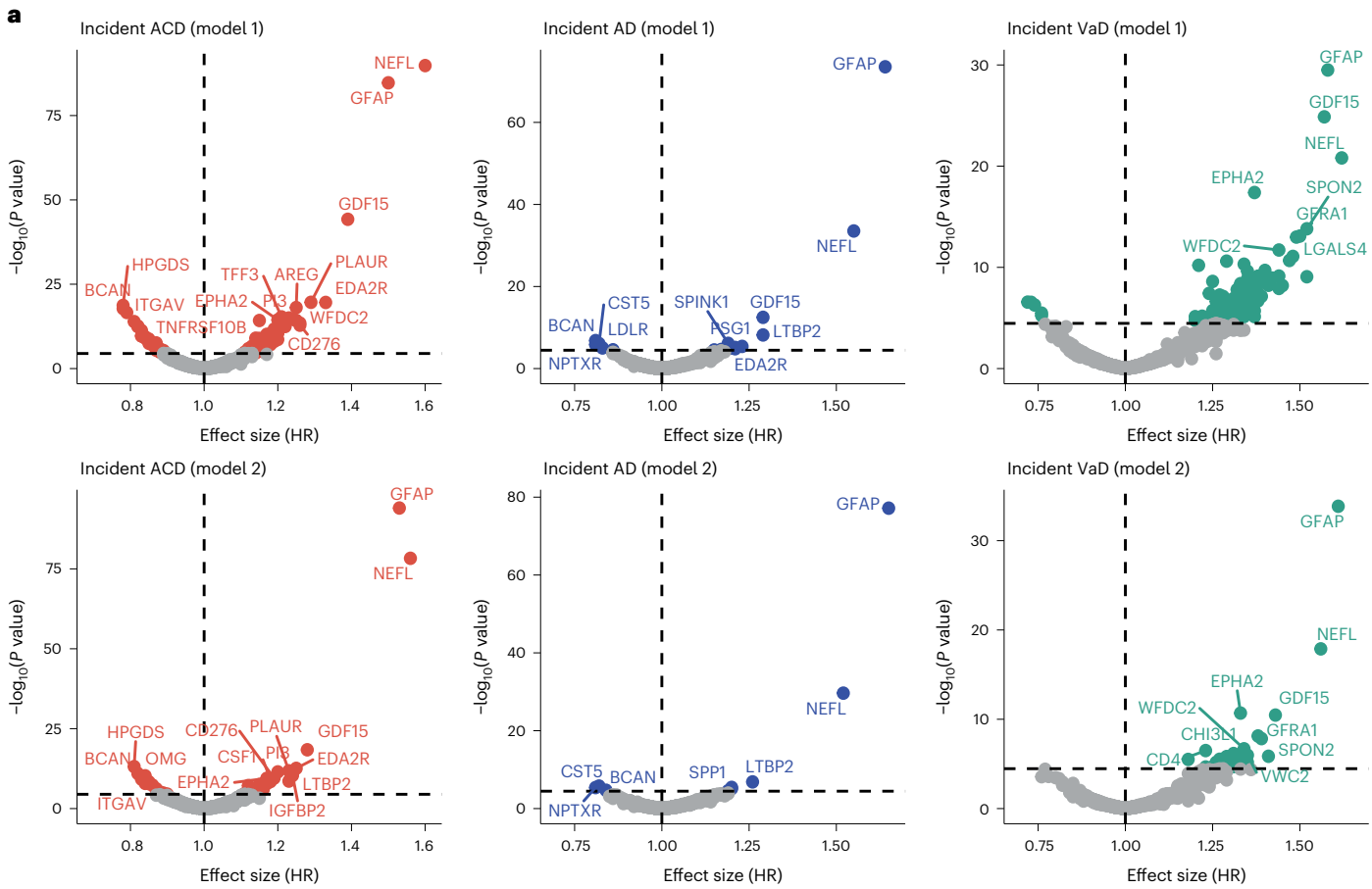
The model fit for predicting all incident AD and VaD resulted in similarly included variables and accuracies (Fig. 3b–c). The combination of GFAP (or GDF15) with demographic characteristics achieved a good prediction for AD (AUC = 0.872) (or VaD (AUC = 0.912)). Further adding cognitive tests did not improve the predictive power significantly (AD: AUC = 0.878; VaD: AUC = 0.914). When applying the same variable models to predict 5-year, 10-year and over 10-year (Fig. 3d–f) incident dementia events or removing individuals who developed dementia in the first two years of follow-up, similar and robust results were produced. Of particular note, combining plasma GFAP with demographic characteristics gave an accurate prediction of incidences of ACD (AUC = 0.872) and AD (AUC = 0.847) over 10 years, with comparable performance to the full model integrating GFAP or the protein panel with demographic and cognitive data. A combination of plasma GDF15 with demographic characteristics (AUC = 0.895) also achieved an excellent prediction for over 10-year incident VaD, with similar accuracy to that of the full model integrating GDF15 or the protein panel with demographic and cognitive data. These data suggest that the combined model we derived could enable an accurate prediction of future dementia even more than 10 years before the diagnosis. Unexpectedly, adding cognition resulted in lower AUC values in some models, particularly for over 10-year incident AD and VaD, although this was generally not significant using both DeLong and bootstrap tests.

Blood proteins and the risk of clinical progression

Next, we investigated the prognostic value of baseline plasma proteins for progression to dementia. Baseline protein levels were dichotomized into high and low groups and the cutoff was derived upon the achievement of the largest Youden index when distinguishing those who experienced clinical progression from those who did not throughout the follow-up period. Detailed thresholds were shown in Supplementary Table 8.

Subjects with higher baseline NEFL (HR = 2.36, $P = 1.27\times10^{-41}$), GFAP (HR = 2.32, $P = 8.68\times10^{-47}$) or GDF15 (HR = 1.70, $P = 2.17\times10^{-15}$) levels presented an elevated risk of developing dementia (Fig. 4a). This finding was also pronounced in AD and VaD (Fig. 4b,c). Remarkably, individuals with higher GFAP levels were 2.91 times more likely to develop AD than those with lower baseline GFAP. The likelihood of developing VaD in the future was 2.45 times greater for those with higher GDF15 levels

https://doi.org/10.1038/s43587-023-00565-0



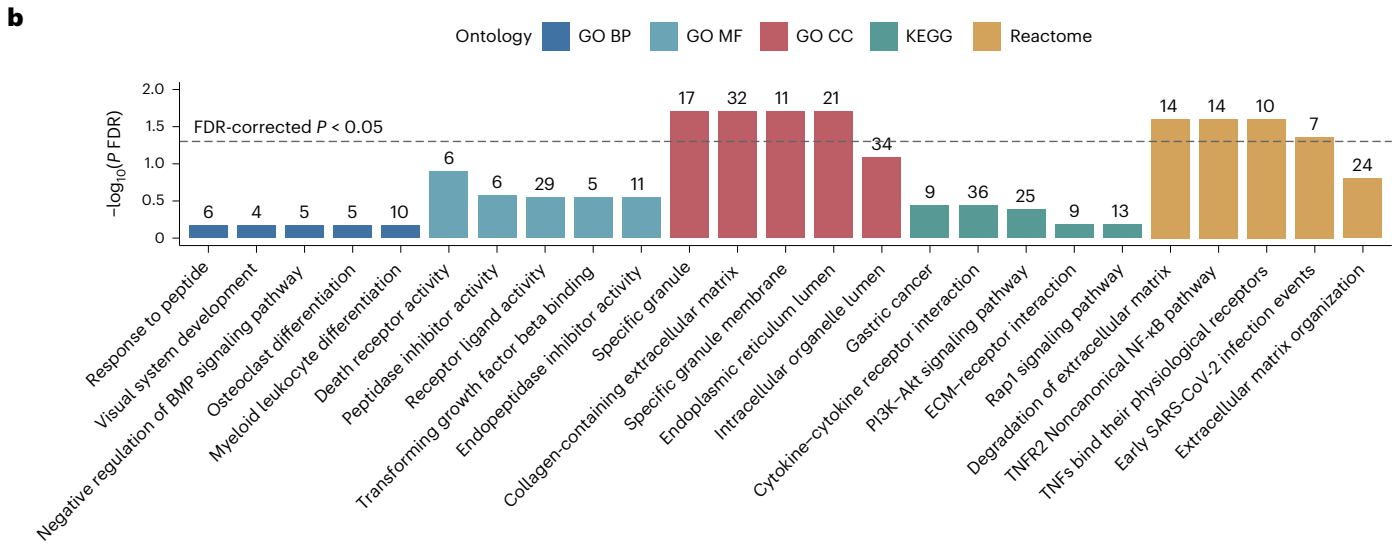


Fig. 1 | Associations of plasma proteins with incident dementia. a, Volcano plots showing the HR (*x* axis) and $-\log_{10}(P\text{ value})$ (*y* axis) for the global associations of 1,463 proteins with incident ACD, AD and VaD. All results for both Cox proportional hazard regression models 1 and 2 are shown here. Model 1 was adjusted for age, sex, education and *APOE* ε4 alleles. Model 2 was additionally adjusted for systolic blood pressure, hypertension treatment, history of diabetes, smoking status, prevalent cardiovascular disease (atrial fibrillation, coronary heart disease, heart failure, stroke or peripheral artery disease), total and HDL cholesterol and body mass index. *P* values were calculated under two-sided tests and no multiple comparisons were applied. Proteins above the horizontal dotted black line were significantly associated with incident dementia after Bonferroni corrections ($P < 0.05$) taking into account the number of proteins tested ($n = 1{,}463$). **b**, Enrichment for Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome pathways. Significant proteins after Bonferroni correction derived from Cox proportional hazard regressions in model 1 or model 2 were fed into the Enrichr website (https://maayanlab.cloud/Enrichr/) for enrichment analysis using the Olink proteins as background gene set. *P* values were calculated under two-sided tests and statistical significance was defined as a false discovery rate (FDR)-corrected $P < 0.05$ (dotted horizontal line). The number above each bar is the number of observed proteins in each pathway. Detailed results were shown in Supplementary Table 4. BP, biological process; CC, cellular component; ECM, extracellular matrix; MF, molecular function; TNF, tumour necrosis factor.

https://doi.org/10.1038/s43587-023-00565-0

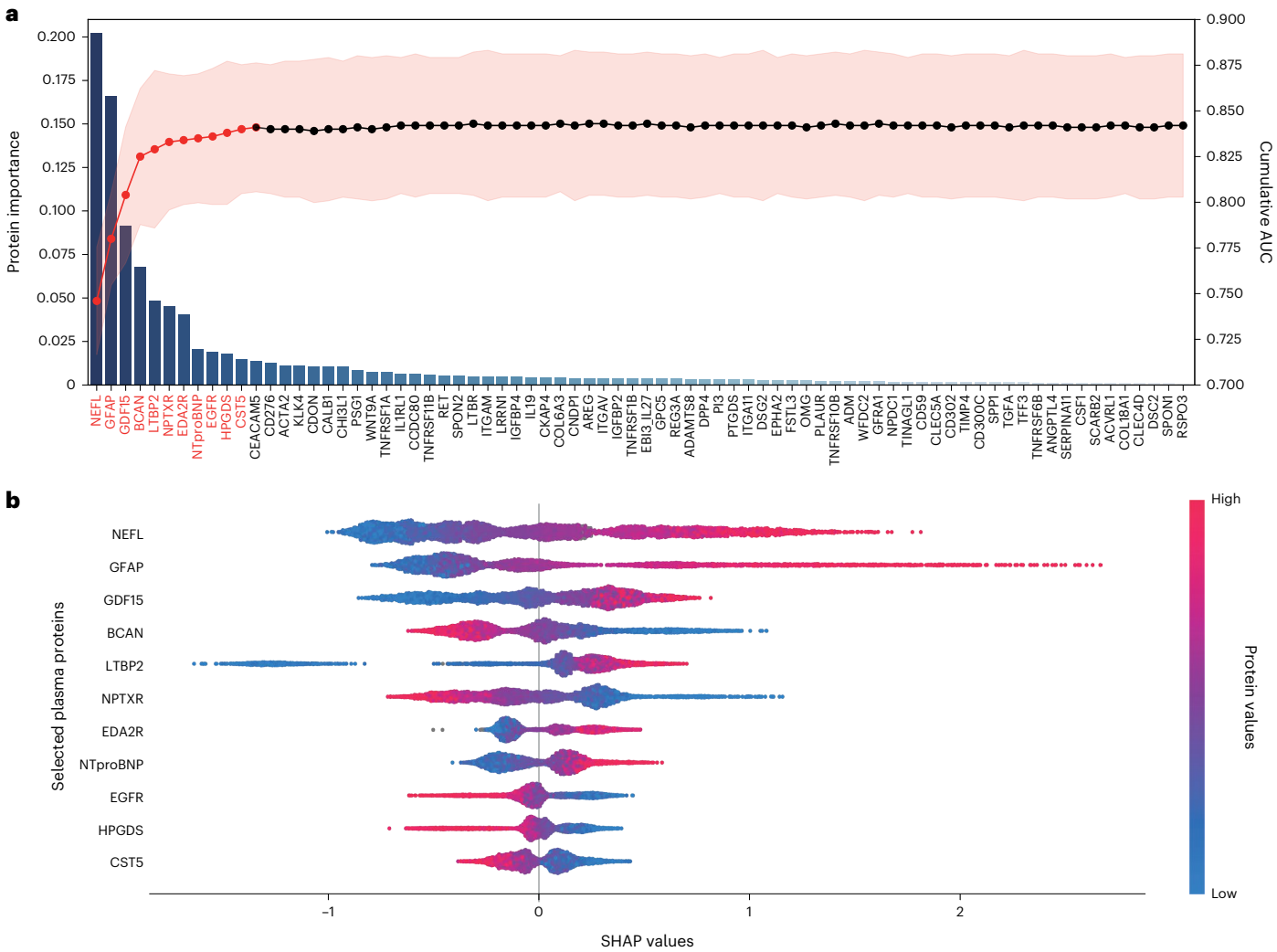


Fig. 2 | Protein importance ranking and SHAP visualization of modeling on all incident dementia populations. a, Sequential forward selection from preselected candidate proteins. The bar chart indicates the importance of the sorted proteins based on their contributions to the prediction of future ACD (as judged by the information gain). The line chart illustrates cumulative AUC values (right axis) upon the inclusion of proteins one by one in each iteration. The top proteins we finally selected are marked in red. Shaded regions represent standard errors derived from cross-validation. **b**, SHAP visualization plot of selected proteins. The width of the range of the horizontal bars can be understood as the extent of the contribution to the prediction of ACD; the wider their range, the greater the contribution. The color of the horizontal bars denotes the magnitude of plasma proteins, which was coded in a gradient from blue (low) to red (high), shown as the color bar on the right-hand side. The direction on the *x* axis indicates the likelihood of developing dementia (right) or being healthy (left).

relative to those with lower levels. Likewise, significant associations were detected for other selected proteins (Supplementary Fig. 3).

In addition, and importantly, we examined the abilities of the plasma proteins in predicting clinical progression to other disease events. As expected, the risk of developing other dementias (except AD and VaD) was elevated for those with higher GFAP levels. It is worth noting that no significant association was observed between baseline GFAP levels and the risk of neurodegenerative diseases (except dementia) (HR [95% confidence interval (CI)] = 1.06 [0.94–1.20], $P > 0.999$), neurological disorders (except dementia) (HR [95% CI] = 0.94 [0.88–1.00], $P = 0.493$) or mental and behavioral disorders (except dementia) (HR [95% CI] = 1.05 [0.95–1.15], $P > 0.999$) (Fig. 4d), indicating that GFAP may be specific for dementia. By contrast, the relationships between baseline GDF15 or NEFL and the risk of almost all these studied disease events were significant. This is also true for other proteins (BCAN, NPTXR, EDA2R, NTproBNP, EGFR, HPGDS and CST5) (Supplementary Fig. 4 and Supplementary Table 8). Subjects with higher baseline LTBP2 levels had a higher risk of developing ACD or AD, but this relationship was not significant for nondementia diseases, which suggested that LTBP2 may be dementia-specific.

Similar trends of disease risk were observed when using the continuous values for each protein in the Cox model (Supplementary Fig. 5).

Results of replication validation analyses

To assess the robustness of our results, we randomly divided the study participants into two-thirds (training set, $n = 35{,}096$) and one-third (testing set, $n = 17{,}549$) sets. The results we obtained (Supplementary Tables 9–11) were consistent with those obtained using the cross-validation strategy that we performed previously. Specifically, from the training set, we reaffirmed the importance of GFAP, NEFL, GDF15 and LTBP2 in the prediction of ACD, AD and VaD. In the testing set, these proteins alone yielded modest predictive accuracies (AUC = 0.7–0.8). Combining GFAP (or GDF15) with demographic data achieved desirable predictions for ACD (AUC = 0.894) and AD (AUC = 0.883) (or VaD (AUC = 0.907)). The same parsimonious models produced AUC values of 0.861, 0.818 and 0.870 for over 10-year ACD, AD and VaD incidence, respectively, with

https://doi.org/10.1038/s43587-023-00565-0

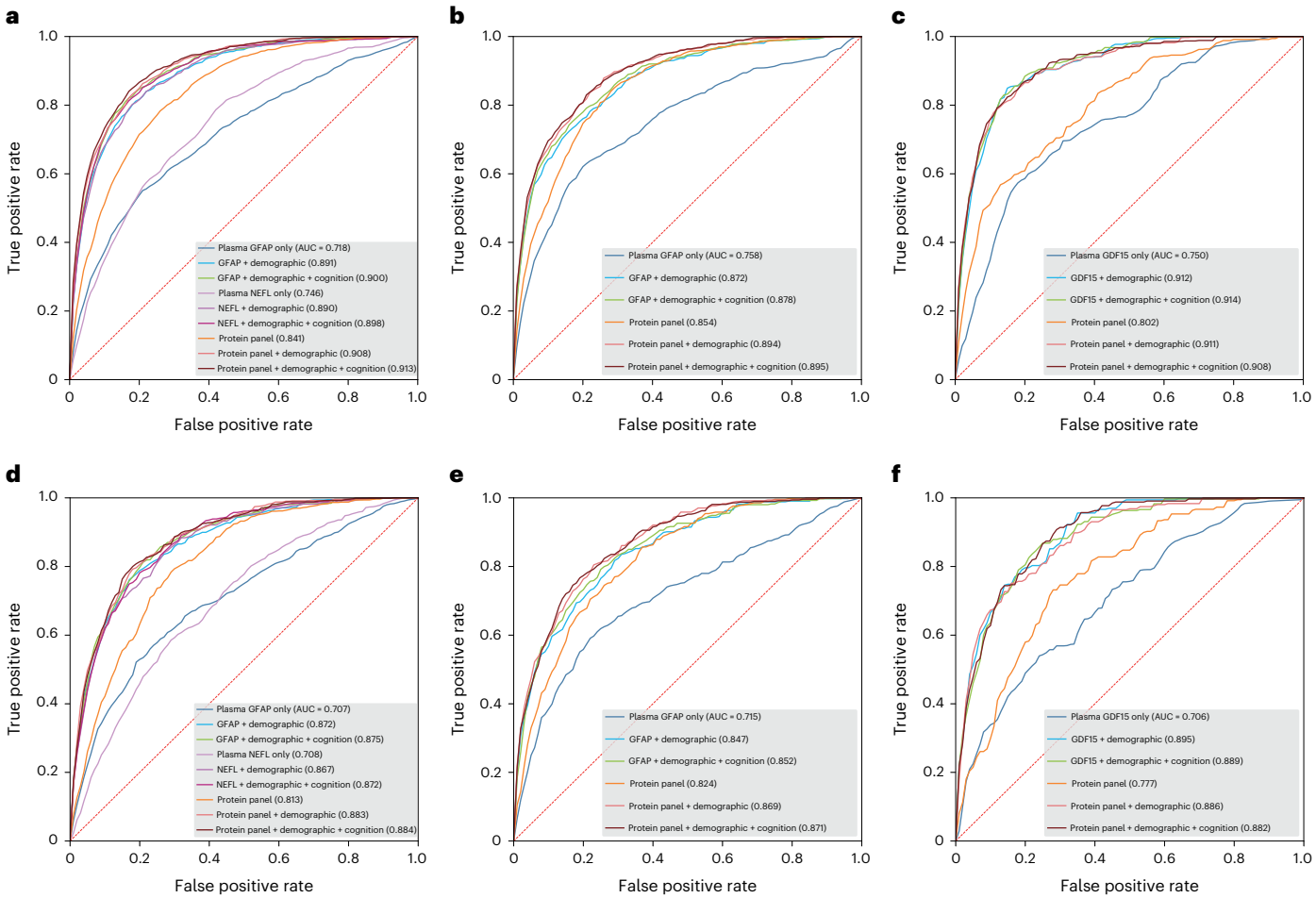


Fig. 3 | Predictive accuracy of plasma proteins, alone or in combination with other variables. a–**f**, Receiver operating curves show the performance of different variable models for predicting all incident ACD (**a**), AD (**b**) and VaD (**c**) as well as the over 10-year incident ACD (**d**), AD (**e**) and VaD (**f**). Within the combined model, demographic indicators included age, sex, education and *APOE* ε4 status, and cognitive tests included pairs matching time and reaction time.

comparable performance to the full model combining the protein panel and clinical information.

Moreover, based on the protein cutoff values obtained from the training set, we reconfirmed the relationships of the important proteins with the risk of clinical progression in the testing set (Supplementary Table 12). In line with our primary results, individuals with higher baseline NEFL and GDF15 levels had a higher risk of developing both dementia and nondementia diseases in the future, whereas the association between baseline GFAP or LTBP2 levels and disease progression was only significant in dementia.

Predementia trajectories for plasma proteins

We finally delineated the temporal trajectories of the plasma proteins starting at dementia diagnosis using a backward 15-year timescale and compared them with protein changes over the same period in those free from dementia. Smoothing splines demonstrated that plasma GFAP, GDF15 and NEFL appeared to deviate from normal values as early as over 10 years before the onset of dementia, either for ACD, AD or VaD (Fig. 5). GFAP (group difference $P = 0.017$) and NEFL ($P = 1.21 \times 10^{-4}$) levels rose more steeply over time for individuals who developed ACD compared with those who did not, whereas there was no obvious difference in the steepness of GDF15 slopes ($P = 0.161$). The slope differences between AD cases and controls were significant when analyzing GFAP ($P = 3.19 \times 10^{-4}$), NEFL ($P = 3.19 \times 10^{-4}$) and GDF15 ($P = 0.044$), but not for the remaining proteins (Supplementary Fig. 6 and Supplementary Table 13). By contrast, for GFAP ($P = 0.855$), NEFL ($P = 0.127$) and GDF15 ($P = 0.127$), the slopes did not significantly deviate between incident VaD individuals and those who remained without dementia.

Discussion

By performing a proteome-wide association study, we identified a wide array of plasma proteins associated with an increased risk of incident ACD, AD and VaD. Of these, GFAP, NEFL, GDF15 and LTBP2 consistently had the most significant associations with future dementia events and showed high importance in prediction tasks in 5-year, 10-year, over 10-year and all-time scenarios. To our knowledge, this study provides the inaugural revelation of the importance ranking of plasma proteins in predicting incident ACD, AD and VaD. These proteins alone demonstrated modest predictive accuracies (AUC = 0.7–0.8). Combining GFAP (or GDF15) with basic demographic data enabled desirable predictions for ACD (AUC = 0.891) and AD (AUC = 0.872) (or VaD (AUC = 0.912)). The same parsimonious models yielded AUC values of 0.872, 0.847 and 0.895 for over 10-year ACD, AD and VaD incidence, respectively, with comparable performance to the full model combining the protein panel and clinical information. Individuals with higher GFAP levels were 2.32 times more likely to develop dementia. Notably, GFAP and LTBP2 were highly specific for dementia prediction, yet not for NEFL and GDF15. Furthermore, changes in GFAP and NEFL began to occur at least 10 years before dementia diagnosis, with concentrations rising

more steeply in individuals with incident ACD or AD than in those who remained dementia-free.

Consistent with previous research, NEFL[20,21] has been reported to be associated with AD and VaD, GFAP[22,23], GDF15 (ref. 24), BCAN[25] and NPTXR[26] with AD, with all associations in the same direction as in our results. The relations of AD with LTBP2 and CST5, as well as the relations of VaD with GFAP, GDF15 and MMP12, were original findings. Several mechanisms could link these proteins to dementia, including reactive astrogliosis, blood–brain barrier and/or glymphatic dysfunction[27], axonal damage[28], inflammation[29,30], synaptic dysfunction and loss[26], amyloid-β clearance[31] and neuron apoptosis[32]. Interestingly, we found the specificity of LTBP2 for dementia prediction, which needs to be studied in depth in the future.

Our results showed the greatest importance of GFAP, NEFL and GDF15 in predicting incident ACD and subtypes in both the long and short terms. Plasma GFAP has been proposed as a promising candidate biomarker for identifying AD[22]. Most previous work on the predictive value of GFAP has primarily focused on the risk conversion from mild cognitive impairment to AD[23,33,34], whereas few studies have predicted the risk transition from normal cognition to AD. Two small longitudinal studies have recently examined this and obtained similar AUC values as ours[35,36]. GFAP is currently a research interest for analyzing AD-specific associations[37]. Former studies have discovered elevated plasma GFAP levels in both AD and nonAD dementias[37–40], demonstrating its poor specificity to AD, which is in agreement with our findings. It is, however, not further studied whether GFAP is specific to dementia. By extending the previous association between higher baseline blood GFAP and increased risk of progression to dementia[9], we proved no relationship between GFAP and the risk of nondementia events to help fill this knowledge gap. The handful of studies assessing the performance of blood GFAP in distinguishing dementia from other diseases also supported our results[41]. In addition, less evidence is available assessing the predictive value of GFAP in incident VaD. Nevertheless, this is plausible because GFAP was also increased in patients with cerebrovascular disease[42,43] and linked to vascular pathologies[44,45], such as white matter hyperintensities and cerebral microbleeds, which we hypothesized would increase the risk of VaD[46,47].

Studies have identified the high predictive value of plasma NEFL for AD, which yielded AUC values slightly lower than ours[35,36,48]. While reinforcing the use of NEFL in predicting AD, our results extended current knowledge by preliminarily noting its predictive value for ACD and VaD. As a marker of axonal injury, NEFL has been reported to be linked to several neurological diseases[49–51], in line with our derived associations between NEFL and the risk of nondementia events. Previous studies reported that plasma GDF15 was associated with cerebrovascular disease burden[24,52] and AD[24], but its relationship with VaD has not been elucidated. Here, we presented preliminary descriptions of its longitudinal relationship with VaD and identified GDF15 as the strongest predictor of incident VaD in the studied plasma proteins. Our findings were supported by a recent proteomics study in middle-aged adults that identified GDF15 as a key marker for predicting dementia many years later[53]. Upregulation of GDF15 in the central nervous system was secondary to vascular brain damage[54] like stroke and cerebral microvascular disease[55], probably exerting anti-inflammatory and neurotrophic effects[29] in response to injuries. As the risk of dementia increased under the influence of vascular brain damage[46,56], it could link high plasma GDF15 with incident AD and VaD. The elevation of plasma GDF15 may also be attributed to risk factors for dementia such as cardiovascular diseases, diabetes[57] and obesity[58].

Other proteins, such as the tau protein MAPT, significantly increased the risk of ACD, AD and VaD when studied alone. However, after taking into account the number of proteins tested ($n = 1{,}463$), the associations lost significance (Bonferroni corrected $P > 0.05$). Previous studies have suggested an inferior prognostic and diagnostic performance of total-tau when compared with NEFL[59–61]. Our study marks the initial extension of the comparison to 1,463 proteins. In addition, the role of total-tau in risk of dementia is also controversial because one study found plasma tau levels did not differ significantly between AD patients and healthy controls[62]. More studies are needed to further elucidate the associations.

Driven by the multifactorial nature of dementia etiology and the heterogeneity of clinical manifestations, plasma protein alone is unlikely to attain the highest predictive accuracy[8]. Accordingly, there is a need to combine plasma proteins with other measures to generate the most accurate prediction of future dementia and establish an optimal predictive algorithm that is noninvasive, cost-effective and easily accessible. We found that the parsimonious model combining GFAP (or GDF15) with basic demographic indicators could achieve a desirable prediction for ACD and AD (or VaD). The results were robust in predicting 5-year, 10-year, over 10-year and all-time scenarios, which is important but has not been explored before. Our proposed model may offer considerable cost benefits compared with using lumbar punctures or imaging scans to screen eligible participants, particularly in primary care.

Generally, adding proteins to demographic features significantly improved the prediction, which reflected the complementary information that proteins hold over demographic measures. This finding may largely translate into the potential clinical utility of proteins as an additional source of discriminatory information to refine future dementia prediction. Moreover, it is worth noting that, both the AUC values of demographic features alone and single key proteins were high (>0.70), which could lead to a mild increase in AUC values when combining them.

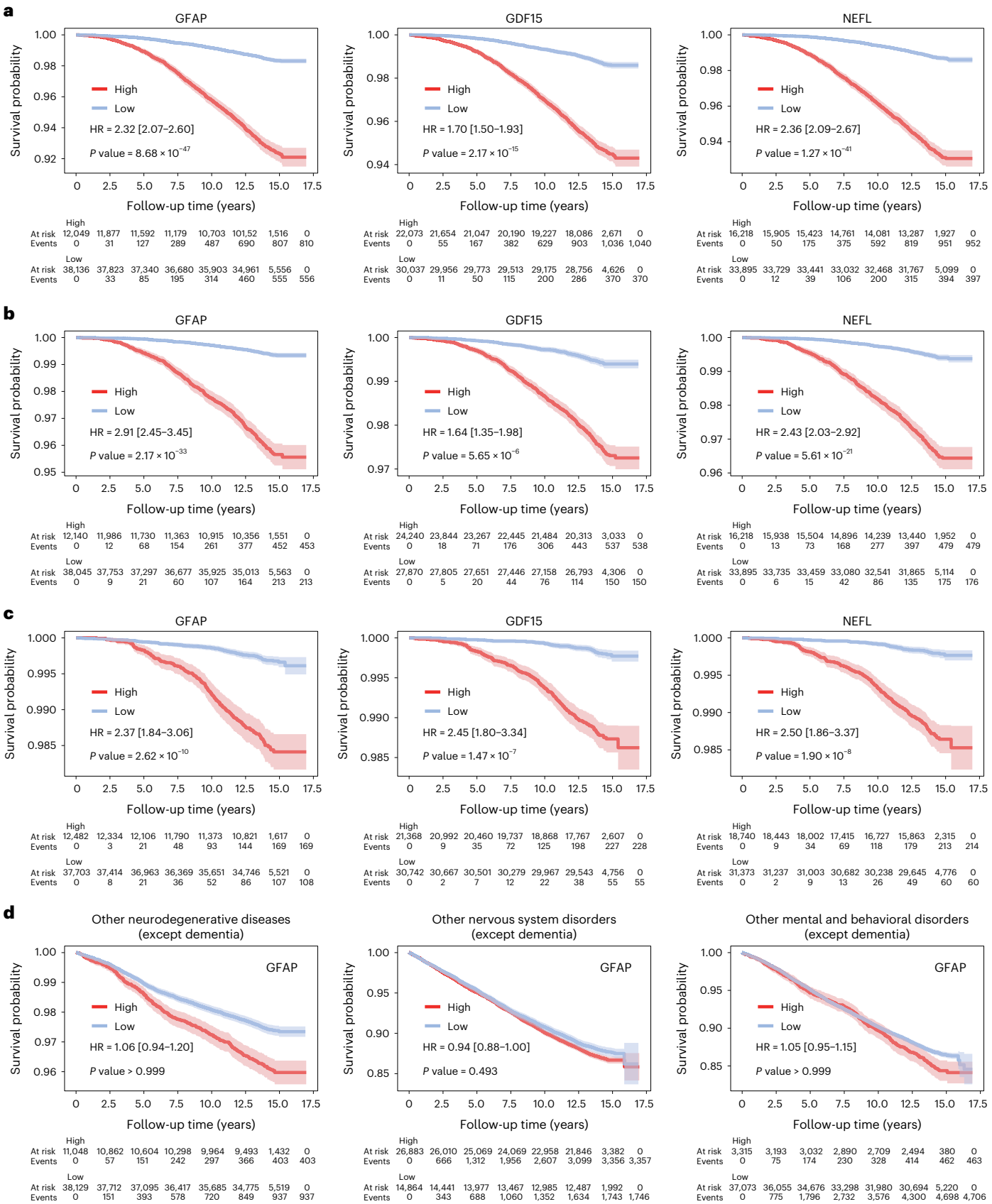
Adding additional cognitive tests to the model has little improvement in prediction accuracy, indicating that substantial parts of the cognitive tests' discriminatory information are shared with protein indicators. Moreover, considering the AUC values of single proteins, protein panel and demographic features alone were high, combining them with cognition is reasonable to attain a mild increase in AUC values. In some cases, adding additional cognitive tests to the model did not yield an increased predictive accuracy and this phenomenon was particularly present in over 10-year models. Because cognitive decline occurs late in the course of dementia, cognitive tests might hold limited predictive information on over 10-year dementia outcomes.

This study has clear implications for the use of plasma GFAP, NEFL and GDF15 as prognostic and/or monitoring biomarkers in the preclinical phase of dementia. The dynamic changes in GFAP, NEFL and GDF15 provide observable evidence of early signs of dementia beginning more than 10 years before the diagnosis. All three proteins were strongly associated with the risk of progression to dementia, and the effect sizes of GFAP and NEFL were the largest. In particular, for AD, the effect sizes of GFAP and NEFL were 1.8 and 1.5 times that of GDF15. These results suggest that GFAP and NEFL are stronger prognostic markers for risk

Fig. 4 | Predictive performance of baseline protein levels on the risk of clinical progression. a–**d**, Unadjusted Kaplan–Meier curves present the clinical progression to ACD (**a**), AD (**b**), VaD (**c**) and nondementia diseases (**d**) over time, visualized for individuals with low (blue line) and high (red line) baseline plasma GFAP, GDF15 or NEFL levels. The cutoffs splitting the high- and low-protein level groups were calculated by the achievement of the largest Youden index. The number of individuals at risk per 2.5-year interval is listed below the curve. Cox proportional hazard models estimate the association between baseline dichotomized protein and disease risk, where HRs and *P* values are calculated after adjusting for age, sex, education and *APOE* ε4 alleles. Shaded regions represent standard errors derived from survival proportions. *P* values were calculated under two-sided tests. Bonferroni corrections were applied to assess significant associations ($P < 0.05$), taking into account the number of proteins tested here ($n = 12$).

https://doi.org/10.1038/s43587-023-00565-0

of incident ACD and AD than GDF15, as partially reported previously[9]. Moreover, we found steeper rises in GFAP, NEFL and GDF15 levels in individuals with incident AD but not VaD than in those who remained dementia-free. This adds evidence to previous studies showing that plasma GFAP and NEFL concentrations rise at a constantly higher rate in people who will develop AD[20,23]. VaD-related research needs to be further explored in the future. Together with our findings that GFAP was specific in predicting dementia onset, GFAP might be more



https://doi.org/10.1038/s43587-023-00565-0

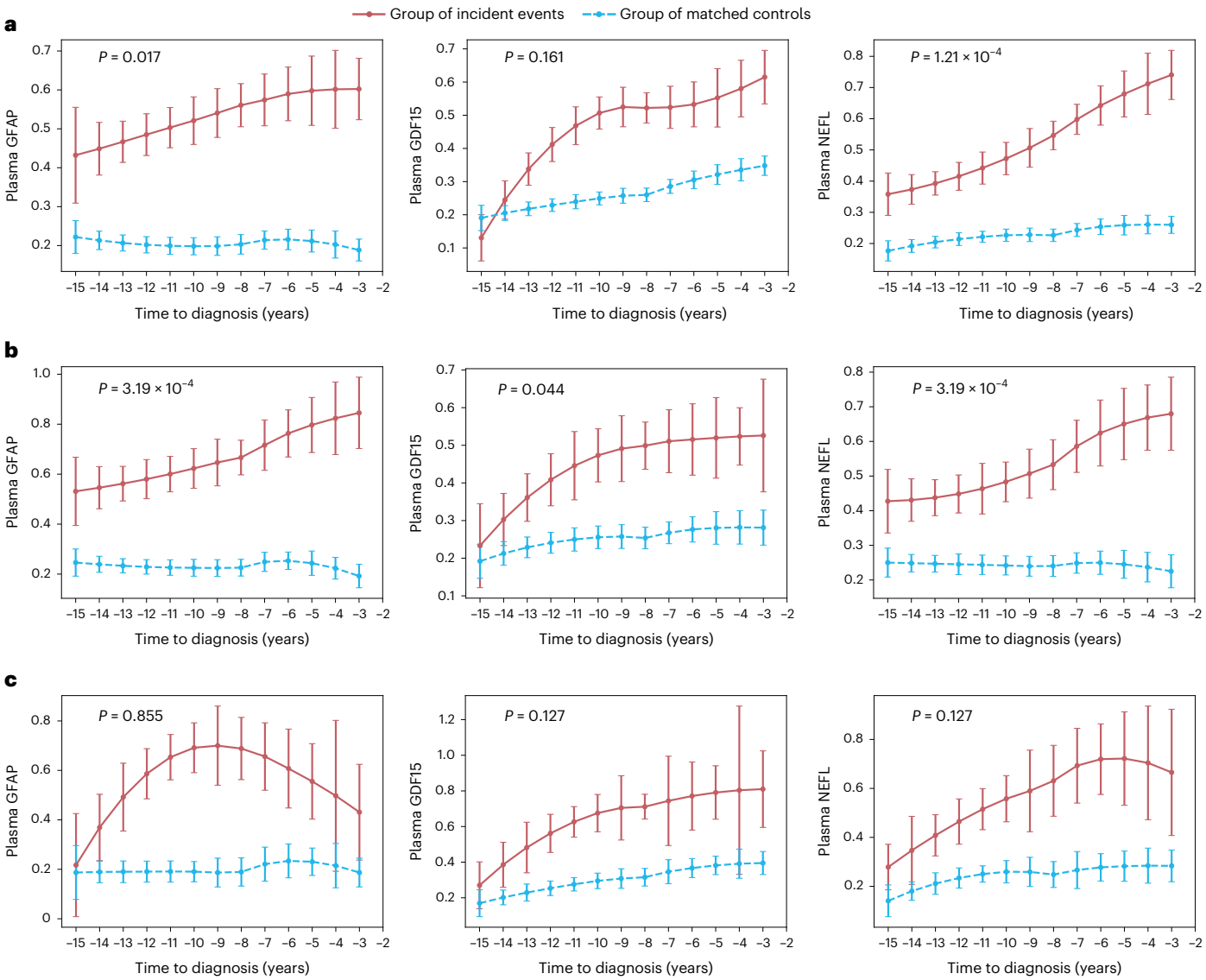


Fig. 5 | Temporal trajectories of plasma proteins preceding the dementia diagnosis. a–**c**, The dynamic changes of plasma GFAP, GDF15 and NEFL before developing ACD (**a**), AD (**b**) and VaD (**c**) were plotted. Nested case–control studies were employed to match an individual with incident events (within 15-year observation period) to five healthy controls under matching criteria of age (±2 years), sex and *APOE* ε4 alleles. ACD versus non-ACD controls: $n = 1{,}399$ versus 6,995; AD versus non-AD controls: $n = 677$ versus 3,385; VaD/non-VaD controls: $n = 281$ versus 1,405. Red curves correspond to patients with dementia, and blue curves correspond to controls. The mean values of protein concentrations by time to the index date were fitted using locally weighted smoothing curves. Error bars represent standard errors. The number of cases in the 1, 2 and 3 years before dementia diagnosis was small and thus aggregated. Mann–Kendall trend tests were performed to examine the slope differences of plasma protein levels over time for individuals with incident dementia compared with those for individuals who did not develop dementia, where *P* values were calculated under two-sided tests and no multiple comparisons were applied.

valuable when measured repeatedly during the preclinical phase of dementia as an indicator of therapeutic effects in future clinical trials. If treatment-induced reductions in GFAP towards normal values are clearly correlated with clinically beneficial effects, then future trials targeting early-stage dementia could incorporate plasma GFAP as a potential surrogate endpoint.

The strengths of our study include the long follow-up and high-throughput proteome analysis of a large community-based sample, which enabled us to replicate previous findings and discover unprecedented plasma biomarkers. Some limitations should be acknowledged when interpreting the results. First, although a comprehensive assessment of circulating proteins is provided by the UKB, not all of the human proteome is captured within this platform, and biases may be present in the priority of measuring secreted proteins. For example, there is a lack of data on plasma amyloid and tau-related proteins, which limits our ability to investigate their predictive values in dementia. Second, the incidence of dementia is lower than that in other reported cohorts given that the UKB participants tend to be younger at enrollment. However, when compared with other studies at certain same age groups (for example, the 65–69 years group)[63], the incidence rates were similar. To identify potential patients with dementia, we have included patients from hospital admissions, death registers and primary care, as suggested previously[64]. Third, validation using independent external datasets would be ideal. However, obtaining proteomics data from large-scale prospective external cohorts with long follow-up that perform Olink assays is currently unavailable. To alleviate this issue, we performed a replication analysis by randomly splitting the UKB into training and testing sets. The replication results were consistent with those obtained using the cross-validation strategy that we performed previously, which further suggests the robustness of

our findings. Fourth, our study identified important protein biomarkers with high accuracies for incident dementia prediction, which would have significant implications for screening of high-risk populations for dementia and early intervention. Regarding the generalization of the protein thresholds, many practical circumstances need to be taken into account, because different detection techniques, antibodies, sample handling procedures and many other factors can affect the measured protein concentrations. Fifth, there may be uncertainty in the diagnosis of UKB. For example, VaD often coexists with AD pathology, and it is difficult to know how different the VaD category is from AD in the absence of additional biomarkers.

Utilizing a data-driven proteomics strategy, we innovatively identified important plasma biomarkers for future dementia prediction from the largest prospective community-based cohort with long-term follow-up to date. Our findings strongly support the associations of plasma GFAP, NEFL, GDF15 and LTBP2 with incident ACD, AD and VaD, and emphasize their importance in dementia prediction. Of note, GFAP and the emerging biomarker LTBP2 could serve as promising predictive biomarkers specific to dementia. Combining GFAP with basic demographic indicators achieved a desirable prediction for dementia, even more than 10 years before the diagnosis. These findings are poised to yield significant implications for screening people at high risk for dementia and for early intervention.

Methods

Study population

The UKB is an ongoing large population-based prospective cohort study with extensive and in-depth proteomic and phenotypic data. Recruitment of over 500,000 individuals aged 39–70 years was successfully achieved during 2006–2010, and their health is being followed over the long term. Enrollees, registered under the UK National Health Service, were sourced from 22 assessment centers across the United Kingdom. For the purpose of analyses, we excluded subjects with dementia at baseline or self-reported dementia and those with missing proteomic data, yielding an analytic cohort of 52,645 participants without dementia (median age 58 years and 53.9% females) (Extended Data Fig. 1).

Blood proteomics

The baseline blood sample collection was completed at 22 local assessment centers across the UK in 2007–2010. As previously described[65,66], the majority of blood samples were randomly collected during the baseline visit from UKB participants, and the remaining were gathered from members of the UKB Pharma Plasma Proteome consortium and individuals participating in the COVID-19 repeat-imaging study at multiple visits. For each participant, blood samples were collected in EDTA tubes and then immediately centrifuged at 2,500*g* for 10 min at 4 °C to isolate plasma. Afterward, the supernatant was divided into aliquots and stored at −80 °C as soon as possible until further processing. Samples were transported on dry ice to the Olink Analysis Service in Sweden and then uniformly quantified using the antibody-based Olink Explore Proximity Extension Assay[67]. Proteomic profiling was done on plasma samples from 54,306 UKB participants spanning April 2021 to February 2022. Experimenting investigators were blinded to all sample characteristic or clinical data. More detailed sample handling and storage procedures were reported in previous publications[68]. Following stringent quality control procedures (biobank.ndph.ox.ac.uk/ukb/ukb/docs/PPP_Phase_1_QC_dataset_companion_doc.pdf), 1,463 unique proteins were measured across four panels containing cardiometabolic, inflammation, neurology and oncology proteins. The inter- and intraplate coefficients of variation for all Olink panels were lower than 20% and 10%, respectively. The protein levels were provided by translating them into Normalized Protein eXpression (NPX) values (https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/Olink_1536_B0_to_B7_Normalization.pdf). To generate NPX values, the counts for each sample and each assay were divided by the counts for the extension control, and the ratio was further log-transformed. Intra- and interplate variations were minimized through accounting for the median of extension control-normalized counts, batch-specific median NPX value, and difference of the assay specific median NPX value of each batch[67] (https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/Olink_1536_B0_to_B7_Normalization.pdf). Because proteomics data were preprocessed to log-transformed NPX values, data distribution (Supplementary Fig. 1) was assumed to be normal but this was not formally tested.

Dementia outcomes

Curated disease phenotypes were defined using reports from hospital admissions, primary care and death registry records (Supplementary Table 1). Primary outcomes include incident events due to ACD, AD and VaD, ascertained from data records of first occurrence reports (fields 131036-37, 130836-43), algorithm definitions (fields 42018-25), death registrations (fields 40001-02) and hospital inpatient data summaries (fields 41270, 41280). The outcome date for dementia diagnosis was established using the earliest recorded date of any aforementioned data sources[69]. Follow-up visits commenced from the date of attendance at the assessment center (field 53) until the earliest recorded date of diagnosis, the date of mortality or the last available date supplied by the hospital or general practitioner, whichever occurred first[64]. The last recorded date was March 2023. The diagnosis data were linked to UK electronic health records among which the dementia cases were reported by professional clinicians in hospitals, family doctors in the primary care system or by staff in the death register system of the UK. According to the International Classification of Diseases (ICD)-9 and ICD-10 codes, dementia cases were diagnosed and classified. Self-reported disease cases were excluded from this paper. All of these ensure the reliability of dementia diagnosis.

Statistics and reproducibility

No statistical methods were used to predetermine sample size, but our sample size is similar to or even higher than those reported in previous publications[18,53]. In the descriptive analysis of variables of interest, between-group (incident ACD/AD/VaD versus control) differences were compared using chi-squared tests for categorical variables and Student's *t*-tests for continuous variables. First, Cox proportional hazard regression models were conducted to estimate the associations of each plasma protein NPX value (scaled) with incident dementia (ACD, AD and VaD). HR values, 95% CIs and *P* values were reported. Model 1 was adjusted for age, sex, education and *APOE* ε4 alleles. Model 2 was additionally adjusted for systolic blood pressure, hypertension treatment, history of diabetes, smoking status, prevalent cardiovascular disease (atrial fibrillation, coronary heart disease, heart failure, stroke or peripheral artery disease), total and high-density lipoprotein (HDL) cholesterol, and body mass index[20,70]. Bonferroni corrections were applied to assess significant associations ($P < 0.05$), taking into account the number of proteins tested ($n = 1{,}463$).

Enrichment analysis was performed on the significant proteins after Bonferroni correction derived from Cox proportional hazard regressions in model 1 or model 2. We employed the Enrichr[71] using the full set of Olink proteins as the reference to glean a deeper biological understanding. Statistical significance was presented as the *P* value of Fisher's exact test, followed by false discovery rate corrections via the Benjamini–Hochberg procedure[72].

Important proteins were then determined in two steps: variable importance ranking and sequential forward selection. The significant proteins after Bonferroni corrections derived from Cox proportional hazard regressions in both model 1 and model 2 were fed into a preliminary trained light gradient boosting machine (LGBM) classifier[73], and each protein was ranked according to its contribution to model performance (as judged by the information gain), which can be considered

as the protein's ability to identify future dementia onset. Afterward, a sequential forward selection approach was employed, adding proteins to the newly developed LGBM classifier one at a time in succession based on their ranking of importance[64]. The selection procedure ceased once the optimal performance of the AUC was attained, artificially defined as when no incremental performance was detected in two consecutive DeLong tests. In this scenario, no significant improvement in model performance could be observed when additional proteins were added. The selected important proteins were then visualized using SHAP plots.

Next, receiver operating characteristic analyses were performed to evaluate the accuracy of the above-selected important proteins in predicting dementia, both alone and in combination with other measures (demographic indicators: age, sex, education and *APOE* ε4 status; cognitive tests: reaction time and pairs matching time[64]). Reaction time and pairs matching time were measured by two different cognitive tests, the snap game (https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=400; https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=20023) and pairs matching game (https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=100030), respectively. Detailed descriptions have been supplied in the previous literature[64]. To assess the combined power of the proteins, we further utilized the protein panel (or ProRS) to replace the single protein. The protein panel referred to direct modeling using the combination of the above-selected important proteins during the protein ranking procedure. The ProRS was a predicted risk score generated from the LGBM model utilizing the above-selected important proteins during the protein ranking process, and it was developed to further exploit the ensembled proteins as a whole to predict future dementia. DeLong tests[74] and bootstrap tests with 2,000 iterations implemented using R package pROC[75] were adopted to compare whether the AUC values differed significantly between models.

To evaluate the generalizability of the selected proteins and the robustness of their predictive accuracy, we repeated the above analyses among these target populations: 5-year/10-year/over 10-year/all incident dementia, 10-year/over 10-year/all incident AD and 10-year/over 10-year/all incident VaD. When performing the corresponding analyses, individuals who developed dementia after the timestamps, either 5 or 10 years, were treated as healthy individuals (for example, for 5-year prediction analysis, people who developed dementia after 5 years were treated not having dementia as beyond 5-year incidences were unknown under the specific observation period). Notably, for the over 10-year analysis, individuals who developed dementia within 10 years were excluded because it assumed the observation period was longer than 10 years, aiming to investigate the proteins' predictive utility for those who developed dementia more than 10 years after baseline assessments. The analyses were repeated following the exclusion of individuals who experienced ACD, AD or VaD within the initial 2 years of follow-up.

The model establishment and evaluations were implemented through internal leave-one-region-out cross-validation. In brief, we split the dataset into ten folds based on the geographical locations (East Midlands, London, North East, North West, Scotland, South East, South West, Wales, West Midlands, and Yorkshire and Humber) of the 22 assessment centers during participants' recruitment. Each time, nine folds of data were utilized as a training set and the rest as a testing set, and we repeated this process ten times by shifting the folds of data as training and testing sets. To further reduce overfitting, we tuned hyperparameters within the training data (ninefold of data) itself under each cross-validation loop by randomly splitting 80% and 20% for model development and model validation. In addition, we conducted feature selection, dramatically reducing the pool of proteins engaged in the prediction task to alleviate potential overfitting concerns. The testing sets of the leave-one-region-out cross-validation were kept untouched and merely used for model evaluations.

Subsequently, we constructed Kaplan–Meier survival curves, delineating high and low baseline top protein levels (the optimal cutoff was determined by maximizing the Youden index), to visualize the clinical progression of dementia events over time[9]. Cox proportional hazard models adjusted for age, sex, education and *APOE* ε4 alleles were undertaken to estimate the differences in the prognostic value of dichotomized protein concentrations. Of note, we explored the relationships of certain plasma proteins with clinical progression in nondementia diseases to assess their specificity in dementia events.

In addition, we randomly split the studied population into a two-thirds and a one-third set. The significant proteins derived from Cox proportional hazard regressions in both model 1 and model 2 were fed into LGBM classifiers. Here we used the training set to ascertain important proteins for dementia prediction and then evaluated the predictive accuracies of these proteins in the testing set. We reported the performance metrics using median AUCs and 95% CIs, which were obtained through bootstrap strategy. When investigating the clinical progression of disease events over time, the protein cutoffs were calculated using the training set and defined by the achievement of the largest Youden index. Cox proportional hazard models were then applied to estimate the associations between baseline proteins and dementia risk within the testing set.

Lastly, temporal trajectories were depicted to observe the dynamic evolution of plasma proteins during the 15 years preceding dementia diagnosis. To achieve this, a nested case–control study design was carried out. Similar to a previous article[76], all incident cases of dementia identified during follow-up were considered as cases, whereas the nested controls were selected by incidence density sampling among cohort members who remained dementia-free at follow-up. Controls were matched to patient cases in a ratio of five controls per patient case by age (±2 years), sex and *APOE* ε4 alleles. The date of observation for nested controls was set as the same for their matched dementia cases. Locally weighted scatterplot smoothing curves were employed to plot the mean concentrations of proteins concerning the time leading up to the index date for both cases and controls[76]. Mann–Kendall trend tests were employed to examine the existence of monotonic trends in plasma levels over time between those who developed dementia and those who did not.

Data analyses and visualizations were implemented with libraries of lifelines (v.0.27.4), LightGBM (v.3.3.2), scikit-learn (v.1.0.2), pyMannKendall (v.1.4.2) and Shap (v.0.40.0) under Python (v.3.9) and R (v.4.0.3). We considered two-tailed $P < 0.05$ to be significant.

Ethics statement

This research adhered to the Declaration of Helsinki. Before participation, all individuals provided written consent, and approval was obtained from the North West Multi-Center Research Ethics Committee (11/NW/0382; https://www.ukbiobank.ac.uk/learn-more-about-ukbiobank/about-us/ethics)[65,66]. Subjects were compensated commensurate with the amount of research procedures accomplished and the duration of involvement. The study received approval from the UKB under application number 19542.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data used in the present study are available from UK Biobank with restrictions applied. Data were used under license and are thus not publicly available. Access to the UK Biobank data can be requested through a standard protocol (https://www.ukbiobank.ac.uk/register-apply/). Data used in this study are available in the UK Biobank under application number 19542. All data supporting the findings described in this manuscript are available in the article and in the supplementary

materials and from the corresponding author upon request. Source data are provided with this paper.

Code availability

All software used in this study is publicly available. The code used in this study can be accessed at https://github.com/jasonHKU0907/DementiaProteomicPrediction.

References

1. Scheltens, P. et al. Alzheimer's disease. *Lancet* **397**, 1577–1590 (2021).
2. Swaddiwudhipong, N. et al. Pre-diagnostic cognitive and functional impairment in multiple sporadic neurodegenerative diseases. *Alzheimers Dement.* **19**, 1752–1763 (2023).
3. Shah, H. et al. Research priorities to reduce the global burden of dementia by 2025. *Lancet Neurol.* **15**, 1285–1294 (2016).
4. Teunissen, C. E. et al. Blood-based biomarkers for Alzheimer's disease: towards clinical implementation. *Lancet Neurol.* **21**, 66–77 (2022).
5. Zetterberg, H. Biofluid-based biomarkers for Alzheimer's disease-related pathologies: an update and synthesis of the literature. *Alzheimers Dement.* **18**, 1687–1693 (2022).
6. Hansson, O. et al. The Alzheimer's Association appropriate use recommendations for blood biomarkers in Alzheimer's disease. *Alzheimers Dement.* **18**, 2669–2686 (2022).
7. Nakamura, A. et al. High performance plasma amyloid-beta biomarkers for Alzheimer's disease. *Nature* **554**, 249–254 (2018).
8. Palmqvist, S. et al. Prediction of future Alzheimer's disease dementia using plasma phospho-tau combined with other accessible measures. *Nat. Med.* **27**, 1034–1042 (2021).
9. Verberk, I. M. W. et al. Serum markers glial fibrillary acidic protein and neurofilament light for prognosis and monitoring in cognitively normal older people: a prospective memory clinic-based cohort study. *Lancet Healthy Longev.* **2**, e87–e95 (2021).
10. Kim, K. et al. Clinically accurate diagnosis of Alzheimer's disease via multiplexed sensing of core biomarkers in human plasma. *Nat. Commun.* **11**, 119 (2020).
11. Ashton, N. J. et al. Differential roles of Aβ42/40, p-tau231 and p-tau217 for Alzheimer's trial selection and disease monitoring. *Nat. Med.* **28**, 2555–2562 (2022).
12. Karikari, T. K. et al. Blood phosphorylated tau 181 as a biomarker for Alzheimer's disease: a diagnostic performance and prediction modelling study using data from four prospective cohorts. *Lancet Neurol.* **19**, 422–433 (2020).
13. Mattsson-Carlgren, N. et al. Prediction of longitudinal cognitive decline in preclinical Alzheimer disease using plasma biomarkers. *JAMA Neurol.* **80**, 360–369 (2023).
14. Karikari, T. K. et al. Blood phospho-tau in Alzheimer disease: analysis, interpretation, and clinical utility. *Nat. Rev. Neurol.* **18**, 400–418 (2022).
15. Tanaka, T. et al. Plasma proteomic signatures predict dementia and cognitive impairment. *Alzheimers Dement. (N Y)* **6**, e12018 (2020).
16. Hye, A. et al. Proteome-based plasma biomarkers for Alzheimer's disease. *Brain* **129**, 3042–3050 (2006).
17. Bai, B. et al. Proteomic landscape of Alzheimer's disease: novel insights into pathogenesis and biomarker discovery. *Mol. Neurodegener.* **16**, 55 (2021).
18. Walker, K. A. et al. Large-scale plasma proteomic analysis identifies proteins and pathways associated with dementia risk. *Nat. Aging* **1**, 473–489 (2021).
19. Dubois, B. et al. Clinical diagnosis of Alzheimer's disease: recommendations of the International Working Group. *Lancet Neurol.* **20**, 484–496 (2021).
20. de Wolf, F. et al. Plasma tau, neurofilament light chain and amyloid-β levels and risk of dementia; a population-based cohort study. *Brain* **143**, 1220–1232 (2020).
21. Ma, W. et al. Elevated levels of serum neurofilament light chain associated with cognitive impairment in vascular dementia. *Dis. Markers* **2020**, 6612871 (2020).
22. Benedet, A. L. et al. Differences between plasma and cerebrospinal fluid glial fibrillary acidic protein levels across the Alzheimer disease continuum. *JAMA Neurol.* **78**, 1471–1483 (2021).
23. Cicognola, C. et al. Plasma glial fibrillary acidic protein detects Alzheimer pathology and predicts future conversion to Alzheimer dementia in patients with mild cognitive impairment. *Alzheimers Res. Ther.* **13**, 68 (2021).
24. McGrath, E. R. et al. Growth differentiation factor 15 and NT-proBNP as blood-based markers of vascular brain injury and dementia. *J. Am. Heart Assoc.* **9**, e014659 (2020).
25. Whelan, C. D. et al. Multiplex proteomics identifies novel CSF and plasma biomarkers of early Alzheimer's disease. *Acta Neuropathol. Commun.* **7**, 169 (2019).
26. Dulewicz, M., Kulczyńska-Przybik, A., Słowik, A., Borawska, R. & Mroczko, B. Neurogranin and neuronal pentraxin receptor as synaptic dysfunction biomarkers in Alzheimer's disease. *J. Clin. Med.* **10**, 4575 (2021).
27. O'Connor, A. et al. Plasma GFAP in presymptomatic and symptomatic familial Alzheimer's disease: a longitudinal cohort study. *J. Neurol. Neurosurg. Psychiatry* **94**, 90–92 (2023).
28. Kuhle, J. et al. Serum neurofilament light chain is a biomarker of human spinal cord injury severity and outcome. *J. Neurol. Neurosurg. Psychiatry* **86**, 273–279 (2015).
29. Fuchs, T. et al. Macrophage inhibitory cytokine-1 is associated with cognitive impairment and predicts cognitive decline – the Sydney Memory and Aging Study. *Aging Cell* **12**, 882–889 (2013).
30. Babu, H. et al. Systemic inflammation and the increased risk of inflamm-aging and age-associated diseases in people living with HIV on long term suppressive antiretroviral therapy. *Front. Immunol.* **10**, 1965 (2019).
31. Castellano, J. M. et al. Low-density lipoprotein receptor overexpression enhances the rate of brain-to-blood Aβ clearance in a mouse model of β-amyloidosis. *Proc. Natl Acad. Sci. USA* **109**, 15502–15507 (2012).
32. Li, W. et al. Dl-3-*n*-butylphthalide reduces cognitive impairment induced by chronic cerebral hypoperfusion through GDNF/GFRα1/Ret signaling preventing hippocampal neuron apoptosis. *Front. Cell Neurosci.* **13**, 351 (2019).
33. Oeckl, P. et al. Serum GFAP differentiates Alzheimer's disease from frontotemporal dementia and predicts MCI-to-dementia conversion. *J. Neurol. Neurosurg. Psychiatry* **93**, 659–667 (2022).
34. Kivisäkk, P. et al. Plasma biomarkers for diagnosis of Alzheimer's disease and prediction of cognitive decline in individuals with mild cognitive impairment. *Front. Neurol.* **14**, 1069411 (2023).
35. Beyer, L. et al. Amyloid-beta misfolding and GFAP predict risk of clinical Alzheimer's disease diagnosis within 17 years. *Alzheimers Dement.* **19**, 1020–1028 (2023).
36. Stocker, H. et al. Association of plasma biomarkers, p-tau181, glial fibrillary acidic protein, and neurofilament light, with intermediate and long-term clinical Alzheimer's disease risk: results from a prospective cohort followed over 17 years. *Alzheimers Dement.* **19**, 25–35 (2023).
37. Oeckl, P. et al. Glial fibrillary acidic protein in serum is increased in Alzheimer's disease and correlates with cognitive impairment. *J. Alzheimers Dis.* **67**, 481–488 (2019).
38. Heller, C. et al. Plasma glial fibrillary acidic protein is raised in progranulin-associated frontotemporal dementia. *J. Neurol. Neurosurg. Psychiatry* **91**, 263–270 (2020).

39. Verberk, I. M. W. et al. Combination of plasma amyloid beta$_{(1-42/1-40)}$ and glial fibrillary acidic protein strongly associates with cerebral amyloid pathology. *Alzheimers Res. Ther.* **12**, 118 (2020).
40. Rajan, K. B. et al. Remote blood biomarkers of longitudinal cognitive outcomes in a population study. *Ann. Neurol.* **88**, 1065–1076 (2020).
41. Katisko, K. et al. GFAP as a biomarker in frontotemporal dementia and primary psychiatric disorders: diagnostic and prognostic performance. *J. Neurol. Neurosurg. Psychiatry* **92**, 1305–1312 (2021).
42. Katsanos, A. H. et al. Plasma glial fibrillary acidic protein in the differential diagnosis of intracerebral hemorrhage. *Stroke* **48**, 2586–2588 (2017).
43. Undén, J. et al. Explorative investigation of biomarkers of brain damage and coagulation system activation in clinical stroke differentiation. *J. Neurol.* **256**, 72–77 (2009).
44. Elahi, F. M. et al. Plasma biomarkers of astrocytic and neuronal dysfunction in early- and late-onset Alzheimer's disease. *Alzheimers Dement.* **16**, 681–695 (2020).
45. Shir, D. et al. Association of plasma glial fibrillary acidic protein (GFAP) with neuroimaging of Alzheimer's disease and vascular pathology. *Alzheimers Dement. (Amst.)* **14**, e12291 (2022).
46. Vermeer, S. E. et al. Silent brain infarcts and the risk of dementia and cognitive decline. *N. Engl. J. Med.* **348**, 1215–1222 (2003).
47. Prins, N. D. & Scheltens, P. White matter hyperintensities, cognitive impairment and dementia: an update. *Nat. Rev. Neurol.* **11**, 157–165 (2015).
48. Cullen, N. C. et al. Plasma biomarkers of Alzheimer's disease improve prediction of cognitive decline in cognitively unimpaired elderly populations. *Nat. Commun.* **12**, 3555 (2021).
49. Fyfe, I. Neurofilament light chain – new potential for prediction and prognosis. *Nat. Rev. Neurol.* **15**, 557 (2019).
50. Pilotto, A. et al. Plasma neurofilament light chain predicts cognitive progression in prodromal and clinical dementia with Lewy bodies. *J. Alzheimers Dis.* **82**, 913–919 (2021).
51. Gisslén, M. et al. Plasma concentration of the neurofilament light protein (NFL) is a biomarker of CNS injury in HIV infection: a cross-sectional study. *EBioMedicine* **3**, 135–140 (2016).
52. Chai, Y. L. et al. Growth differentiation factor-15 and white matter hyperintensities in cognitive impairment and dementia. *Medicine (Baltimore)* **95**, e4566 (2016).
53. Walker, K. A. et al. Proteomics analysis of plasma from middle-aged adults identifies protein markers of dementia risk in later life. *Sci. Transl. Med.* **15**, eadf5681 (2023).
54. Schindowski, K. et al. Regulation of GDF-15, a distant TGF-β superfamily member, in a mouse model of cerebral ischemia. *Cell Tissue Res.* **343**, 399–409 (2011).
55. Andersson, C. et al. Associations of circulating growth differentiation factor-15 and ST2 concentrations with subclinical vascular brain injury and incident stroke. *Stroke* **46**, 2568–2575 (2015).
56. van Dijk, E. J. et al. Progression of cerebral small vessel disease in relation to risk factors and cognitive consequences: Rotterdam Scan study. *Stroke* **39**, 2712–2719 (2008).
57. Conte, M. et al. GDF15, an emerging key player in human aging. *Ageing Res. Rev.* **75**, 101569 (2022).
58. Wang, D. et al. GDF15: emerging biology and therapeutic applications for obesity and cardiometabolic disease. *Nat. Rev. Endocrinol.* **17**, 592–607 (2021).
59. Marks, J. D. et al. Comparison of plasma neurofilament light and total tau as neurodegeneration markers: associations with cognitive and neuroimaging outcomes. *Alzheimers Res. Ther.* **13**, 199 (2021).
60. Cousins, K. A. Q. et al. ATN incorporating cerebrospinal fluid neurofilament light chain detects frontotemporal lobar degeneration. *Alzheimers Dement.* **17**, 822–830 (2021).
61. Illán-Gala, I. et al. Plasma tau and neurofilament light in frontotemporal lobar degeneration and Alzheimer disease. *Neurology* **96**, e671–e683 (2021).
62. Jiang, Y. et al. Large-scale plasma proteomic profiling identifies a high-performance biomarker panel for Alzheimer's disease screening and staging. *Alzheimers Dement.* **18**, 88–102 (2022).
63. Prince, M. et al. World Alzheimer Report 2015—The Global Impact of Dementia: An Analysis of Prevalence, Incidence, Cost and Trends. *Alzheimer's Disease International* www.alzint.org/resource/world-alzheimer-report-2015/ (2015).
64. You, J. et al. Development of a novel dementia risk prediction model in the general population: a large, longitudinal, population-based machine-learning study. *EClinicalMedicine* **53**, 101665 (2022).
65. Sun, B. B. et al. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
66. Dhindsa, R. S. et al. Rare variant associations with plasma protein levels in the UK Biobank. *Nature* **622**, 339–347 (2023).
67. Wik, L. et al. Proximity extension assay in combination with next-generation sequencing for high-throughput proteome-wide analysis. *Mol. Cell Proteomics* **20**, 100168 (2021).
68. Elliott, P., Peakman, T. C. & UK Biobank The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int. J. Epidemiol.* **37**, 234–244 (2008).
69. You, J. et al. Plasma proteomic profiles predict individual future health risk. *Nat. Commun.* **14**, 7817 (2023).
70. Tynkkynen, J. et al. Association of branched-chain amino acids and other circulating metabolites with risk of incident dementia and Alzheimer's disease: a prospective study in eight cohorts. *Alzheimers Dement.* **14**, 723–733 (2018).
71. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
72. Korthauer, K. et al. A practical guide to methods controlling false discoveries in computational biology. *Genome Biol.* **20**, 118 (2019).
73. Guolin K. et al. LightGBM: a highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (eds von Luxburg, U., Guyon, I., Bengio, S., Wallach, H. & Fergus, R.) 3149–3157 (Curran Associates Inc., 2017).
74. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
75. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
76. Fang, F. et al. Lipids, apolipoproteins, and the risk of Parkinson disease. *Circ. Res.* **125**, 643–652 (2019).

Acknowledgements

We thank all the participants and researchers from the UK Biobank. W.C. was funded by National Key Research and Development Program of China (grant no. 2023YFC3605400). J.T.-Y. was funded by grants from the Science and Technology Innovation 2030 Major Projects (grant no. 2022ZD0211600), National Natural Science Foundation of China (grant nos. 82071201, 92249305), Research Start-up Fund of Huashan Hospital (grant no. 2022QD002) and Excellence 2025 Talent Cultivation Program at Fudan University (grant no. 3030277001). J.F.-F. was funded by National Key R&D Program of China (grant nos. 2018YFC1312904, 2019YFA0709502), Shanghai Municipal Science and Technology Major Project (grant no. 2018SHZDZX01) and the

111 Project (no. B18015). J.Y. was funded by Shanghai Pujiang Talent Program (grant no. 23PJD006). The funders had no role in study design, data collection and analysis; decision to publish or preparation of the manuscript. Further, we would like to thank the support from the ZHANGJIANG LAB, Tianqiao and Chrissy Chen Institute, and the State Key Laboratory of Neurobiology and Frontiers Center for Brain Science of Ministry of Education, Fudan University.

Author contributions

J.-T.Y. undertook conceptualization and design of the study, interpretation of the data and revision of the manuscript. Y.G., J.Y. and Y.Z. collected, analyzed and interpreted the data, and drafted and revised the manuscript. J.-F.F., W.C. and J.-T.Y. were responsible for funding, administrative, technical or material support. All authors carried out revision of the manuscript. All authors had full access to all the study data and accepted responsibility for submitting it for publication.

Competing interests

The authors declare no competing interests.

Additional information

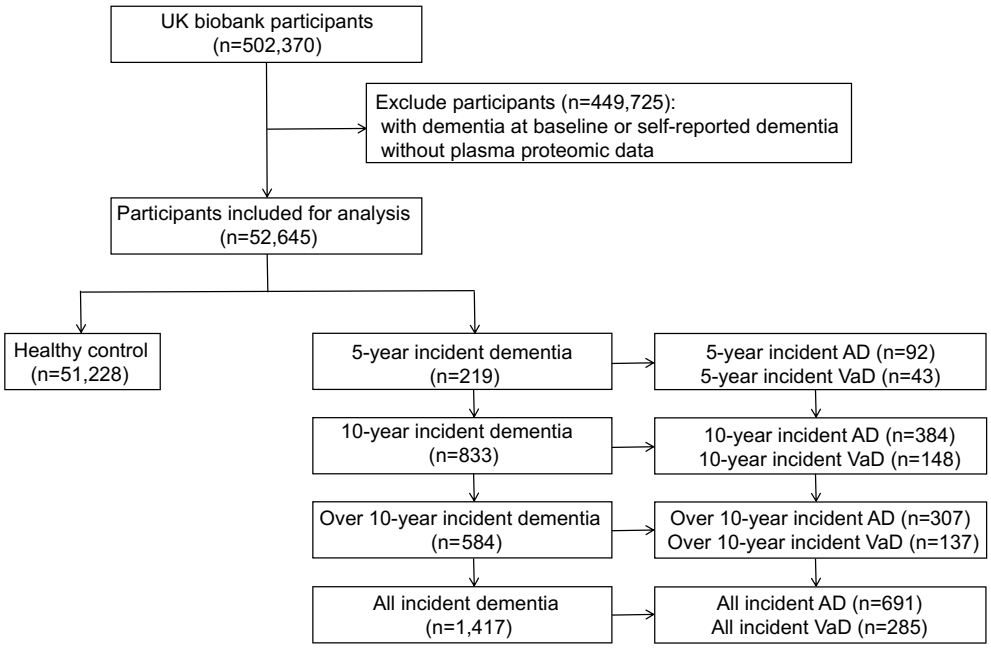
Extended data is available for this paper at https://doi.org/10.1038/s43587-023-00565-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43587-023-00565-0.

Correspondence and requests for materials should be addressed to Jian-Feng Feng, Wei Cheng or Jin-Tai Yu.

Peer review information *Nature Aging* thanks Keenan Walker and Alexa Pichet Binette for their contribution to the peer review of this work.

https://doi.org/10.1038/s43587-023-00565-0



Extended Data Fig. 1 | Flowchart for participants' enrollment. From the UK Biobank cohort, we excluded individuals with dementia at baseline or with self-reported dementia and those who did not undergo plasma proteomic assay. The remaining participants were classified based on their first reported years of ACD or AD or VaD after baseline. Abbreviations: ACD, all-cause dementia; AD, Alzheimer's disease; VaD, vascular dementia.

nature portfolio

Corresponding author(s): Jin-Tai Yu, Wei Cheng, Jian-Feng Feng

Last updated by author(s): Dec 19, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Python v3.9 and R v4.0.3
Data analysis	Most data analyses were performed using Python v3.9. The code used in this study can be assessed at https://github.com/jasonHKU0907/DementiaProteomicPrediction. R version 4.0.3 packages: pROC (v1.18.4), ggplot2 (v3.4.3), ggsurvfit (v0.3.0), survival (3.5-5).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data used in the present study are available from UKB with restrictions applied. Data were used under license and are thus not publicly available. Access to the UKB data can be requested through a standard protocol (https://www.ukbiobank.ac.uk/register-apply/). Data used in this study are available in the UK Biobank

under application number 19542. All data supporting the findings described in this manuscript are available in the article and in the supplementary materials and from the corresponding author upon request. Source Data underlying Fig. 1 to Fig. 5 are available with the paper.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We took sex into consideration in our study and our analyses were adjusted for sex. Sex (FIELD ID 31) in the UK Biobank was determined based on self-reporting data via questionnaires, and all included participants gave written informed consent for sharing of individual-level data.
Reporting on race, ethnicity, or other socially relevant groupings	Ethic background (FIELD ID 21000) information of the UK Biobank was collected based on self-reported data via questionnaires, and most participants were white and of European origin.
Population characteristics	This study included 52,645 adults without dementia at baseline, with a median age of 58 [50-64] years, of whom 28,393 (53.9%) were female and 49,353 (93.7%) were of white ancestry. During a median follow-up of 14.1 [13.4-14.8] years, 1,417 (2.7%) incident dementia cases were identified. The baseline characteristics of the participants are summarized in Table 1. We described continuous variables as median [IQR] and categorical variables as number (percentage) .
Recruitment	The UK Biobank enrolled the participants aged 40-69 years between 2006 and 2010 for baseline assessments in 22 centers across the UK. The assessment visits comprised interviews and questionnaires covering lifestyle and health conditions, physical measures, biological samples, imaging, and genotyping. The database is linked to national health databases, including primary care, hospital inpatient, death, and cancer registration data.
Ethics oversight	This research adhered to the Declaration of Helsinki. Prior to participation, all individuals provided written consent, and approval was obtained from the North West Multi-Center Research Ethics Committee (11/NW/0382; https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics). Subjects were compensated commensurate with the amount of research procedures accomplished and the duration of involvement. The study received approval from the UKB under application number 19542.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample sizes, but our sample size is similar to or even higher than those reported in previous publications. 52,645 eligible participants were finally included.
Data exclusions	We excluded participants with dementia at baseline or self-reported dementia and those with missing proteomic data.
Replication	All available data were used to maximize statistical power of the analysis therefore we did not repeat the analysis.
Randomization	Covariates including baseline age, sex, education, and APOE ε4 alleles were adjusted in our study. As a sensitivity analysis, vascular-related factors were additionally adjusted.
Blinding	Blinding is not applicable to this study as this study is observational.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.