Blood protein levels predict leading incident diseases and mortality in UK Biobank

Danni A. Gadd[1,2], Robert F. Hillary[1,2], Zhana Kuncheva[1,3], Tasos Mangelis[1,3], Romi Admanit[4,5], Jake Gagnon[4,5], Tinchi Lin[4], Kyle Ferber[4], Heiko Runz[4], Biogen Biobank Team, Riccardo E. Marioni*[1,2], Christopher N. Foley*[1,3], Benjamin B. Sun[4,6]*

[1] Optima Partners, Edinburgh, EH2 4HQ, UK.

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

[3] Bayes Centre, The University of Edinburgh, Edinburgh, EH8 9BT, UK.

[4] Translational Sciences, Research and Development, Biogen Inc. Cambridge, MA, USA.

[5] Global Analytics and Data Science, Research and Development , Biogen Inc., Cambridge, MA, USA

[6] Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, UK.

* Equal contributions.

Correspondence: benjamin.sun@biogen.com and riccardo.marioni@ed.ac.uk.

Abstract

The circulating proteome offers insights into the biological pathways that underlie disease. Here, we test relationships between 1,468 Olink protein levels and the incidence of 23 age-related diseases and mortality, ascertained over 16 years of electronic health linkage in the UK Biobank (N=49,234). We report 3,123 associations between 1,052 protein levels and incident diseases ($P_{Bonferroni} < 5.4x10^{-6}$). Forty-four proteins are indicators of eight or more morbidities. Next, protein-based scores (ProteinScores) are developed using penalised Cox regression. When applied to test sets, eight ProteinScores improve Area Under the Curve (AUC) estimates for the 10-year onset of incident outcomes ($P_{Bonferroni} < 0.0025$) beyond age, sex and additional health and lifestyle covariates. The type 2 diabetes ProteinScore outperforms HbA1c ($P = 5.7x10^{-12}$) – a clinical marker used to monitor and diagnose type 2 diabetes. A maximal type 2 diabetes model including the ProteinScore, HbA1c and a polygenic risk score has AUC = 0.90 and Precision-Recall AUC = 0.76. These data characterise early proteomic contributions to major age-related disease.

Introduction

Omics signatures are increasingly used to hone clinical trial design [1], while also opening up avenues for more personalised healthcare [2,3]. Of all the omics layers that can be measured from a single blood test, proteomics arguably holds the most intrinsic predictive potential, given that proteins are the intermediary effectors of health maintenance and disease and are often the targets of pharmacological interventions. Several studies have shown that circulating proteins can discriminate disease cases from controls and delineate risk of incident diagnoses [4–11]. Screening the proteome against incident outcomes has been shown to identify sets of individual protein markers – some of which have then been causally-implicated in disease [8,12–14]. This demonstrates the value protein data have in informing therapeutic targeting and reflecting the internal processes occurring in the body that precede formal diagnoses.

While singular protein markers offer insight into the mediators of disease, harnessing multiple proteins simultaneously can be expected to generate predictive tools with even greater clinical utility [15]. Although cross-sectional case-control studies can inform on the molecular signatures of diagnosed diseases, longitudinal approaches that assess early biomarker signatures relating to time-to-disease are more suited to risk stratification. Scores developed through statistical learning stratify where individuals lie on the disease-risk continuum for a population. While proteomic and metabolomics scores have been developed for certain time-to-event outcomes in isolation [9,16–20], these predictors are rarely developed and tested at scale. Proteomic predictors have been trained using the SomaScan platform for diabetes and cardiovascular event risk and multiple lifestyle and health indicators [21]. Metabolomics data have been recently shown to facilitate incident disease prediction in the UK Biobank [22]. However, no study has systematically assessed proteomic score generation for multiple incident morbidities.

Here, we quantify how large-scale proteomic sampling can identify candidate protein targets and facilitate the prediction of incident outcomes in the UK Biobank **(Fig. 1)**. We use 1,468 Olink plasma protein measurements in 49,234 individuals available as part of the UK Biobank Pharma Proteomics Project (UKB-PPP) [23]. First, Cox proportional hazards (PH) models are used to characterise associations between each protein and 23 incident diseases, ascertained via data linkage to primary and secondary care records and mortality over 16 years of follow-up. Next, the dataset is split into independent training and testing subsets to assess the utility of proteomic scores (ProteinScores) for modelling either 5-year or 10-year onset of the 20 incident outcomes that had a minimum of 150 cases available.

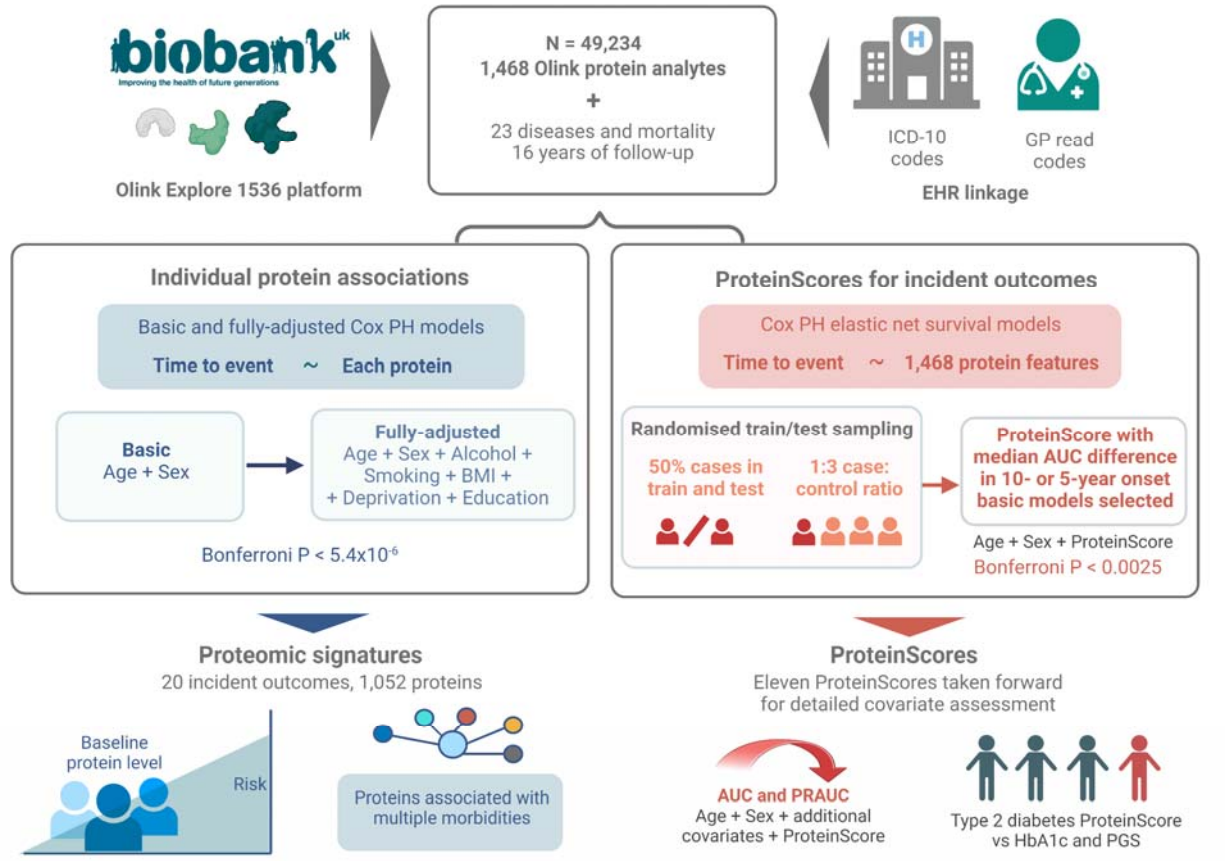

Figure 1. Proteomic assessment of 23 incident diseases and mortality in the UK Biobank (N=49,234). First, individual Cox proportional hazards (PH) models were used to profile relationships between baseline protein analytes and incident diseases or death. Associations that had $P_{Bonferroni} < 5.4x10^{-6}$ in both basic and fully-adjusted models were retained. Proteins associated with multiple morbidities were identified. Next, proteomic predictors (ProteinScores) were trained using Cox PH elastic net regression for 20 of the time-to-event outcomes that had a minimum of 150 cases. Fifty ProteinScore iterations with randomised train and test sample allocations and case:control ratio of 1:3 were run for each trait. ProteinScores that yielded the median incremental difference to the AUC of a basic model for 5-year or 10-year incidence (depending on suitability of the time-to-onset distribution for traits) in the test set were selected. The eleven ProteinScores that significantly improved AUC ($P_{Bonferroni} < 0.0025$) in basic models were taken forward for analyses with a more detailed set of covariates. HbA1c (a clinically used biomarker) and a polygenic risk score (PGS) were further examined for the type 2 diabetes ProteinScore.

Results

The UKB-PPP sample

Of the 1,472 protein levels available in the UKB-PPP sample, 1,463 are unique, due to CXCL8, IL6 and TNF having multiple analyte measurements (annotation information provided in **Supplementary Table 1**). After quality control and removal of outliers, measurements for 54,189 individuals were available. In this study, a total sample of 49,234 individuals with 1,468 protein analytes was used, after exclusions for related individuals and missing data **(Supplementary Fig. 1, Methods)**. The 1,468 analyte measurements correspond to 1,459 unique protein levels. Demographic and phenotypic information for these individuals are presented in **Supplementary Table 2**. Principal components analyses indicated that the first 385 components explained a cumulative variance of 80% in the protein levels **(Supplementary Table 3)**.

Distinct and overlapping protein associations with incident outcomes

First, differential plasma protein levels that were associated with the onset of 23 diseases enriched for leading causes of disability, morbidity and reductions in healthy life expectancy [24–26] were identified, up to 16 years prior to formal diagnoses. Time-to-mortality was also considered as an outcome (4,580 individuals had died during the 16-year follow-up period). A total of 35,232 associations were tested (1,468 analytes and 24 outcomes). The number of cases, controls and mean time-to-onset from baseline for each outcome in basic- and fully-adjusted Cox PH models are summarised in **Table 1**.

In basic (age- or age- and sex-adjusted) models, there were 4,916 associations between 1,203 unique protein analytes and 21 outcomes that had Bonferroni-adjusted $P < 5.4 \times 10^{-6}$ **(Supplementary Table 4).** In fully-adjusted models that further adjusted for health and lifestyle risk factors (body mass index (BMI), alcohol consumption, social deprivation, education status and smoking status), 3,123 of

the basic model associations had $P_{Bonferroni} < 5.4x10^{-6}$ **(Fig. 2a, Supplementary Table 5)**. The 3,123 associations involved 1,052 unique protein analytes and 20 outcomes, ranging from one association for Alzheimer's dementia, gynaecological cancer and multiple sclerosis, to 694 for mortality. No associations were found for brain/CNS cancer, chronic cystitis, major depression and schizophrenia. **Supplementary Table 6** summarises the 1,052 unique protein analytes selected across the 3,123 fully-adjusted associations by disease and by direction of effect (i.e. Hazard Ratio (HR) < 1 or HR > 1). The most statistically significant association was observed between higher GDF15 levels and the incidence of death (HR = 1.97 per SD of the rank transformed protein, 95% CI = [1.89, 2.04], P = 2.6 x 10^{-260}). Multimorbidity profiling identified 44 proteins that had associations with eight or more incident morbidities **(Fig. 2b);** in all instances, elevated levels of the proteins were associated with the increased incidence of disease or death. Of the 44 proteins, GDF15, PLAUR, ST6GAL1 and IL6 had the largest number of associations (10 incident outcomes).

Of the 3,123 associations retained from the fully-adjusted models, 1,964 failed to satisfy the local (i.e., protein) Cox PH assumption over the 16-year follow-up period (Schoenfeld residual test P < 0.05). Due to the presence of sparsity in the final six years of linkage records resulting from censoring, sensitivity analyses that computed hazard ratios over successive yearly follow-up intervals for cases were performed. In these analyses, 1,395 and 415 of the 3,123 associations failed the local PH assumption when restricted to 10-year and 5-year follow-up intervals, respectively **(Supplementary Table 7)**. Relatively minor deviations to the magnitude and statistical significance of the effect size were observed across follow-up intervals. Summary statistics for each of the 3,123 associations split by year of follow-up for cases are available in **Supplementary Table 8**. The results from these sensitivity analyses can be visualised for every association tested in a Shiny app at: https://protein-disease-ukb.optima-health.technology [Username: ukb_disease, Password: shinyappUKB]. The app also includes an interactive network for the 3,123 fully-adjusted

associations that can be manipulated to view multiple proteins and examine their associations with multiple incident morbidities.

Incident event	Cases basic (N)	Controls basic (N)	Years to onset basic (mean, sd)	Cases full (N)	Controls full (N)	Years to onset full (mean, sd)
Chronic cystitis	85	26,403	5.7 (3.8)	83	26,130	5.7 (3.8)
Multiple sclerosis	95	48,811	5.8 (3.3)	93	48,255	5.8 (3.3)
Brain/CNS cancer	122	49,094	6 (3.5)	119	48,518	6 (3.6)
Schizophrenia	126	48,952	6.2 (3.7)	123	48,379	6.3 (3.8)
Systemic lupus erythematosus	157	48,703	4.7 (2.8)	153	48,135	4.7 (2.9)
Endometriosis	164	26,018	4.7 (3.5)	160	25,750	4.8 (3.5)
Vascular dementia	208	26,106	8.5 (3.1)	206	25,809	8.6 (3.1)
Amyotrophic lateral sclerosis	262	48,934	5.4 (3)	252	48,370	5.5 (3)
Liver fibrosis/cirrhosis	305	48,799	7.1 (3.7)	298	48,230	7.1 (3.7)
Inflammatory bowel disease	326	48,425	6.3 (3.7)	324	47,855	6.3 (3.7)
Gynaecological cancer	365	25,702	5.7 (3.7)	361	25,437	5.7 (3.7)
Major depression	376	47,961	3.3 (2.6)	369	47,398	3.3 (2.6)
Alzheimer's dementia	499	25,790	7.5 (3.3)	491	25,502	7.6 (3.3)
Lung cancer	551	48,636	7 (3.7)	543	48,064	6.9 (3.7)
Parkinson's disease	699	48,408	5.5 (3.5)	682	47,847	5.5 (3.5)
Rheumatoid arthritis	702	47,960	6.8 (3.6)	690	47,408	6.8 (3.6)
Colorectal cancer	716	48,278	7 (3.8)	706	47,712	7 (3.8)
Ischaemic stroke	1,015	48,012	7 (3.7)	989	47,469	7 (3.7)
Breast cancer	1,044	24,586	6.2 (3.6)	1,033	24,329	6.2 (3.6)
Prostate cancer	1,117	21,276	7 (3.6)	1,107	20,986	7 (3.6)
Chronic obstructive pulmonary disease	1,992	46,403	6.6 (3.6)	1,949	45,889	6.6 (3.6)
Type 2 diabetes	2,275	44,519	6.4 (3.5)	2,236	44,034	6.3 (3.5)
Ischaemic heart disease	3,123	44,011	6.3 (3.7)	3,086	43,512	6.3 (3.7)
Death	4,580	44,654	7.9 (3.5)	4,471	44,183	7.9 (3.5)

Table 1. Counts for cases and controls with mean time-to-onset for 23 incident morbidities and mortality in the UK Biobank (N=49,234). Mean time-to-onset is summarised for each disease over a 16-year follow-up period. Summary information is provided for basic and fully-adjusted Cox PH models. Alzheimer's and vascular dementia were restricted to age of event of 65 years or above. Sex-specific traits are stratified. CNS: central nervous system.

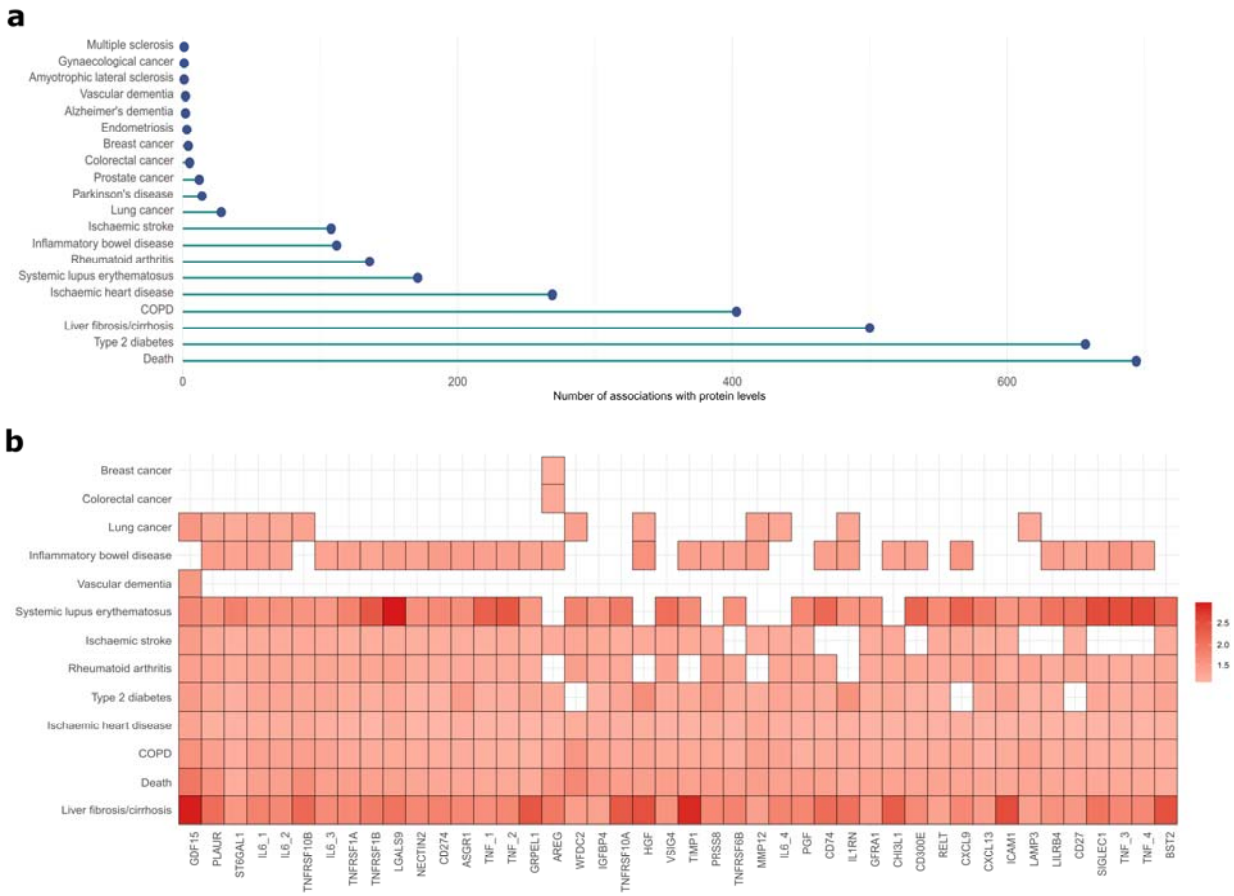

Figure 2. Individual protein associations with incident outcomes in the UK Biobank (N=49,234). a, Number of associations between protein analytes and time-to-onset for 20 outcomes that had $P_{Bonferroni} < 5.4x10^{-6}$ in both basic and fully-adjusted Cox PH models. There were 3,123 associations in total involving 1,052 protein analytes. **b,** Hazard ratios (HR) are plotted for the 44 protein analytes that were associated with eight or more outcomes in fully-adjusted Cox PH models. Each association is represented by a rectangle. Every association identified for these proteins had HR > 1 (red) and associations are shaded based on HR effect size (darkest colouration indicating larger magnitude of effect). The largest HR shown is for the association between LGALS9 levels and liver fibrosis/cirrhosis (HR =2.98). COPD: chronic obstructive pulmonary disease.

ProteinScores for incident outcomes

ProteinScores for 20 incident outcomes with a minimum of 150 cases available were trained using Cox PH elastic net regression with cross-validation in a training subset. Performance was quantified via incremental Cox PH models in the test subset, to obtain onset probabilities for calculation of AUC and Precision Recall AUC (PRAUC) estimates **(see Methods)**. This approach was repeated with fifty randomly sampled train and test subset combinations for each outcome with case:control ratios of 1:3 (**see Supplementary Fig. 4)** ProteinScores with the median difference in AUC were selected for each outcome **(see Methods)**. Cumulative time-to-onset distributions **(Supplementary Figs. 2-3)** indicated that amyotrophic lateral sclerosis, endometriosis, major depression and systemic lupus erythematosus were better-suited to 5-year onset assessments. All remaining ProteinScores were tested in the context of 10-year onset. **Fig. 3** provides an overview of training, feature selection and basic model evaluation for the ProteinScore selected for Alzheimer's dementia. Summaries of protein features for the 20 ProteinScores assessed in basic models are available in **Supplementary Tables 9-10**.

In tests for significant differences between receiver operating characteristic (ROC) curves between Cox models with basic adjustments with/out the selected ProteinScore, the addition of eleven ProteinScores resulted in $P_{Bonferroni} < 0.0025$ **(Supplementary Table 11)**. Differences in AUC and PRAUC for these eleven ProteinScores were then quantified across five incremental Cox PH models with increasing covariate adjustments (**Supplementary Table 12**). **Fig. 4** shows AUC and PRAUC estimates for basic and fully-adjusted models with/out the ProteinScore. Eight of the eleven ProteinScores (death, type 2 diabetes, ischaemic heart disease, Alzheimer's dementia, Parkinson's disease, liver fibrosis/cirrhosis, ischaemic stroke and COPD) had $P_{Bonferroni} < 0.0025$ in ROC model comparisons assessing the addition of the ProteinScores to fully-adjusted models. All eight of the best-performing ProteinScores were assessed for stratification of 10-year onset. The remaining three

scores (amyotrophic lateral sclerosis, rheumatoid arthritis and lung cancer) had nominal $P < 0.05$ in fully-adjusted model comparisons.

A final series of models considered only the ProteinScore for each trait tested **(Supplementary Table 12)**. For nine of the 11 traits, the ProteinScore model had a higher AUC than the models with basic and additional covariate adjustments. **Supplementary Figs. 5-15** visualise ROC and precision-recall curves for all of the incremental Cox PH models tested for each of the eleven ProteinScores assessed in fully-adjusted models.

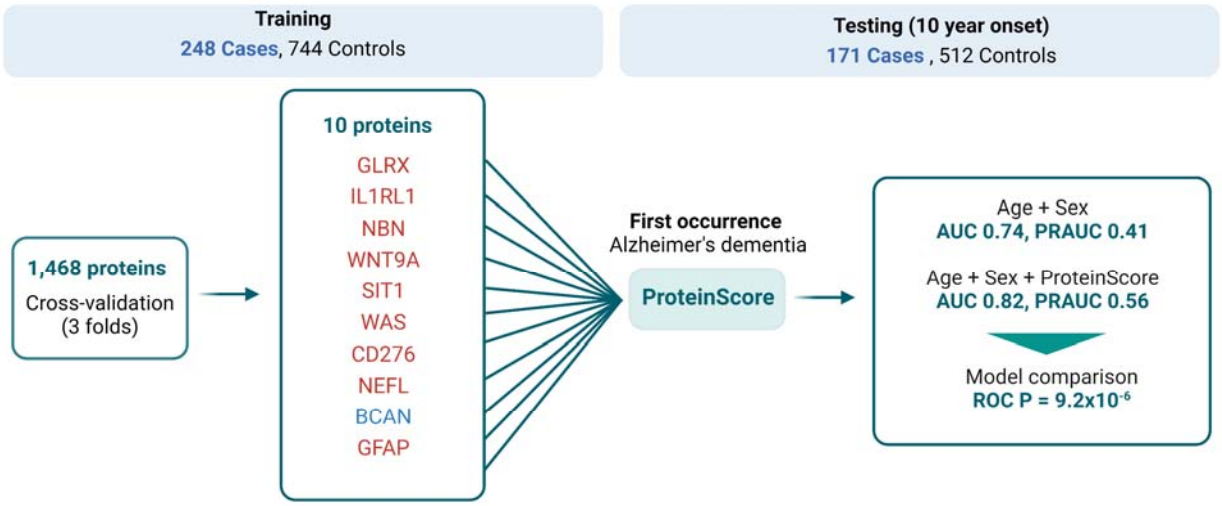

Figure 3. Example of feature selection and basic model evaluation for the Alzheimer's dementia ProteinScore. Of the 1,468 proteins considered, ten were selected and assigned weighting coefficients (positive = red, negative = blue) via Cox PH elastic net regression with 3-fold cross-validation (N_{cases} < 500) in the training sample. Weighting coefficients from this step were used to derive the ProteinScore in the test set. A model comparison between 10-year first occurrence of Alzheimer's dementia in the test sample with and without the ProteinScore yielded ROC P = $9.2x10^{-6}$, beyond adjustments for age and sex. Differences of 0.08 in AUC and 0.15 in PRAUC were observed due to the addition of the ProteinScore to the basic model.

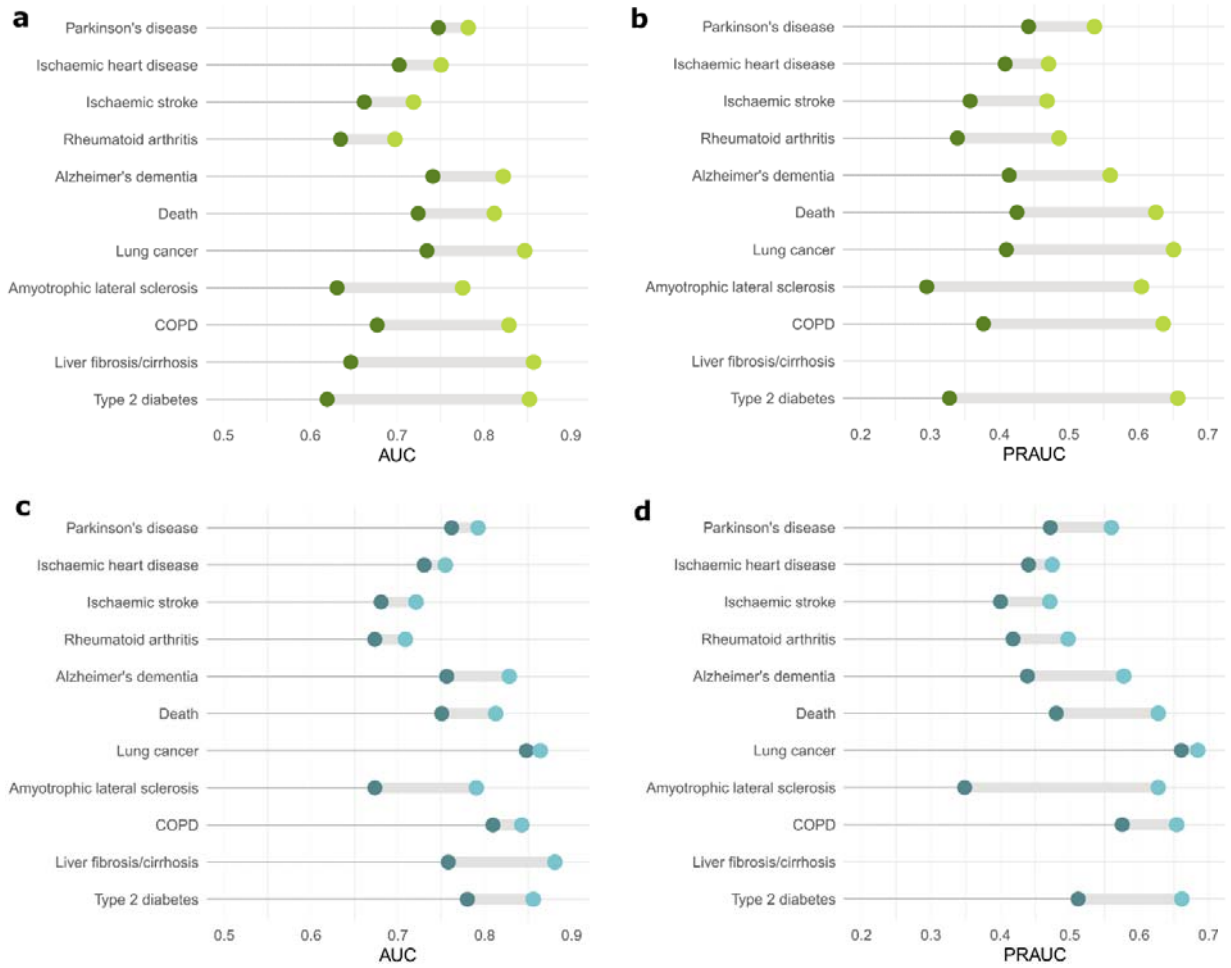

Figure 4. Predictive value offered by ProteinScores for incident outcomes in the UK Biobank. **a,** Differences in AUCs resulting from the addition of ProteinScores to basic models (adjusting for age and sex) – outcomes ordered by increasing AUC differences. **b,** Differences in PRAUCs resulting from the addition of ProteinScores to basic models. **c,** Differences in AUCs resulting from the addition of ProteinScores to basic models with additional risk factor adjustments (BMI, alcohol consumption, social deprivation, educational attainment and smoking status). **d,** Difference in PRAUCs resulting from the addition of ProteinScores to basic models with additional risk factor adjustments. All 11 of the ProteinScores had ROC $P_{Bonferroni} < 0.0025$ in basic model comparisons. Eight ProteinScores (all except those for lung cancer, amyotrophic lateral sclerosis and rheumatoid arthritis) had ROC $P_{Bonferroni} < 0.0025$ in fully-adjusted model comparisons. All ProteinScore performances shown correspond to 10-year onset, except amyotrophic lateral sclerosis that was assessed for 5-year onset.

Exploration of the type 2 diabetes ProteinScore clinical utility

The clinical utility that ProteinScores may offer was explored for type 2 diabetes, given the availability of a well-validated biomarker – Glycated haemoglobin (HbA1c) – in UK Biobank. HbA1c measures average long-term glucose over two to three months and is widely employed clinically to monitor pre-clinical diabetes risk (42-47mmol/mol) and diagnose the disease (with two repeated measurements >48mmol/mol) [27,28].

The HbA1c and type 2 diabetes ProteinScore markers were assessed individually and concurrently in 10-year Cox PH models in those in the type 2 diabetes test sample that had HbA1c available (873 cases, 2,542 controls, with a mean time-to-event of 5.4 (SD 2.8) years). There was a strong correlation between the ProteinScore and HbA1c (Pearson r=0.50 for rank-based inverse normal transformed variables). A contour plot of both variables grouped by those who went on to be diagnosed with type 2 diabetes over a 10-year period is presented in **Fig. 5a**. HbA1c levels increased across ProteinScore risk deciles, with individuals in the upper deciles of the ProteinScore falling within the clinical HbA1c screening threshold (42-47mmol/mol) for diabetes **(Fig. 5b)**.

In incremental Cox PH models for the 10-year onset of type 2 diabetes **(Fig. 5c)** the singular use of either the ProteinScore or HbA1c outperformed the singular use of the PGS (AUCs = 0.66, 0.84 and 0.85 for the PGS, HbA1c and ProteinScore, respectively). Adding the ProteinScore to a model with HbA1c, basic and additional health and lifestyle covariates significantly improved performance (AUC = 0.87, PRAUC = 0.73 for the model without the ProteinScore versus AUC = 0.89, PRAUC = 0.75, $P_{ROC\ Comparison} = 5.7x10^{-12}$ for the model with the ProteinScore). When the PGS was added to this model, a further improvement was observed (AUC = 0.90, PRAUC = 0.76 and $P_{ROC\ Comparison} = 9.4x10^{-6}$). **Supplementary Table 13** summarises AUC, PRAUC and ROC P statistics for the ten incremental models tested.

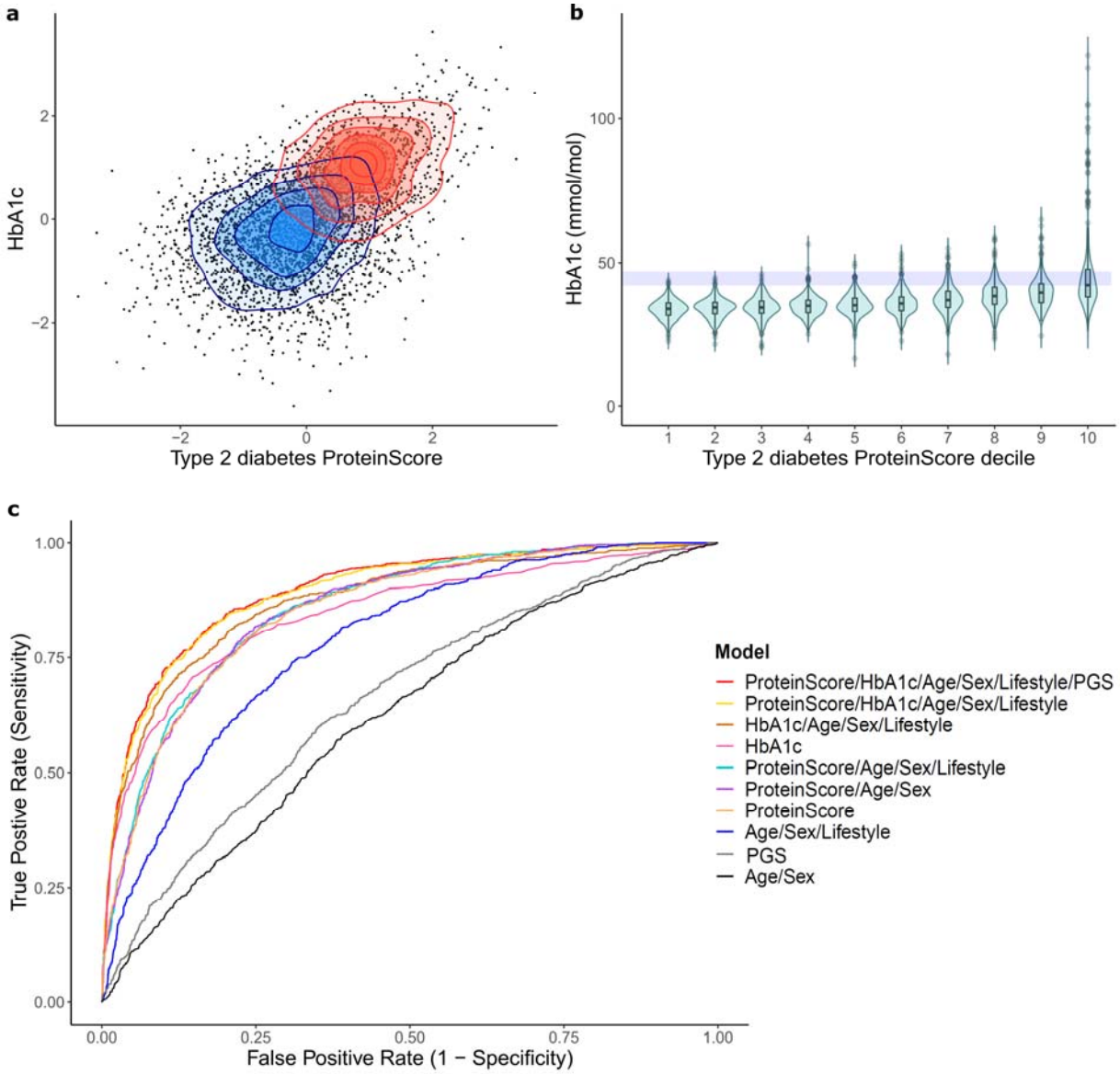

Figure 5. Clinical value of the type 2 diabetes ProteinScore beyond HbA1c and PGS in a subset of the test sample (N=3,608). a, Case (red) and control (blue) discrimination for HbA1c and the type 2 diabetes ProteinScore. Both markers were rank-based inverse normalised and scaled to have a mean of 0 and standard deviation of 1. Data for 873 cases and 2,542 controls with ProteinScores and HbA1c available are shown. **b,** HbA1c (mmol/mol) per decile of the type 2 diabetes ProteinScore. Shaded rectangle indicates the type 2 diabetes HbA1c screening threshold (42-47 mmol/mol). **c,** ROC curves for incremental 10-year onset models incorporating HbA1c and the type 2 diabetes ProteinScore and PGS individually and concurrently, alongside relevant covariates such as age, sex and detailed health and lifestyle factors.

Discussion

Developing scores that identify individuals at risk of a future event is a priority for prevention-based medicine during ageing [29]. Our study shows that protein-based scores (ProteinScores) significantly improve risk classification in the UK Biobank for 10-year onset of eight outcomes when adjusting for common risk factors. The type 2 diabetes ProteinScore outperformed both the corresponding PGS and a clinically-validated blood marker for type 2 diabetes monitoring and diagnosis, HbA1c. Over 16 years of follow-up, 3,123 significant associations between 1,052 circulating proteins and time-to-onset for 20 outcomes were also profiled, identifying circulating proteins that were indicative of multiple morbidities.

The breadth of electronic health data linkage and protein data available in UK Biobank provides a unique resource for profiling early molecular signatures of age-related disease. The ProteinScores developed in this study demonstrate that subsets of relatively few circulating proteins can add predictive value, up to a decade prior to formal clinical diagnoses. As available cases increase, it is likely that the performance of ProteinScores will be enhanced. Nonetheless, the improvement in AUC resulting from concurrent modelling of HbA1c and the type 2 diabetes ProteinScore suggests that the latter may provide additional clinical value. Although the increase over-and-above HbA1c, age, sex and lifestyle factors offered by the ProteinScore was modest (AUC 0.87 to 0.89, PRAUC 0.73 to 0.75), it was a significant improvement ($P = 5.7 \times 10^{-12}$). For the majority of outcomes, modelling the ProteinScore in isolation resulted in higher AUCs than models with basic and lifestyle covariate adjustments. This suggests that ProteinScores absorb a large proportion – if not all – of the signal and may offer a streamlined set of metrics to proxy for an individual's health status. This often-enhanced predictive quality of the scores presents an exciting opportunity to reconsider how best to formulate (and maintain) modern clinical prediction models. This is an important consideration given that self-reported measures are known to be variable in accuracy and are often

misreported [30]. Additionally, while much interest is currently devoted to employing PGS for disease prediction, they neglect environmental components of disease risk and may therefore be limited in the context of complex age-related disease [31,32]. Our ProteinScore for type 2 diabetes (that relies on 97 proteins) outperformed the PGS in this study, which is likely due to proteins representing an interface that captures genetic, environmental and lifestyle contributions to disease risk. However, a further modest improvement to the fully-adjusted type 2 diabetes model resulted from the addition of the PGS (AUC 0.89 to 0.90, PRAUC 0.75 to 0.76, with $P = 9.4x10^{-6}$). Deriving ProteinScores for multiple diseases within the same individuals may also facilitate an improved understanding of multimorbidity. For example, if an individual falls within the top 5% of the ProteinScore distributions for type 2 diabetes and Alzheimer's dementia, this information may enhance personalised intervention plans. The ProteinScore for Alzheimer's dementia was also largely unchanged upon addition of lifestyle covariates to incremental models. As therapeutic interventions for neurodegenerative diseases have greater efficacy when implemented earlier in the disease pathogenesis [33–35], the ProteinScore for Alzheimer's dementia (that relies on 10 proteins) may hone trial recruitment.

The method for ProteinScore generation selects proteins that, in combination, are predictive of outcomes, but these do not necessarily represent the most probable drivers of disease. It is likely that a subset of the 3,123 individual protein-disease associations we report represent direct mediators of disease. These proteins should be interrogated further via methods such as Mendelian randomisation to determine their causal roles and suitability as therapeutic targets. Elevated growth differentiation factor 15 (GDF15) was one of three proteins that associated with the largest number of morbidities, which is in concordance with previous screening of the circulating proteome against multimorbidity and mortality [36,37]. All 44 of the proteins that were associated with eight or more morbidities had associations with hazard ratios greater than 1, indicating that elevated levels of these proteins may serve as early warning signatures of disease onset. Similarly, increased levels of neurofilament light

(NEFL) were associated with higher incidence of multiple neurological traits (Parkinson's disease, Alzheimer's dementia, multiple sclerosis, amyotrophic lateral sclerosis and ischaemic stroke). These diseases are hallmarked by neuron degradation and NEFL may therefore be a consequential marker that is released into the blood upon breakdown of synapses [38,39] . R-spondin-1 (RSPO1) – the most significant association for endometriosis – is a regulator of endometrial mesenchymal stem-like cells during menstruation [40]. Stanniocalcin 2 (STC2) was the top candidate (based on P values) for breast cancer and is being explored as a therapeutic target for tumour regulation [41]. Similarly, TSPAN1 – the lead association for prostate cancer – is known to increase prostate cancer cell migration [42]. In instances such as these, where biopsies are invasive and early diagnosis is critical for survival, blood-based markers may offer value.

Across the 16-year window of follow-up in individual Cox PH models, a subset of associations violated the Cox PH assumption at the local (protein) level. The majority of violations occurred within the final six-year window of linkage, which has a high degree of sparsity and case-control imbalance due to censoring at June 2016 (whereas ICD linkage for cases extended to 2022). Our Shiny app https://protein-disease-ukb.optima-health.technology [Username: ukb_disease, Password: shinyappUKB] provides visualisations for sensitivity analyses run across cases over successive years of follow up, allowing for interrogation of individual protein-outcome relationships. In these sensitivity analyses, fewer violations were observed when time-to-event was restricted to 5 years for cases than at 16 years, indicating that hazard ratios may be less variable in near-term risk stratification. This is often of importance to clinicians and patients for behaviour change and intervention strategies. The Shiny app also visualises the 3,123 fully-adjusted associations in a network view, allowing users to view overlapping signatures between multiple proteins and the onset of multiple diseases.

This study has several limitations. First, a subset of 6,385 individuals in the UKB-PPP sample were selected by consortium members for enrichment of certain diagnoses and this non-random selection can introduce biases. Second, as UK Biobank currently represents the largest population with comprehensive Olink proteomics and electronic health data linkage, it was not possible to source an external test set for the ProteinScores. Third, variation in protein analyte levels across measurement technologies has been reported [43]. Results should therefore be corroborated across panels in future. Fourth, the UK Biobank population studied here is largely comprised of individuals with European ancestry and a restricted age range (40-71 years, with a mean of 57 years); future studies in equally well-characterized cohorts will be needed to assess how well ProteinScores translate to other populations and ethnicities. Fifth, non-linear trajectories of blood-based protein signatures are known to exist across the life course in the context of ageing [44]. Similarly, the presence of morbidities and medication use at baseline was not accounted for. These factors should be considered in disease-specific analyses in future. Finally, although a comprehensive set of major age-related morbidities were studied, many diseases were not included in this work. Continued linkage and proteomic sampling will expand the applications of ProteinScores to further diseases.

In conclusion, this study quantified circulating proteome signatures that are reflective of multiple individual disease states across mid-to-later life. ProteinScores for the incidence of eight incident outcomes significantly improved AUCs for 10-year onset beyond common health and lifestyle factors ($P_{Bonferroni} < 0.0025$), with the type 2 diabetes ProteinScore improving AUC beyond both a PGS and HbA1c. A total of 3,123 individual protein-disease associations were also profiled across the 16-year follow-up period. These data suggest that proteomic features are powerful tools for honing risk stratification.

Methods

The UK Biobank sample population

UK Biobank (UKB) is a population-based cohort of around 500,000 individuals aged between 40-69 years that were recruited between 2006 and 2010. Genome-wide genotyping, exome sequencing, electronic health record linkage, whole-body magnetic resonance imaging, blood and urine biomarkers and physical and anthropometric measurements are available. More information regarding the full measurements can be found at: https://biobank.ndph.ox.ac.uk/showcase/. The UK Biobank Pharma Proteomics Project (UKB-PPP) is a precompetitive consortium of 13 biopharmaceutical companies funding the generation of blood-based proteomic data from UKB volunteer samples.

Proteomics in the UK Biobank

The UKB-PPP sample includes 54,306 UKB participants and 1,474 protein analytes measured across four Olink panels (Cardiometabolic, Inflammation, Neurology and Oncology: annotation information provided in **Supplementary Table 1**) [23]. A randomised subset of 46,673 individuals were selected from baseline UKB, with 6,385 individuals selected by the UKB-PPP consortium members and 1,268 individuals included that participated in a COVID-19 study. The randomised samples have been shown to be highly representative of the wider UKB population, whereas the consortium-selected individuals were enriched for 122 diseases [23]. Details on sample selection for UKB-PPP, in addition to processing and quality control information for the Olink assay are provided in **Supplementary Information**. Of 54,309 individuals that had protein data measured, there were 54,189 that were available after quality control exclusions with 1,474 Olink protein analytes measured (annotations in **Supplementary Table 1**) [23]. The sample is predominantly of European

ancestry (93%), but also has individuals with black/black British, Asian/Asian British, Chinese, mixed, other and missing ethnic backgrounds (7%).

Supplementary Fig. 1 summarises the processing steps applied to this dataset to derive a complete set of measurements for use. Briefly, of 107,161 related pairs of individuals (calculated through kinship coefficients > 0 across the full UKB cohort), 1,325 pairs were present in the 54,189 individuals. After exclusion of 108 individuals in multiple related pairs, in addition to one individual randomly selected from each of the remaining pairs, there were 52,962 individuals. A further 3,728 individuals were excluded due to having >10% missing protein measurements. Four proteins that had >10% missing measurements (CTSS.P25774.OID21056.v1 and NPM1.P06748.OID20961.v1 from the neurology panel, PCOLCE.Q15113.OID20384.v1 from the cardiometabolic panel and TACSTD2.P09758.OID21447.v1 from the oncology panel) were then excluded. The remaining 1% of missing protein measurements were imputed by K-nearest neighbour (k=10) imputation using the impute R package (Version 1.60.0) [45]. The final dataset consisted of 49,234 individuals and 1,468 protein analytes. Assessments of protein batch, study centre and genetic principal components suggested that these factors had minimal effects on protein levels (lowest correlation between protein levels and residuals of 0.94) **(Supplementary Information)**. Therefore, protein levels were not adjusted for these factors.

Phenotypes in the UK Biobank

Demographic and phenotypic information for the 49,234 individuals with complete protein data for 1,468 analytes are available in **Supplementary Table 2**. Baseline measurements of several covariates were used in fully-adjusted models: BMI (weight in kilograms divided by height in metres squared), alcohol intake frequency (1 = Daily or almost daily, 2 = Three-Four times a week, 3 = Once or twice a week, 4 = One-Three times a month, 5 = Special occasions only, 6 = Never), the Townsend index of deprivation (higher score representing greater levels of deprivation) and smoking

status (0 = Never, 1 = Previous, 2 = Current) and education status (1 = college/university educated, 0 = all other education). Of the 49,234 individuals with complete protein data, there were 55, 55, 244 and 62 missing entries for alcohol, smoking, BMI and deprivation, respectively. No imputation of missing data was performed. There were an additional 188 and 59 individuals that answered 'prefer not to answer' and were excluded from smoking and alcohol variables, respectively. Twenty genetic principal components were available for 48,821 of the individuals that had genetic information. Study centre was also available in these 48,821 individuals and included 22 centres where blood sampling and clinical assessments took place.

Electronic health data linkage in the UK Biobank

Electronic health linkage to NHS GP and hospital records was used to collate incident disease diagnoses. Death information was sourced from the death registry data available through the UK Biobank. The following 23 diseases were included: liver fibrosis/cirrhosis, systemic lupus erythematosus, type 2 diabetes, amyotrophic lateral sclerosis, Alzheimer's dementia, endometriosis, chronic obstructive pulmonary disease (COPD), inflammatory bowel disease, rheumatoid arthritis, ischaemic stroke, Parkinson's disease, vascular dementia, ischaemic heart disease, major depressive disorder, schizophrenia, multiple sclerosis, chronic cystitis and lung, prostate, breast, gynaecological, brain/CNS and colorectal cancers. These represent a selection of leading age-related causes of morbidity, mortality and disability. A full summary of the methodology and UK Biobank sources used to derive incident case status is provided in **Supplementary Information**. Code lists used to extract traits from GP (read2/read3) and hospital (ICD9/ICD10) level linkage were sourced from the HDR phenotype library CALIBER disease code lists [available at: https://phenotypes.healthdatagateway.org/]. These have been collated in a recent study that characterised 308 conditions in 4 million individuals across the English National Health Service [46]. In all cases, any relevant self-report of the disease at baseline was used to ensure cases occurred post-

baseline. Non-specific codes were excluded and the final set of ICD and read codes used to define disease cases are summarised in **Supplementary Tables 14-36**. Gynaecological cancers were grouped together as a single outcome, consisting of ovarian, uterine and cervical primary malignancies, in addition to cervical carcinomas *in situ*. In all analyses involving sex-specific diseases, the population was stratified to males or females and sex was not included as a covariate in incremental Cox PH assessments. Traits that were stratified included gynaecological cancer, breast cancer, endometriosis and chronic cystitis (all female-stratified) and prostate cancer (male-stratified). Inflammatory bowel disease included both Crohn's disease and ulcerative colitis linkage codes.

Incident disease calculation in the UK Biobank

Dates of diagnoses for each disease were ascertained through electronic health linkage **(Supplementary Tables 14-36)**. Using the date of baseline appointment, time-to-first-onset for each diagnoses in years was calculated. Although ICD code linkage was available up to 2022, the date of censoring for controls was set to June 2016 – the most recent GP linkage extraction date. Time-to-onset for controls was defined as the time from baseline to censoring date. If controls had died prior to the censor date, age at death was taken forward for censoring calculations. Any cases that were prevalent at baseline were excluded. The same approach was used to define time-to-onset for death as an outcome, with a censor date set to December 2021 – the last linkage update for the death registry data. Alzheimer's and vascular dementias were restricted to age at onset or censoring of 65 years or older in all analyses. Sex-specific traits were stratified across all analyses.

Individual Cox proportional hazards analyses

Cox proportional hazards models were run between each protein and each incident disease using the 'survival' package (Version 3.4-0) [47] in R (Version 4.2.0) [48]. Protein levels were rank-based inverse normalised and scaled to have a mean of 0 and standard deviation of 1 prior to analyses. Basic Cox PH models for sex-stratified traits included age at baseline as a covariate, whereas the remaining

models adjusted for age and sex. Fully-adjusted models further controlled for education status, BMI, smoking status, social deprivation rank and alcohol intake frequency. A Bonferroni threshold for multiple testing based on the 385 components that explained 80% of the cumulative variance in the 1,468 protein analyte levels **(Supplementary Table 3)** and 24 outcomes tested was applied across all Cox PH models ($P < 0.05/(385 \times 24) = 5.4 \times 10^{-6}$). Proportional hazards assumptions were checked through examination of protein-level Schoenfeld residuals. A sensitivity analysis was performed for each of the 35,232 fully-adjusted associations tested, restricting cases to successive years of follow-up. These sensitivity analyses were visualised using the Shiny package (Version 1.7.3) [49] in R. A network visualisation was also created within the Shiny interface to highlight the fully-adjusted associations that had $P_{Bonferroni} < 5.4 \times 10^{-6}$ using networkD3 (Version 3.0.4) [50] and igraph (Version 1.3.5) [51] packages.

ProteinScore training

MethylPipeR [52] is an R package with accompanying user interface that we have previously developed for systematic and reproducible development of incident disease predictors. Using MethylPipeR, ProteinScores that considered 1,468 Olink protein levels were trained using Cox PH elastic net regression via the R package Glmnet (Version 4.1-4) [53]. Penalised regression minimises overfitting by the use of a regularisation penalty and the best shrinkage parameter (λ) was chosen by cross-fold validation with alpha fixed to 0.5. Of the 24 outcomes featured in the individual Cox PH analyses, 20 that had a minimum case count of 150 were selected for ProteinScore development. The ProteinScore training strategy is summarised in **Supplementary Fig. 4**. Briefly, 50 training iterations were performed that randomised sample selection by seeds (randomly sampled between 1 and 5000). For each iteration, cases and controls were randomly split into 50% groups for training and testing. From the 50% training control population, a subset of controls were then randomly sampled to give a case:control ratio of 1:3 in order to balance the datasets. For traits with over 1000 cases in training

samples 10 folds were used. For traits with between 500 and 1000 cases in training, five folds were used. Three folds were used when there were fewer than 500 cases in the training sample. Protein levels were rank-based inverse normalised and scaled to have a mean of 0 and standard deviation of 1 in the training set. The linear combination of weighting coefficients for selected protein features from cross-validation within the folds of the training set were then used to generate a ProteinScore for each individual in the test samples. Of the 50 training iterations tested, models that had no features selected were documented **(Supplementary Table 11)**.

Assessment of ProteinScore performance

Cumulative time-to-onset distributions for cases **(Supplementary Figs. 2-3)** indicated that amyotrophic lateral sclerosis, endometriosis, major depression and systemic lupus erythematosus were better-suited to 5-year onset assessments in the test sample. All remaining ProteinScores were tested in the context of 10-year onset. Across each of the 50 ProteinScore iterations for each trait, 50% of cases and controls that were not randomly selected for training were reserved for testing. For a visualisation of the test set sampling and assessment strategy, see **Supplementary Fig. 4**. In the test set, cases that had time-to-event up to or including the 5-year or 10-year thresholds used for onset prediction were selected, while cases beyond the threshold were placed with the control population, which was then randomly sampled in a 1:3 ratio. Weighting coefficients for features selected during ProteinScore training were used to project scores into the test sample. Incremental Cox PH models were run in the test sample to obtain cumulative baseline hazard and onset probabilities, which were used to obtain AUC and PRAUC estimates. The test set sampling strategy ensured that while the majority of cases occurred up to the onset threshold, there were a small proportion (~3%) of cases included in Cox PH models with onset times after the 10- or 5-year threshold, to simulate a real-world scenario for risk stratification. If cases fell beyond the 5-year or 10-year threshold for onset, they were recoded as controls in the AUC calculation. Cumulative

baseline hazard probabilities were calculated using the 'gbm' package (Version 2.1.8.1) [54]. Survival probabilities were then generated through taking the exponential of the negative cumulative baseline hazard at 5 or 10 years to the power of the Cox PH prediction probabilities. ProteinScore onset probabilities were calculated as one minus these survival probabilities. AUC, PRAUC and ROC statistics were extracted for the survival probabilities using the calibration function from the 'caret' package (Version) [55] and the evalmod function from the 'MLmetrics' package (Version 1.1.1) [56].

ProteinScores that yielded the median incremental difference to the AUC of a basic model (adjusting for age- or age- and sex) were selected from the 50 possible ProteinScores for each trait. If no features were selected during training, models were weighted as performance of 0 in the median model selection. In some instances, features were selected during training and incremental Cox PH models were run successfully, but the random sampling of the test set did not include a case with time-to-event at or after the 5-year or 10-year onset threshold. Therefore, these models were excluded as cumulative baseline hazard distributions did not reach the onset threshold and could not be extracted for AUC and PRAUC calculations. The number of models, with minimum and maximum performance was documented **(Supplementary Table 11)**. Taking this approach mitigated against the presence of extreme case:control profiles driving ProteinScore performance and minimised the possibility of bias being introduced by selecting train and test samples based on matching for specific population characteristics.

ROC P-value tests were used to ascertain whether the improvements offered by 20 selected ProteinScores for each outcome were statistically significant, beyond a basic Cox PH model. A Bonferroni threshold for ROC P was used based on the 20 traits ($P < 0.05/20 = 0.0025$). Eleven ProteinScores were therefore taken forward for analysis with a more detailed set of covariates. Differences in AUC and PRAUC were then quantified for the addition of the ProteinScores to basic and fully-adjusted models. Fully-adjusted models included further adjustment for health and lifestyle

covariates (education status, BMI, smoking status, social deprivation rank and alcohol intake frequency). A series of models that included only the ProteinScore were also considered for each outcome. AUC and PRAUC statistics were extracted using the onset probabilities from these incremental models. A comparison of ROC curves was also performed between fully-adjusted models, with and without the addition of the ProteinScores and ProteinScores with $P_{Bonferroni} < 0.0025$ in these model comparisons were reported. The 'precrec' package (Version 0.12.9) [57] was used to generate ROC and Precision-Recall curves for each ProteinScore.

Clinical value of the type 2 diabetes ProteinScore

Glycated Haemoglobin (HbA1c) is a blood-based measure of chronic glycemia that is highly predictive of type 2 diabetes events and is recommended as a test of choice for the monitoring and diagnosis of type 2 diabetes [27,28]. HbA1c (mmol/mol) measurements (fieldID 30750) and the type 2 diabetes polygenic risk score (PGS) available in UK Biobank (fieldID 26285) were extracted. A contour plot showing both variables grouped by those who went on to be diagnosed with type 2 diabetes over a 10-year period was created. HbA1c levels were also plotted against ProteinScore risk deciles. HbA1c and the ProteinScore levels were rank-based inverse normalised and assessed individually and concurrently in incremental models for 10-year onset of type 2 diabetes. There were 873 type 2 diabetes cases and 2,542 controls that had HbA1c, PGS and ProteinScore measures available (mean time-to-event of 5.4 (SD 2.8) years). A Pearson correlation coefficient (*r*) between the transformed HbA1c and ProteinScore levels was calculated. The 10-year incremental Cox PH models were used to derive onset probabilities for calculation of AUCs and PRAUCs after adding the ProteinScore to models adjusting for HbA1c, as well as basic and additional health and lifestyle covariates and the type 2 diabetes PGS. Model comparisons were used (test of the difference in ROC curves) to quantify the value added beyond age, sex, additional covariates, PGS and HbA1c that the type 2 diabetes ProteinScore offered. Models that included HbA1c or the ProteinScore with no other

covariates were also considered, in addition to a model that considered the PGS alongside age and sex.

Ethics declarations

All participants provided informed consent. This research has been conducted using the UK Biobank Resource under approved application numbers 65851, 20361, 26041, 44257, 53639, 69804.

Data availability

Datasets generated in this study are made available in **Supplementary Tables**. These data will also be deposited in an open-access repository upon publication. Proteomics data is available in UK Biobank under return dataset [return dataset ID and URL will be provided upon publication, depending on time of official publication].

Code availability

Code is available with open access at the following Github repository:

https://gitfront.io/r/user-2029007/EKPBzX5Fh2KV/UKB-project/

Acknowledgements

This research was funded in whole, or in part, by the Wellcome Trust [108890/Z/15/Z]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

We thank the participants, contributors, and researchers of UK Biobank for making data available for this study – with special thanks to Lauren Carson, John Busby, Naomi Allen and Rory Collins for making the study possible. We are grateful to the research & development leadership teams at the thirteen participating UKB-PPP member companies (Alnylam Pharmaceuticals, Amgen,

AstraZeneca, Biogen, Bristol-Myers Squibb, Calico, Genentech, Glaxo Smith Klein, Janssen Pharmaceuticals, Novo Nordisk, Pfizer, Regeneron, and Takeda) for funding the study. We thank the Legal and Business Development teams at each company for overseeing the contracting of this complex, precompetitive collaboration – with particular thanks to Erica Olson of Amgen, Andrew Walsh of GSK, and Fiona Middleton of AstraZeneca. The Biogen team is especially thankful to Helen McLaughlin for her project management support. Finally, we thank the team at Olink Proteomics (Philippa Pettingell, Klev Diamanti, Cindy Lawley, Linda Jung, Sara Ghalib, Ida Grundberg and Jon Heimer) for their consistent logistic support throughout the project – with special thanks to Evan Mills for co-championing the project and leading internal activities at Olink.

R.E.M. is supported by Alzheimer's Society major project grant AS-PG-19b-010. R.F.H is supported by a MRC IEU Fellowship. D.A.G. is supported by the Wellcome Trust Translational Neuroscience programme [**108890/Z/15/Z]**.

Author contributions

D.A.G., R.F.H., R.E.M., B.S., C.F., and Z.K., conceptualised the study design and consulted on methods and results. D.A.G., carried out all analyses. D.A.G., R.F.H., B.B.S., and R.E.M., drafted the article. R.A., and J.G., conducted preliminary analyses. T.L., and K.F., performed quality control on the proteomics dataset. All authors reviewed and approved of the manuscript.

Competing interests

B.B.S., R.A., J.G., T.L., K.F., and H.R., are employed by Biogen. C.N.F., Z.K., D.A.G., and T.M., are employed by Optima partners. D.A.G., R.F.H., and R.E.M., have received consultancy fees from Optima Partners. R.E.M. is an advisor to the Epigenetic Clock Development Foundation. R.F.H., has received consultant fees from Illumina. All other authors declare no competing interests.

Materials and correspondence

Correspondence and material requests should be sent to Dr Benjamin Sun at benjamin.sun@biogen.com or Prof Riccardo Marioni at riccardo.marioni@ed.ac.uk.

References

1. Fountzilas, E., Tsimberidou, A. M., Vo, H. H. & Kurzrock, R. Clinical trial design in the era of precision medicine. *Genome Med.* **14**, 101 (2022).
2. Duarte, T. T. & Spencer, C. T. Personalized Proteomics: The Future of Precision Medicine. *Proteomes* **4**, 29 (2016).
3. Al-Nesf, M. A. Y. *et al.* Prognostic tools and candidate drugs based on plasma proteomics of patients with severe COVID-19 complications. *Nat. Commun.* **13**, 946 (2022).
4. Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* **9**, 3268 (2018).
5. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet. 2021 5312* **53**, 1712–1721 (2021).
6. Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* eabj1541 (2021).
7. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
8. Gudmundsdottir, V. *et al.* Circulating protein signatures and causal candidates for type 2 diabetes. *Diabetes* **69**, 1843–1853 (2020).
9. Nurmohamed, N. S. *et al.* Targeted proteomics improves cardiovascular risk prediction in secondary prevention. *Eur. Heart J.* **43**, 1569–1577 (2022).
10. Huth, C. *et al.* Protein markers and risk of type 2 diabetes and prediabetes: a targeted proteomics approach in the KORA F4/FF4 study. *Eur. J. Epidemiol.* **34**, 409–422 (2019).
11. LaFramboise, W. A. *et al.* Serum protein profiles predict coronary artery disease in symptomatic patients referred for coronary angiography. *BMC Med.* **10**, 157 (2012).
12. Mendelian Randomization Studies in Stroke: Exploration of Risk Factors and Drug Targets With Human Genetic Data | Stroke. https://www.ahajournals.org/doi/full/10.1161/STROKEAHA.120.032617.
13. Ritchie, S. C. *et al.* Integrative analysis of the plasma proteome and polygenic risk of cardiometabolic diseases. *Nat. Metab.* **3**, 1476–1483 (2021).

14. Sathyan, S. *et al.* Plasma proteomic profile of age, health span, and all-cause mortality in older adults. *Aging Cell* **19**, e13250 (2020).

15. Borrebaeck, C. A. K. Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer. *Nat. Rev. Cancer* **17**, 199–204 (2017).

16. Ganz, P. *et al.* Development and validation of a protein-based risk score for cardiovascular outcomes among patients with stable coronary heart disease. *JAMA* **315**, 2532–2541 (2016).

17. Wang, Z. *et al.* Metabolomic Pattern Predicts Incident Coronary Heart Disease. *Arterioscler. Thromb. Vasc. Biol.* **39**, 1475–1482 (2019).

18. Machado-Fragua, M. D. *et al.* Circulating serum metabolites as predictors of dementia: a machine learning approach in a 21-year follow-up of the Whitehall II cohort study. *BMC Med.* **20**, 334 (2022).

19. Eiriksdottir, T. *et al.* Predicting the probability of death using proteomics. *Commun. Biol.* **4**, 758 (2021).

20. Lind, L. *et al.* Large-Scale Plasma Protein Profiling of Incident Myocardial Infarction, Ischemic Stroke, and Heart Failure. *J. Am. Heart Assoc.* **10**, e023330 (2021).

21. Williams, S. A. *et al.* Plasma protein patterns as comprehensive indicators of health. *Nat. Med.* **25**, 1851–1857 (2019).

22. Buergel, T. *et al.* Metabolomic profiles predict individual multidisease outcomes. *Nat. Med.* 1–12 (2022) doi:10.1038/s41591-022-01980-3.

23. Sun, B. B. *et al.* Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *bioRxiv* **20**, 2022.06.17.496443 (2022).

24. Kyu, H. H. *et al.* Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet* **392**, 1859–1922 (2018).

25. James, S. L. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 354 Diseases and Injuries for 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet* **392**, 1789–1858 (2018).

26. Feigin, V. L. *et al.* Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* **18**, 459 (2019).

27. Sherwani, S. I., Khan, H. A., Ekhzaimy, A., Masood, A. & Sakharkar, M. K. Significance of HbA1c Test in Diagnosis and Prognosis of Diabetic Patients. *Biomark. Insights* **11**, 95–104 (2016).

28. WHO. Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus. Abbreviated Report of a WHO Consultation. WHO/NMH/CHP/CPM/11.1.

29. Next Steps For Risk Stratification in the NHS. NHS England. Available at: https://www.england.nhs.uk/wp-content/uploads/2015/01/nxt-steps-risk-strat-glewis.pdf.

30. Bollepalli, S., Korhonen, T., Kaprio, J., Anders, S. & Ollikainen, M. EpiSmokEr: A robust classifier to determine smoking status from DNA methylation data. *Epigenomics* **11**, 1469–1486 (2019).

31. Li, R., Chen, Y., Ritchie, M. D. & Moore, J. H. Electronic health records and polygenic risk scores for predicting disease risk. *Nat. Rev. Genet.* **21**, 493–502 (2020).

32. Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* **12**, 44 (2020).

33. Barnett, J. H., Lewis, L., Blackwell, A. D. & Taylor, M. Early intervention in Alzheimer's disease: a health economic study of the effects of diagnostic timing. *BMC Neurol.* **14**, 101 (2014).

34. Crous-Bou, M., Minguillón, C., Gramunt, N. & Molinuevo, J. L. Alzheimer's disease prevention: from risk factors to early intervention. *Alzheimers Res. Ther.* **9**, 71 (2017).

35. Foster, L. A. & Salajegheh, M. K. Motor Neuron Disease: Pathophysiology, Diagnosis, and Management. *Am. J. Med.* **132**, 32–37 (2019).

36. Tanaka, T. *et al.* Plasma proteomic biomarker signature of age predicts health and life span. *eLife* **9**, 1–24 (2020).

37. Bao, X. *et al.* Growth differentiation factor-15 is a biomarker for all-cause mortality but less evident for cardiovascular outcomes: A prospective study. *Am. Heart J.* **234**, 81–89 (2021).

38. Alirezaei, Z. *et al.* Neurofilament light chain as a biomarker, and correlation with magnetic resonance imaging in diagnosis of CNS-related disorders. *Mol. Neurobiol.* **57**, 469–491 (2020).

39. Wu, J. *et al.* Plasma neurofilament light chain: A biomarker predicting severity in patients with acute ischemic stroke. *Medicine (Baltimore)* **101**, e29692 (2022).

40. Xu, S., Chan, R. W. S., Li, T., Ng, E. H. Y. & Yeung, W. S. B. Understanding the regulatory mechanisms of endometrial cells on activities of endometrial mesenchymal stem-like cells during menstruation. *Stem Cell Res. Ther.* **11**, 239 (2020).

41. Qie, S. & Sang, N. Stanniocalcin 2 (STC2): a universal tumour biomarker and a potential therapeutical target. *J. Exp. Clin. Cancer Res.* **41**, 161 (2022).

42. Munkley, J. *et al.* The cancer-associated cell migration protein TSPAN1 is under control of androgens and its upregulation increases prostate cancer cell migration. *Sci. Rep.* **7**, 5249 (2017).

43. Pietzner, M. *et al.* Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nat. Commun. 2021 121* **12**, 1–13 (2021).

44. Lehallier, B. *et al.* Undulating changes in human plasma proteome profiles across the lifespan. *Nat. Med.* **25**, 1843–1850 (2019).

45. Hastie, T., Tibshirani, R., Narasimhan, B. & Chu, G. Package 'impute' Title impute: Imputation for microarray data. R package version 1.60.0. (2022).

46. Kuan, V. *et al.* A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *Lancet Digit. Health* **1**, e63–e77 (2019).

47. Therneau, T. M. A Package for Survival Analysis in R. R package version 3.2-7, https://CRAN.R-project.org/package=survival. Accessed April 2021. (2020).

48. (2017), R. C. T. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

49. Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B. shiny: Web Application Framework for R. R package version 1.7.3.9002, https://shiny.rstudio.com/.

50. J.J. Allaire, Christopher Gandrud, Kenton Russell and CJ Yetman. networkD3: D3 JavaScript Network Graphs from R. R package. https://CRAN.R-project.org/package=networkD3. (2017).

51. Csardi G, Nepusz T. The igraph software package for complex network research, InterJournal, Complex Systems 1695. https://igraph.org. (2006).

52. Cheng, Y. *et al.* DNA Methylation scores augment 10-year risk prediction of diabetes. *medRxiv* 2021.11.19.21266469 (2021) doi:10.1101/2021.11.19.21266469.

53. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J. Stat. Softw.* **39**, (2011).

54. Greenwell B, Boehmke B, Cunningham J, Developers G. gbm: Generalized Boosted Regression Models. R package version 2.1.8.1. (2022).

55. Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton, Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew, & Ziem, Luca Scrucca, Yuan Tang and Can Candan. caret: Classification and Regression Training. R package version 6.0-71. (2016).

56. Yan, Y. MLmetrics: Machine Learning Evaluation Metrics. R package version 1.1.1. (2016).

57. Saito, T. & Rehmsmeier, M. Precrec: fast and accurate precision–recall and ROC curve calculations in R. *Bioinformatics* **33**, 145–147 (2017).