Serum proteomic profiles of UK Biobank participants with Multiple Sclerosis

Benjamin M Jacobs [1,2], Nicola Vickaryous [1], Gavin Giovannoni [1, 2], Sheena Waters [1], Ruth Dobson [1,2]

Abstract

Discovery of serum protein biomarkers for Multiple Sclerosis risk and severity could shed light on pathophysiological mechanisms, suggest new drug targets, enhance the speed of diagnosis, and improve risk stratification algorithms for use in clinical practice or in the identification of high-risk participants for prevention trials. Prior proteome-wide studies of Multiple Sclerosis have yielded inconsistent results due to small sample sizes. We aimed to describe serum protein biomarkers of Multiple Sclerosis risk and to explore protein biomarkers of disease severity using radiological outcome measures. Multiple Sclerosis cases and controls were identified in UK Biobank, a longitudinal cohort study of ~500,000 British adults. Serum proteins were assayed in ~50,000 UK Biobank participants using the Olink proximity extension assay. We performed case-control association testing to examine the association between ~1300 proteins and Multiple Sclerosis, using linear models adjusted for confounding covariates. We performed a range of sensitivity analyses and validated our findings in a semi-independent testing set. We then compared results across a range of other autoimmune and neurodegenerative diseases within UK Biobank to evaluate disease specificity. Finally, associations with radiological lesion burden and brain volume were determined using a subset with available magnetic resonance imaging, using normalised T2-hyperintensity volume or whole brain volume as the outcome measure. 380 prevalent Multiple Sclerosis cases and 38,263 healthy controls were included. Cases and controls were reasonably well-matched for age, sex, ethnicity, and deprivation status. We discovered 100 proteins associated with Multiple Sclerosis at a Bonferroni-adjusted p-value of 0.05. The only proteins elevated in Multiple Sclerosis at this significance level were interleukin-15 and neurofilament light chain. 98 proteins were decreased in Multiple Sclerosis, including several integrin proteins and cystatin D. Suggestive (False Discovery Rate < 1%) positive associations were found for an additional 13 proteins, including Glial Fibrillary Acidic Protein, interleukin-17 receptor beta, and the interleukin-1 receptor accessory protein. A substantial minority of these proteins, including cystatin D, were not associated with other autoimmune or neurodegenerative disorders and so could be considered MS-specific. We found evidence ($P < 0.05$) for an association between higher serum Neurofilament Light Chain and T2 lesion load. We report the largest serum proteomic screen of Multiple Sclerosis, replicating important known associations and strengthening the evidence for the role of serum interleukin-15. While these findings require external validation, they demonstrate the power of biobank-scale datasets for discovering new biomarkers for Multiple Sclerosis.

Author affiliations

1. Preventive Neurology Unit, Wolfson Institute of Population Health, Queen Mary University of London, UK
2. Department of Neurology, Royal London Hospital, London, UK

Correspondence to:
Benjamin M Jacobs
B.jacobs@qmul.ac.uk
Preventive Neurology Unit, Wolfson Institute of Population Health, Queen Mary University of London, UK

Running title: Serum proteomic analysis of Multiple Sclerosis

Keywords: biomarkers; proteomic; interleukin-15; neurofilament; cystatin D

Introduction

Blood-based protein biomarkers have several properties which make them attractive adjuncts in the diagnosis, monitoring, and prognostication of Multiple Sclerosis (MS). Proteins are often abundant in serum, stable in solution, straightforward to quantify, and serum is easier to obtain than cerebrospinal fluid or Magnetic resonance imaging (MRI) scans, particularly if regular measurements are required for ongoing monitoring.

Serum proteins do not currently feature routinely in the investigation or management of MS[1], although serum neurofilament light chain (NFL) is increasingly entering research and clinical practice. Prior studies have shown that serum markers of neuroaxonal damage - NFL and glial fibrillary acidic protein (GFAP) - are elevated early in the MS disease course and show potential use for tracking disease activity and response to treatment[2,3]. However, neither of these markers is specific to MS[4–6]. Incorporating information from across the entire spectrum of plasma proteins - the proteome - may help to shed light on aspects of MS biology beyond neuronal loss, refine these tools for use on an individual basis, and suggest new targets for therapeutic intervention.

Previous efforts to explore the MS proteome have discovered proteins which correlate with either disease status or disease progression. However, the small sample size of these studies, differing technologies, and different protein targets have made it challenging to integrate these datasets[7–11]. A Swedish study compared the serum proteome of 52 MS cases with 17 healthy controls using the Olink assay[8], reporting higher adenosylhomocysteinase (AHCY) and corticotropin-releasing hormone (CRH) in MS, and lower fatty acid-binding protein 2 (FABP2). An earlier study using a smaller set of proteins focussed on inflammation compared 111 MS cases and 46 healthy controls in the discovery cohort and a replication dataset of 94 cases and 47 controls. In the discovery cohort, oncostatin M (OSM) and hepatocyte growth factor (HGF) were increased in MS, whereas fibroblast growth factor 21 (FGF-21), fibroblast growth factor 23 (FGF-23), and cystatin-D (CST5 / CST-D) were decreased[9]. Of these signals, only the OSM and HGF signals replicated[9]. In the Dutch GeneMSA cohort, Malekzaheh *et al.* compared 90 MS cases and 20 controls using the SomaLogic assay: the clearest difference between MS and controls was the downregulation of matrix metalloproteinase-3 (MMP3) in MS cases[7]. Studies of serum proteomic correlates of disease severity have shown consistent associations with elevated markers of neuronal damage (GFAP and NFL) and myelin oligodendrocyte glycoprotein (MOG)[2,3,12], but other markers have shown variable results between studies.

Recent advances have made it feasible to assay over a thousand different proteins in a sample simultaneously using either aptamer- or antibody-based technology. UK Biobank - a longitudinal cohort study of >500,000 British adults - has recently applied the Olink proximity extension assay technology to measure the serum level of 1,500 proteins in >50,000 of its participants, providing an invaluable resource for examining the associations between genetics,

protein levels, and health outcomes[13–15]. The prevalence of MS in UK Biobank is similar to the UK population (0.5%), and so this resource is also of significant value for MS proteomics research[16].

In this study, we analysed proteomic data from MS cases and healthy controls in UK Biobank to search for proteomic signatures associated with MS disease status and with MS severity.

Materials and methods

Cohort

UK Biobank is a longitudinal cohort study of ~500,000 British adults. Participants aged 40-69 were recruited between 2006 and 2010. A range of data is available for UKB participants, including self-reported lifestyle questionnaire data, linked electronic healthcare records, and genotyping/genome sequencing, described in detail elsewhere[17]. Serum samples for proteomics were obtained at enrolment during the baseline visit.

Case-control definitions

Disease definitions were derived from the UK Biobank 'first occurrence' fields (UKB category 1712), which combine Hospital Episode Statistics (HES), Primary Care data, mortality records, and self-reported data to define the earliest report of each ICD10 code per participant in the dataset. MS cases were defined as participants with >=1 occurrence of the ICD10 code G35, an MS diagnostic code in primary care, HES, or self-reported MS. We distinguished prevalent and incident cases based on whether the first diagnostic code occurred before or within 10 years of recruitment to allow for diagnostic and recording lag.

Using the same approach, we identified participants with neurodegenerative disorders - Parkinson's Disease (G20), Alzheimer's Disease (G30), Motor Neuron Disease (G12) - and other autoimmune disorders - Systemic Lupus Erythematosus (M32), Ankylosing Spondylitis (M45), Coeliac Disease (K90), Crohn's Disease (K50), Ulcerative Colitis (K51), Rheumatoid Arthritis (M05), and psoriasis (L40).

We excluded cases with missing values for the date of earliest diagnostic code report or with age at diagnosis <= 10 years old. For the primary analysis, we included all prevalent MS cases and excluded controls with prevalent or incident diagnoses of any listed autoimmune / neurodegenerative conditions. Counts for each disease are shown in supplementary figure 1.

For the primary analysis, we excluded samples processed as part of the COVID or pilot batches to minimise biases due to the non-random sampling of the cohort[14]. These samples include controls selected for follow-up as part of ongoing COVID studies and so do not represent a random cohort selection, hence our decision to exclude these samples from the initial analysis. In

addition, the risk of batch effects is higher with the pilot batch as there are a lower number of samples included, so these samples were also excluded, in line with previous analysis strategies employed by the UK Biobank Pharma Proteomics Project. We set aside these samples as a semi-independent validation set to attempt replication of the results from the primary analysis. The replication dataset included 41 MS cases and 6277 controls.

Proteomics data

In April 2023, UK Biobank released proteomic assay data covering ~55,000 participants[14]. The methods used to generate these data and in-house quality control have been previously described[14]. In brief, UKB performed multiplexed proteomic assays on ~55,000 serum samples obtained (for the majority) at baseline visits using the Olink platform, which exploits dual barcoded antibody technology to produce specific and semiquantitative readouts for >1500 proteins in multiplex, selected to provide coverage across a range of systems and diseases. We used proteomic data derived from blood taken at the baseline visit. From the 1463 protein assays available, we excluded proteins with >10% of values missing or >20% of values below the lower limit of detection, resulting in 1286 protein assays retained for analysis. Missing values in the remaining 1286 proteins were imputed with the mean value for the whole cohort.

Statistical analysis

Case-control comparison

To assess the relationship between serum protein levels and Multiple Sclerosis, we filtered the dataset to prevalent MS cases and healthy controls and constructed linear regression models for each protein of the form:

Protein level (NPX; normalised protein expression) ~ age + sex + age^2 + age x sex + age^2 x sex + batch + disease status (MS vs control)

The outcome for each regression model was the NPX value, a normalised value corresponding to the log fold change in protein expression between samples. Linear models were fitted using the *limma* R package, which uses empirical Bayes estimators to calibrate the per-protein variance using information from all proteins. We reported results as statistically significant if they surpassed a Bonferroni-adjusted alpha threshold of 0.05 ($P < 3.9 \times 10^{-5} = 0.05 / 1286$).

We performed the following sensitivity analyses to explore the impact of various possible confounders:

- Adjustment for BMI: Protein level (NPX; normalised protein expression) ~ age + sex + age^2 + age x sex + batch + age^2 x sex + BMI + disease status (MS vs control)
- Matched case-control analysis: we repeated the analysis in a matched case-control subsection of the cohort, matching each MS case to four healthy controls on age at

recruitment (rounded to the nearest year) and sex. We used the same model specification as for the primary analysis.

MS-specific biomarkers

To determine whether these biomarkers are specific to MS, we then performed case-control association testing between each protein biomarker and a variety of neurodegenerative and autoimmune disorders. For these models, we used the same model specification as in the primary analysis (adjusting for age, sex, batch, age^2, age x sex, and age^2 x sex). For each disease, we again compared prevalent cases with healthy controls (without neurodegenerative or autoimmune disorders) and excluded incident cases.

We then considered whether the MS-associated proteins identified were MS-specific using a stringent definition:

- The protein must be associated with MS at $P_{Bonferroni}<0.05$ and
- The protein must not show association in the same direction with **any** other tested autoimmune or neurodegenerative disease at a nominal $P<0.1$.

Proteomic correlates of MS lesion load and brain atrophy

To identify serum protein biomarkers of lesion volume in MS, we used MRI scans acquired on a subset of ~45,000 UK Biobank participants[18,19]. We used the pre-computed total volume of brain T2 hyperintensities (UK Biobank field ID 25781) derived from T2-weighted FLAIR images. To account for variation due to head size, we calculated the normalised T2 lesion volume by dividing this total volume by the total intracranial volume (the sum of the brain volume and the CSF volume). Normalised T2 lesion volumes were rank-inverse normalised before model fitting to ensure linear model assumptions were satisfied. We then tested for association between the 1286 proteins and normalised T2 lesion volume within the prevalent MS cases, using the same model specification as in the primary analysis. Using the same approach, we examined the proteomic correlates of brain atrophy in MS by assessing the association between each of 1286 proteins and the total volume of grey and white matter normalised for head size (UK Biobank field ID 25009), a cross-sectional readout of brain volume. As MRI scans were captured after the baseline visit, for both of these analyses we included the age at scan as an additional covariate in the models (i.e. we adjusted for age at recruitment, sex, age at recruitment^2, age at recruitment x sex, age at recruitment^2 x sex, batch, disease status (MS vs control), and age at scan). We confirmed there was no correlation between age at recruitment and age at scan (Pearson's correlation coefficient -0.04, P=0.6) prior to model fitting to quantify the risk of multicollinearity.

.

Gene/protein set enrichment analysis

We performed gene/protein set enrichment analysis to infer which pathways were altered in MS versus healthy controls. To do so, we used the fast gene set enrichment analysis R package (fgseaR)[20], which estimates empirical P values for enrichment using a permutation-based approach. We explored pathways for which there was a minimum of 15 proteins overlapping between our set of tested proteins and the gene/protein set, used 10,000 permutations, and stipulated a maximum gene/protein set size of 500. Proteins were ranked using the product of the sign of the correlation coefficient and the negative log10 of the P value, i.e. the proteins with the smallest P values and positive beta coefficients were ranked top, and those with the smallest P values and negative beta coefficients were ranked bottom. We report statistically significant results at a False Discovery Rate of 5%. As a reference, we used pathways from the Kyoto Encyclopaedia of Genes and Genomes (KEGG)[21].

Computing & code availability

All analyses were conducted in R version >4.2.2. This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT[22]. All code used for analysis is available at github.com/benjacobs123456/ukb_proteomics and in the supplementary R markdown document.

Data availability

UK Biobank data are available on request from https://www.ukbiobank.ac.uk/. This research was conducted under approved application 78867.

Reporting guidelines

This research was conducted in accordance with the STOBE guidelines on observational studies. The completed STROBE checklist is provided in the supplementary materials.

Results

Demographics

From the 52,705 participants with proteomic data derived from their baseline sample, we excluded 2,189 participants with missing data for >=20% of all proteins available and 7,335 participants included as part of the COVID substudy or in the pilot batch, leaving 43,181 participants for analysis (supplementary figure 2). We selected healthy controls by excluding 1317 controls with incident or prevalent neurodegenerative disorders (PD, AD, or ALS), and 3101 controls with incident or prevalent other autoimmune diseases (Coeliac disease, Ankylosing Spondylitis, Rheumatoid Arthritis, Ulcerative Colitis, Crohn’s disease, Psoriasis). We excluded participants with a diagnosis of MS coded more than ten years after enrolment (i.e. ‘incident’ cases), leaving 380 prevalent MS cases and 38263 healthy controls.

Cases and controls were reasonably well-balanced in terms of age at recruitment, sex, body mass index (BMI), Townsend deprivation scores (a proxy for socioeconomic status), and self-reported ethnicity (table 1). There were statistically-significant but quantitatively minor differences in the age at recruitment between cases and controls (median 59 vs 58 respectively), BMI (median 25.9 vs 26.8 respectively) and Townsend scores (median -2.6 vs -2.0 respectively). The median age at MS report was 43.4 years old (IQR 15.1), broadly consistent with previous estimates of the peak incidence of MS derived from electronic health records[23]. Most MS cases (301, 79.2%) had at least two separate sources of diagnostic code report (i.e. two of self-report, primary care, or hospital episode statistics).

Table 1: demographic characteristics of included participants.

	MS	Control	P[a]
N	380	38263	
Age at recruitment (median, IQR)	59 (11)	58 (13)	0.002
Female gender (n, %)	221 (58.2%)	20740 (54.2%)	0.14
Self-reported White British ethnicity (n, %)	340 (89.5%)	33241 (86.9%)	0.18
Townsend deprivation index (median, IQR)	-2.6 (4.3)	-2.0 (4.4)	0.006
Body mass index (BMI, median, IQR)	25.9 (5.8)	26.8 (5.7)	3.9x10-5
Age at MS diagnostic report (years, median, IQR)	43.4 (15.9)		
Two or more diagnostic codes for MS (n, %)	301 (79.2%)		

[a]P values indicate statistics for unpaired Wilcoxon tests (continuous variables) and Chi-squared tests (categorical variables).

Elevation of serum interleukin-15 and neurofilament light chain in Multiple Sclerosis

To search for protein biomarkers of Multiple Sclerosis, we compared the serum proteomic profiles of 380 prevalent MS cases and 38,263 healthy control participants in UK Biobank.

Of the 1286 proteins tested, 100 were associated with Multiple Sclerosis at a stringent family-wise error rate of 5% ($\square_{Bonferroni}<0.05$, $P<3.9x10^{-6}$). Seven were increased in MS, and 93 decreased in MS (figure 1, supplementary table 1). Serum samples from MS cases tended to have higher levels of the neuroaxonal damage marker Neurofilament light chain (NFL), interleukin-15 (IL15), interleukin-17 receptor beta (IL17RB), insulin-like growth factor binding protein 2 (IGFBP2), ghrelin (GHRL), cathepsin F (CTSF), and folate receptor alpha 1 (FOLR1) than controls (table 2). Several proteins were present at lower levels in MS cases, including integrin proteins (alpha-V [ITGAV], alpha-M [ITGAM], beta-2 [ITGB2], alpha-11 [ITGA11]), cystatin D (CST5), and the glycoprotein cartilage oligomeric matrix protein (COMP). Other notable proteins showing suggestive evidence (FDR < 1%) for an increase in MS included the interleukin-1 receptor subunit (IL1RAP), glial fibrillary acidic protein (GFAP), fractalkine/C-

X3-C chemokine ligand 1 (CX3CL1), and complement component 4 binding protein beta (C4BPB) (supplementary table 1).

Table 2: elevated serum proteins in Multiple Sclerosis.

Protein	Beta	Unadjusted P	FDR	Alpha
NEFL[a]	0.169772	6.73E-10	4.30E-08	8.66E-07
IGFBP2	0.201984	1.13E-06	2.50E-05	0.001452
GHRL	0.234808	2.81E-06	5.23E-05	0.003608
IL15[a]	0.082278	7.17E-06	1.21E-04	0.00922
CTSF	0.106203	7.30E-06	1.22E-04	0.009387
FOLR1	*0.07295*	*1.24E-05*	*1.82E-04*	*0.015977*
IL17RB	*0.136411*	*1.43E-05*	*2.04E-04*	*0.018384*
GFAP	0.096535	4.39E-05	5.43E-04	0.056465
ALPP	0.278381	1.01E-04	0.001088	0.129622
IL1RAP	0.099911	1.62E-04	0.00164	0.2083
NPTX1	0.083349	5.05E-04	0.00414	0.649969
MET	0.039593	7.81E-04	0.006084	1
CX3CL1	0.0618	9.85E-04	0.007448	1
CGA	0.217257	9.90E-04	7.45E-03	1
C4BPB	0.076231	1.02E-03	0.007594	1

Serum proteins found to be elevated in Multiple Sclerosis cases versus healthy controls. Results are shown for proteins achieving statistical significance at a False Discovery Rate of <1%. The columns show (in order) the protein symbol, the beta coefficient from linear regression (where the positive values indicate a higher mean protein abundance in Multiple Sclerosis), the raw unadjusted P value, the Benjamini-Hochberg False Discovery Rate, and the Bonferroni-adjusted P value (i.e. P divided by the number of tests, 1286). Italicised proteins are those that achieved a stringent Bonferroni alpha threshold of 5%.

[a]Denotes proteins which retained a strong (alpha < 5%) association with MS after adjustment for BMI.

Given the profound effects of BMI on the serum proteome[14], and the difference in BMI between MS cases and controls, we performed sensitivity analyses adjusting for BMI. Most proteins (70/100, 70%) remained strongly associated with MS on adjustment for BMI (□ < 5%, supplementary table 1). Five of the seven proteins identified as elevated in MS showed a weaker association once BMI was considered: only serum NFL and IL-15 remained strongly associated (□ < 5%). To account for possible confounding due to differences in age and sex, we repeated the analysis in a nested matched case-control cohort (380 MS and 1520 age and sex-matched controls). Despite the lower statistical power due to the smaller number of controls, we observed concordant effects for both NFL and IL15 (supplementary table 1). We observed similar concordant results in the validation set (41 MS, 6277 controls, supplementary table 1).

Gene set enrichment analysis reaffirmed the MS-associated decrease in proteins involved in leukocyte migration, interaction with the extracellular matrix, regulation of the actin

cytoskeleton, and cell-cell adhesion (KEGG terms showing negative normalised enrichment scores [NES] at FDR < 5%, figure 1). We also observed suggestive evidence for an increase in proteins involved in JAK-STAT signalling (NES 1.4, nominal P = 0.07), although this result did not surpass the multiple testing threshold (FDR >5%, figure 1). This pathway-level enrichment was primarily driven by cytokine and cytokine receptor proteins IL-15, IL-10, and IL-2RA (supplementary figures).

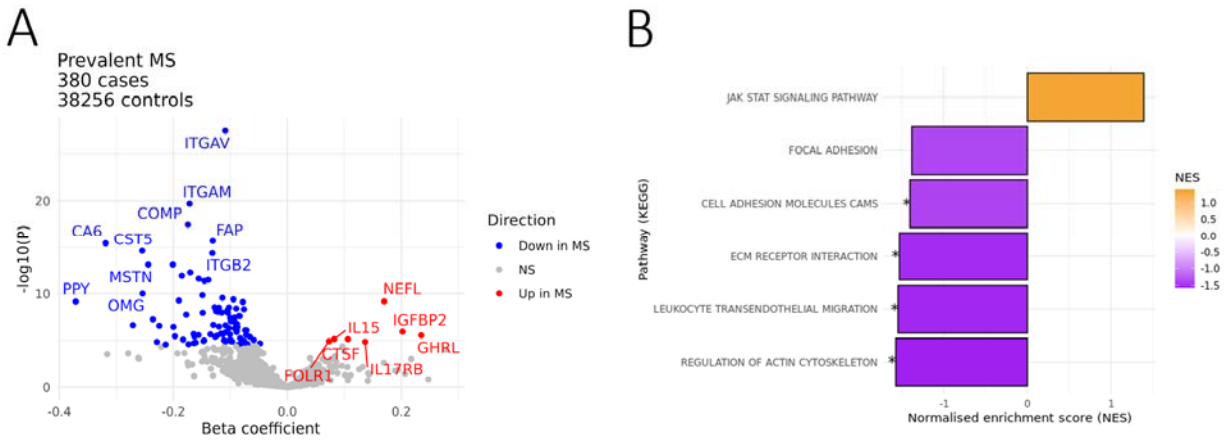

Figure 1: serum proteomic analysis of Multiple Sclerosis.

Comparison with other autoimmune and neurodegenerative disorders

To determine whether any of the 100 protein biomarkers identified were specific to MS, we first explored the association between each of the 1286 serum protein biomarkers and a range of autoimmune and neurodegenerative disorders. We reproduced several well-known associations (supplementary figure 3), such as elevated beta-defensin in psoriasis[24], interleukin-6 in ankylosing spondylitis[25], TNF-alpha in rheumatoid arthritis[26], IL-15 in SLE[27], and neurofilament light chain in MND[28], PD[5], and AD[4]. The proteomic signature of MS showed weak evidence of correlation with other autoimmune diseases - the strongest effect was a weak positive correlation with SLE (R = 0.44, p = 0.1, supplementary figure 4). In contrast, we observed stronger correlations between autoimmune disorders with overlapping aetiological pathways, such as Crohn's and UC. While this may be a consequence of lower power in the MS cohort and so greater measurement error, these data support the concept that the immune system signalling pathways active in the peripheral blood of MS differ from other autoimmune disorders.

Most MS-associated biomarkers were not specific to MS: 50/67 (75%) showed concordant effects in at least one other disorder. As expected, we found that elevation of NFL was also associated with neurodegenerative disorders, with strong effects seen for Parkinson's and Motor Neuron Disease (supplementary figure 3). Increased IL-15 was observed in other autoimmune disorders, including SLE and Crohn's disease. Seventeen MS-associated protein changes

appeared specific to the MS cohort (figure 2). All of these proteins were decreased in MS, with no statistical evidence (i.e. $P < 0.1$) for a decrease in any other disorder tested.

The clearest effect was the reduction in cystatin D (CST5), a serine protease inhibitor with tumour-suppressive action. Cystatin D is induced by serum vitamin D[29], and has been previously reported to be present at lower levels in MS serum, although this result did not replicate in a validation cohort[9]. Given the association between low serum vitamin D and MS, we considered whether the lower levels of cystatin D observed in this MS cohort could be explained by lower serum vitamin D levels. There was no evidence for a difference in serum vitamin D (25-OH-D3) between cases and controls nor for a correlation between serum vitamin D and cystatin D in the whole cohort, including cases and controls (Spearman's rho 0.007, P = 0.18, supplementary figure 5).

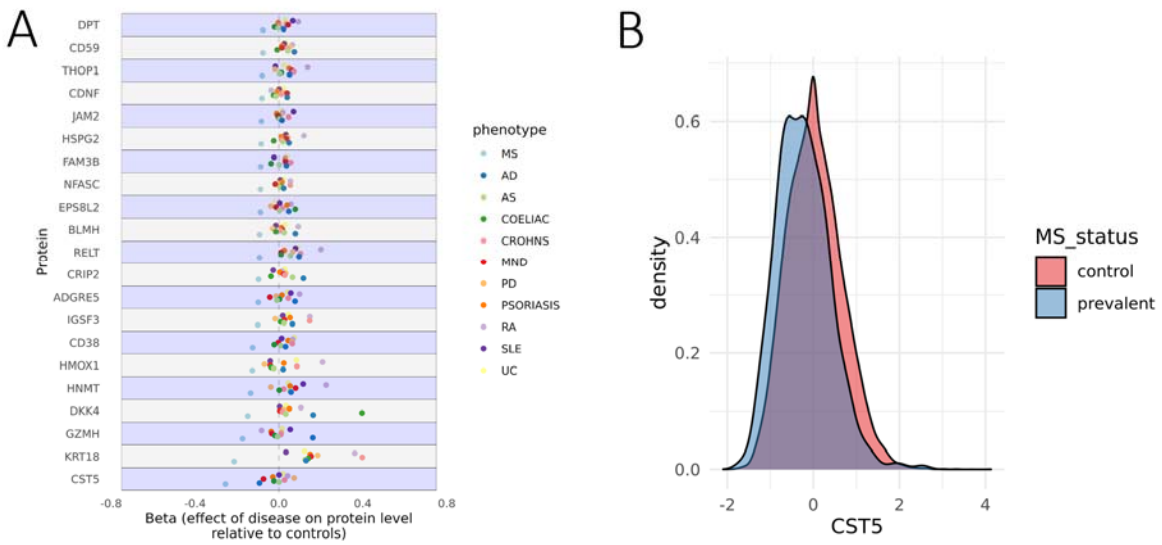

Figure 2: Decrease in serum cystatin D is specific to Multiple Sclerosis.

Proteomic correlates of MS severity

Next, we explored the relationship between the serum proteome and T2 lesion volume using MRI data available for a subset of the cohort. We first confirmed the expected increase in T2 lesion volume in MS cases compared with healthy controls (N=105 MS, N=3627 controls, median 10,995 mm^3 vs 2893mm^3, $P < 2x10^{-16}$, figure 3). There were no proteins associated with T2 lesion volume at a stringent alpha of 5%, however, we found suggestive evidence ($P < 0.05$) for a positive correlation between T2 lesion load and serum NFL, myelin oligodendrocyte glycoprotein (MOG), and SLAMF7 (supplementary table 3).

We also confirmed the expected decrease in total brain volume in MS cases compared with healthy controls (N=105 MS, N=3627 controls, median 1,465.1 cm^3 vs 1,487.7 cm^3, $P < 0.002$,

figure 3). Higher serum levels of osteomodulin (OMD) were associated with lower brain volumes ($P = 9.6 \times 10^{-6}$, $\alpha < 5\%$). Although no other proteins were associated with brain volume at $\alpha < 5\%$, we observed directionally concordant effects for NFL ($P = 0.11$) and SLAMF7 ($P=0.004$), i.e. higher levels of both proteins were associated with lower brain volume. Given the primary role of osteomodulin as a regulator of bone mineralisation[30] and the lack of prior data supporting an association between serum osteomodulin and brain health, we suspect this relationship may reflect a higher degree of frailty and poor bone health among people with more severe MS. Serum vitamin D may act as a confounder in this relationship - however controlling for baseline serum vitamin D did not impact the association, nor did adjustment for BMI measured at the time of the scan.

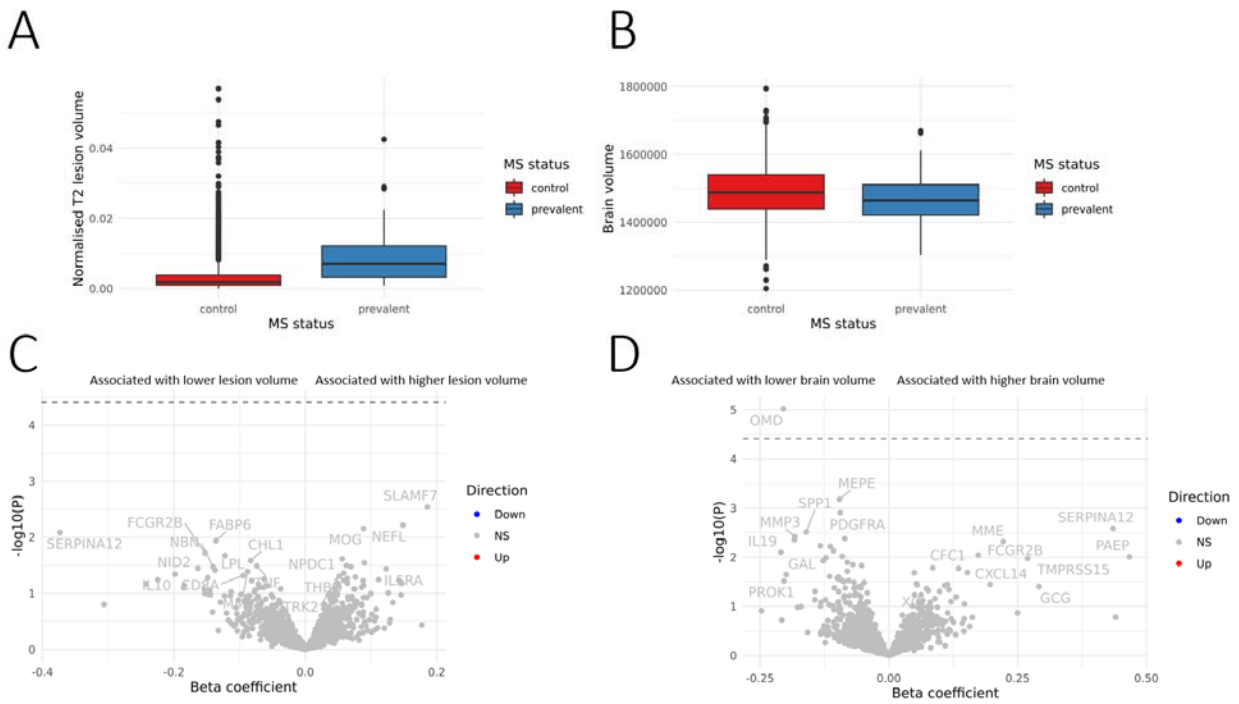

Figure 3: serum proteomic correlates of MRI proxies for MS severity.

Discussion

Using data from ~50,000 UK Biobank participants, we describe alterations in the serum proteome of patients with Multiple Sclerosis. We report that MS patients have higher serum levels of interleukin 15 and neurofilament light chain and lower levels of several proteins, including integrins and cystatin D, compared with healthy controls. While most of these alterations are also seen in other autoimmune or neurodegenerative disorders, we find that the decrease in seventeen of these proteins - including cystatin D - is relatively specific to MS. Finally, using MRI scans from a subset of the cohort, we report that T2 hyperintense lesion burden shows suggestive associations with serum NFL, MOG, and SLAMF7, and the brain volume is associated with osteomodulin.

This is the first study to evaluate MS-associated changes in the serum proteome using biobank-scale data. The key strengths of our study are, firstly, the very large number of healthy controls, minimising the risk of false positive findings, and the broad coverage of the proteome afforded by the Olink platform.

Interleukin-15 (IL-15) is a cytokine in the IL-2 family which plays pleiotropic roles in the governance of T cell homeostasis and has been implicated in a range of autoimmune disorders and cancers[31]. Raised serum IL-15 is a common feature of autoimmune disorders[31], an observation we replicate. Elevation of both serum and cerebrospinal fluid interleukin-15 (IL-15) has been previously reported in smaller studies of MS[32]. IL-15 is a cytokine produced by multiple cell types, including B lymphocytes, astrocytes, and monocytes which promotes the proliferation of CD4+ T cells, polarisation towards cytotoxic phenotypes (e.g. IL-17 production), is capable of signalling via the STAT5 pathway, and may promote chemotaxis into the inflamed Central Nervous System[33–36]. Although there are minimal trial data on therapeutic antagonism of the IL-15 signalling pathway in autoimmune disease, our results support the hypothesis that this approach could be of value in MS.

Serum NFL elevation in MS is a well-established biomarker of neuroaxonal damage, which can be detected early in the course of MS, and maybe a useful biomarker for tracking disease activity and progression over time[2]. The fact that serum NFL was the protein most strongly and consistently associated with MS in our study across a range of sensitivity analyses is a reassuring positive control[15].

We detected a large number of proteins which were present at lower levels in serum from MS patients and in patients with other autoimmune or neurodegenerative disorders, some of which appeared to be specifically downregulated in MS. Cystatin D (CST5) - the strongest MS-specific signal - was previously reported as downregulated in MS patients from a Swedish cohort, but did not replicate in the discovery cohort[9]. Somewhat paradoxically, cystatin D was also associated with relapse in the Swedish cohort and was reported as an early marker of traumatic brain injury in another study[37]. Cystatin D expression is positively regulated by serum vitamin D in cancer cells, but in our cross-sectional study, lower serum vitamin D did not explain the observed decrease of CST5 in MS patients[29]. Other MS-specific decreases in serum proteins were observed for the myeloma marker and plasma cell surface antigen CD38 and the wnt signalling inhibitor dickkopf-4 (DKK4). These results require external validation, as they could plausibly represent statistical noise.

Despite the low statistical power to detect associations between serum proteins and radiological lesion burden, we also detect suggestive associations between serum NFL, MOG, and surface antigen CD319 (SLAMF7), an MS susceptibility gene[38]. The positive control of NFL reinforces

the validity of this approach and suggests that novel targets could be discovered with an increase in sample size. The strong association we observe between brain volume and osteomodulin is somewhat surprising given the lack of prior data to support a role for osteomodulin in either MS or in maintaining brain health[30]. While it is plausible that this merely reflects statistical noise due to the small sample size, we interpret this association as an indicator of poor bone health and overall frailty among people with more severe MS – in support of this hypothesis, a recent longitudinal study described elevation of serum osteopontin, another regulator of bone mineralisation, as a predictor of long-term brain atrophy in MS[12]. Although adjustment for BMI or baseline serum vitamin D could not explain this signal, these metrics are imperfect proxies for frailty.

There are some important limitations to this study. First, although we successfully replicate most MS-associated signals in a separate validation set of samples, these samples are not truly independent, and external replication is required to be certain of the novel findings we describe. Second, as UK Biobank is an all-purpose dataset not specifically designed for MS, we are unable to explore MS phenotypes in granular detail - for instance, we are unable to explore the relationships between the proteome and EDSS, relapse rate, treatment responsiveness, or more MS-specific radiological outcomes such as new/expanding lesion load. There is also some risk of misclassification of cases and controls inherent in using electronic healthcare records and self-reported diagnoses rather than clinically-definite MS diagnosis, although we have previously shown that this cohort is demographically similar to clinical MS cohorts[16]. Third, as this study is cross-sectional and includes participants with established MS, it is challenging to deconvolve the impact of treatment, disease duration, disease stage at the time of sampling, and other time-varying parameters. We mitigate this to some extent by accounting for a variety of important confounders, such as age, sex, BMI, and deprivation status, but the serum proteome is likely highly dynamic, and a cross-sectional snapshot is a gross simplification of this complexity. Fourth, it is now established that participants in UKB are a highly selected population, and so it is plausible that the MS cohort we analysed here represents an unusually ‘healthy’ selection of MS patients[39]. This phenomenon could explain why we see lower levels of CST5 - a marker previously associated with more severe MS - in the cases versus healthy controls. Finally, while the Olink assays give excellent coverage across the proteome, it still only measures a finite set of proteins and may miss a variety of interesting proteins either not present on the panel or which have undergone post-translational modifications, in contrast to the relatively unbiased nature of mass spectrometry.

In summary, we perform the largest analysis of the serum proteome in MS, replicating known biomarkers for diseases such as NFL and GFAP, strengthening the evidence for the role of interleukin-15 in MS, and discovering several new disease-specific negative biomarkers, such as cystatin D. While these findings require replication, they demonstrate the power of biobank-scale datasets for biomarker discovery, even for relatively rare diseases such as Multiple Sclerosis.

Funding

BMJ is funded by a Medical Research Council (MRC) Clinical Research Training Fellowship (CRTF) jointly supported by the UK MS Society (BMJ; grant reference MR/V028766/1), and by Barts Charity. The Preventive Neurology Unit is supported by a grant from Barts Charity.

Competing interests

The authors report no relevant competing interests.

Supplementary material

Supplementary material is available at *Brain* online.

References

1. Thompson, A. J. *et al.* Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol.* **17**, 162–173 (2018).
2. Bjornevik, K. *et al.* Serum Neurofilament Light Chain Levels in Patients With Presymptomatic Multiple Sclerosis. *JAMA Neurol.* **77**, 58–64 (2020).
3. Meier, S. *et al.* Serum Glial Fibrillary Acidic Protein Compared With Neurofilament Light Chain as a Biomarker for Disease Progression in Multiple Sclerosis. *JAMA Neurol.* **80**, 287–297 (2023).
4. Mattsson, N., Andreasson, U., Zetterberg, H., Blennow, K. & Alzheimer's Disease Neuroimaging Initiative. Association of Plasma Neurofilament Light With Neurodegeneration in Patients With Alzheimer Disease. *JAMA Neurol.* **74**, 557–566 (2017).
5. Huang, Y., Huang, C., Zhang, Q., Shen, T. & Sun, J. Serum NFL discriminates Parkinson disease from essential tremor and reflect motor and cognition severity. *BMC Neurol.* **22**, 39 (2022).

6. Shir, D. *et al.* Association of plasma glial fibrillary acidic protein (GFAP) with neuroimaging of Alzheimer's disease and vascular pathology. *Alzheimers. Dement.* **14**, e12291 (2022).

7. Malekzadeh, A. *et al.* Plasma proteome in multiple sclerosis disease progression. *Ann Clin Transl Neurol* **6**, 1582–1594 (2019).

8. Huang, J., Khademi, M., Piehl, F., Olsson, T. & Kockum, I. Plasma protein profiling of multiple sclerosis using proximity extension assays. *bioRxiv* (2022) doi:10.1101/2022.07.29.22278211.

9. Huang, J. *et al.* Inflammation-related plasma and CSF biomarkers for multiple sclerosis. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 12952–12960 (2020).

10. Tremlett, H. *et al.* Serum proteomics in multiple sclerosis disease progression. *J. Proteomics* **118**, 2–11 (2015).

11. Wallin, M. T. *et al.* Serum proteomic analysis of a pre-symptomatic multiple sclerosis cohort. *Eur. J. Neurol.* **22**, 591–599 (2015).

12. Jakimovski, D. *et al.* Proteomics and relationship with axonal pathology in multiple sclerosis: 5-year diffusion tensor imaging study. *Brain Commun* **5**, fcad183 (2023).

13. Petrera, A. *et al.* Multiplatform Approach for Plasma Proteomics: Complementarity of Olink Proximity Extension Assay Technology to Mass Spectrometry-Based Protein Profiling. *J. Proteome Res.* **20**, 751–762 (2021).

14. Sun, B. B. *et al.* Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *bioRxiv* 2022.06.17.496443 (2022) doi:10.1101/2022.06.17.496443.

15. Gadd, D. A. *et al.* Blood protein levels predict leading incident diseases and mortality in UK Biobank. *medRxiv* (2023) doi:10.1101/2023.05.01.23288879.

16. Jacobs, B. M. *et al.* Gene-Environment Interactions in Multiple Sclerosis: A UK Biobank Study. *Neurol Neuroimmunol Neuroinflamm* **8**, (2021).

17. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

18. Alfaro-Almagro, F. *et al.* Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* **166**, 400–424 (2018).

19. Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* **19**, 1523–1536 (2016).

20. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2021) doi:10.1101/060012.

21. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457-62 (2016).

22. King, T., Butcher, S. & Zalewski, L. *Apocrita - High Performance Computing Cluster for Queen Mary University of London*. (2017). doi:10.5281/zenodo.438045.

23. Mackenzie, I. S., Morant, S. V., Bloomfield, G. A., MacDonald, T. M. & O'Riordan, J. Incidence and prevalence of multiple sclerosis in the UK 1990-2010: a descriptive study in the General Practice Research Database. *J. Neurol. Neurosurg. Psychiatry* **85**, 76–84 (2014).

24. Cardner, M. *et al.* Analysis of serum proteomics data identifies a quantitative association between beta-defensin 2 at baseline and clinical response to IL-17 blockade in psoriatic arthritis. *RMD Open* **9**, (2023).

25. Du, J. *et al.* Serum IL-6 and TNF-α Levels Are Correlated with Disease Severity in Patients with Ankylosing Spondylitis. *Lab. Med.* **53**, 149–155 (2022).

26. Edrees, A. F., Misra, S. N. & Abdou, N. I. Anti-tumor necrosis factor (TNF) therapy in rheumatoid arthritis: correlation of TNF-alpha serum level with clinical response and benefit from changing dose or frequency of infliximab infusions. *Clin. Exp. Rheumatol.* **23**, 469–474 (2005).

27. Zhang, T. *et al*. Excessive IL-15 promotes cytotoxic CD4 + CD28− T cell-mediated renal injury in lupus nephritis. *Immun. Ageing* **19**, 1–12 (2022).

28. Davies, J. C. *et al*. Limited value of serum neurofilament light chain in diagnosing amyotrophic lateral sclerosis. *Brain Commun* **5**, fcad163 (2023).

29. Alvarez-Díaz, S. *et al*. Cystatin D is a candidate tumor suppressor gene induced by vitamin D in human colon cancer cells. *J. Clin. Invest.* **119**, 2343–2358 (2009).

30. Lin, W. *et al*. Osteomodulin positively regulates osteogenesis through interaction with BMP2. *Cell Death Dis.* **12**, 147 (2021).

31. Waldmann, T. A., Miljkovic, M. D. & Conlon, K. C. Interleukin-15 (dys)regulation of lymphoid homeostasis: Implications for therapy of autoimmunity and cancer. *J. Exp. Med.* **217**, (2020).

32. Vaknin-Dembinsky, A., Brass, S. D., Gandhi, R. & Weiner, H. L. Membrane bound IL-15 is increased on CD14 monocytes in early stages of MS. *J. Neuroimmunol.* **195**, 135–139 (2008).

33. Pashenkov, Mustafa, Kivisäkk & Link. Levels of interleukin-15-expressing blood mononuclear cells are elevated in multiple sclerosis. *Scand. J. Immunol.* **50**, 302–308 (1999).

34. Saikali, P., Antel, J. P., Pittet, C. L., Newcombe, J. & Arbour, N. Contribution of astrocyte-derived IL-15 to CD8 T cell effector functions in multiple sclerosis. *J. Immunol.* **185**, 5693–5703 (2010).

35. Broux, B. *et al.* IL-15 amplifies the pathogenic properties of CD4+CD28- T cells in multiple sclerosis. *J. Immunol.* **194**, 2099–2109 (2015).

36. Schneider, R. *et al.* B cell-derived IL-15 enhances CD8 T cell cytotoxicity and is increased in multiple sclerosis patients. *J. Immunol.* **187**, 4119–4128 (2011).

37. Hill, L. J. *et al.* Author Correction: Cystatin D (CST5): An ultra-early inflammatory biomarker of traumatic brain injury. *Sci. Rep.* **8**, 4572 (2018).

38. O'Connell, P., Blake, M. K., Godbehere, S., Amalfitano, A. & Aldhamen, Y. A. SLAMF7 modulates B cells and adaptive immunity to regulate susceptibility to CNS autoimmunity. *J. Neuroinflammation* **19**, 241 (2022).

39. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).

Figure legends

Figure 1: serum proteomic analysis of Multiple Sclerosis. A - volcano plot displaying differences in serum levels of proteins measured on the Olink proteomics assay between UK Biobank participants with (n=380) and without (n=38256) MS at the time of sampling. Note that seven controls were excluded due to missing covariate data. The x-axis indicates the log-fold change of the protein, with values above 0 indicating proteins present at higher levels in the MS cohort and those with values below 0 present at lower levels in the control cohort. The y-axis indicates the negative log of the P value for each protein, with higher values indicating a more statistically significant result. Proteins surpassing a Bonferroni-corrected threshold of 5% are shown in colour; other results are shown in grey. B - gene set enrichment analysis (GSEA) results summarising the pathway-level differences in the serum proteome between MS and

healthy controls. The x-axis shows the normalised enrichment score, with positive values equivalent to an increase in the pathway activity in MS and negative values suggesting decreased levels of pathway activity in MS. The y-axis shows the KEGG pathways with nominal evidence ($P < 0.05$) for alteration in MS. The asterisks indicate statistically significant results at an FDR of 5%.

Figure 2: Decrease in serum cystatin D is specific to Multiple Sclerosis. A - forest plot showing seventeen 'MS-specific' protein biomarkers. The x-axis shows the direction of effect (beta coefficient) for the association of each protein with each disease outcome. The y-axis shows the protein tested. Each point is coloured by disease. MS-specific effects were defined as statistically-significant effects (Bonferroni's alpha < 5%) in the MS cohort and did not show suggestive evidence ($P<0.1$) of the same direction of effect in any other disease tested. B - density plots showing the distribution of cystatin D serum levels in healthy controls and MS cases.

Figure 3: serum proteomic correlates of MRI proxies for MS severity. A & B - boxplots showing the relationship between MS status and normalised T2 lesion volume load (A) or total brain volume (B) in the subset of 105 MS cases and 3627 controls with MRI scans available. Normalised lesion load is shown as a proportion of total brain volume. Brain volume is shown in mm^3. C & D - volcano plots demonstrating the association between levels of serum proteins and normalised T2 lesion volume (C) or brain volume (D) in the MS cohort. The dashed line indicates the Bonferroni threshold of 5%.