Proteomic prediction of diverse incident diseases: a machine learning-guided biomarker discovery study using data from a prospective cohort study

Julia Carrasco-Zanini, Maik Pietzner*, Mine Koprulu, Eleanor Wheeler, Nicola D Kerrison, Nicholas J Wareham, Claudia Langenberg*

Summary

Background Broad-capture proteomic technologies have the potential to improve disease prediction, enabling targeted prevention and management, but studies have so far been limited to very few selected diseases and have not evaluated predictive performance across multiple conditions. We aimed to evaluate the potential of serum proteins to improve risk prediction over and above health-derived information and polygenic risk scores across a diverse set of 24 outcomes.

Methods We designed multiple case-cohorts nested in the EPIC-Norfolk prospective study, from participants with available serum samples and genome-wide genotype data, with more than 32 974 person-years of follow-up. Participants were middle-aged individuals (aged 40–79 years at baseline) of European ancestry who were recruited from the general population of Norfolk, England, between March, 1993 and December, 1997. We selected participants who developed one of ten less common diseases within 10 years of follow-up; we also subsampled a randomly drawn control subcohort, which also served to investigate 14 more common outcomes (n>70), including all-cause premature mortality (death before the age of 75 years; case numbers 71–437; controls 608–1556). Individuals were excluded from the current study owing to failed genotyping or proteomic quality control, relatedness, or missing information on age, sex, BMI, or smoking status. We used a machine learning framework to derive sparse predictive protein models for the onset of the the 23 individual diseases and all-cause premature mortality, and to derive a single common sparse multimorbidity signature that was predictive across multiple diseases from 2923 serum proteins.

Findings Participants who developed one of ten less common diseases within 10 years of follow-up included 482 women and 507 men, with a mean age at baseline of 64·56 years (8·08). The random subcohort included 990 women and 769 men, with a mean age of 58·79 years (9·31). As few as five proteins alone outperformed polygenic risk scores for 17 of 23 outcomes (median dfference in concordance index [C-index] 0·13 [0·10–0·17]) and improved predictive performance when added over basic patient-derived information models for seven outcomes, achieving a median C-index of 0·82 (IQR 0·77–0·82). This included diseases with poor prognosis such as lung cancer (C-index 0·85 [+/– cross-validation error 0·83–0·87]), for which we identified unreported biomarkers such as C-X-C motif chemokine ligand 17. A sparse multimorbidity signature of ten proteins improved prediction across seven outcomes over patient-derived information models, achieving performances (median C-index 0·81 [IQR 0·80–0·82]) similar to those of disease-specific signatures.

Interpretation We show the value of broad-capture proteomic biomarker discovery studies across multiple diseases of diverse causes, pointing to those that might benefit the most from proteomic approaches, and the potential to derive common sparse biomarker panels for prediction of multiple diseases at once. This framework could enable follow-up studies to explore the generalisability of proteomic models and to benchmark these against clinical assays, which are required to understand the translational potential of these findings.

Funding Medical Research Council, Health Data Research UK, UK Research and Innovation–National Institute for Health and Care Research, Cancer Research UK, and Wellcome Trust.

Lancet Digit Health 2024; 6: e470–79

*Authors contributed equally

MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge, UK (J Carrasco-Zanini PhD, Prof M Pietzner PhD, M Koprulu MPhil, E Wheeler PhD, N D Kerrison MSci, Prof N J Wareham FMedSci, Prof C Langenberg FFPH); **Computational Medicine, Berlin Institute of Health at Charité-Universitätsmedizin Berlin, Berlin, Germany** (Prof M Pietzner, Prof C Langenberg); **Precision Healthcare University Research Institute, Queen Mary University of London, London, UK** (J Carrasco-Zanini, Prof M Pietzner, Prof C Langenberg)

Correspondence to: Prof Claudia Langenberg, Computational Medicine, Berlin Institute of Health at Charité-Universitätsmedizin Berlin, 10117 Berlin, Germany **claudia.langenberg@qmul.ac.uk**

For more on **EPIC-Norfolk** see https://www.epic-norfolk.org.uk/

Introduction

Blood-based omics have the potential to improve our ability to predict the onset and course of diseases,[1] but systematic and rigorous testing at scale is often lacking. Multiple efforts have successfully developed and validated genetic and polygenic predictors for multiple diseases,[2] but translation for clinical use has been challenging, partly because of gaps in knowledge related to the potential improvement in prediction over and above easily measured clinical parameters.[3] In contrast to our inherited genome, our proteome, which is the central layer of information transfer, changes in response to early disease mechanisms. Characteristics of circulating proteins—such as the wide dynamic range, longer half-life, and mostly endogenous origin compared with other blood-based biomarkers such as metabolites—and easy

Research in context

Evidence before this study

We searched PubMed for omics prediction studies across multiple diseases from inception up to March 20, 2024, using the search terms "proteomic prediction","protein risk scores","incident diseases", "pan-disease", "multimorbidity", and "diverse diseases". Studies have been limited to prediction of selected individual diseases, mostly by the use of polygenic risk scores, and only a few studies have shown protein signatures that improve prediction of onset for individual (mostly cardiovascular) diseases. A few pioneer studies have explored prediction across multiple prevalent indicators of health, or across incident diseases leveraging metabolomic and proteomic profiles, but these have relied on complex omics signatures including hundreds of biomarkers, reducing the translational potential of the findings.

Added value of this study

Our study takes a comprehensive approach, integrating proteomic data, genomic data, and data from electronic health records to systematically derive sparse protein signatures for prediction across 23 diverse incident diseases and all-cause mortality. Our results show that as few as five proteins outperformed polygenic risk scores for the majority of outcomes, and improved the prediction of seven outcomes over common risk factors. We further developed a sparse multimorbidity signature of ten proteins, which improved the prediction of individual diseases over common risk factors.

Implications of all the available evidence

The increasing availability of a high-throughput proteomics platform promises to improve biomarker discovery and prediction strategies; however, evaluation across different diseases has not been done systematically. Our study highlights the potential of broad-capture proteomics for the development of sparse signatures to improve prediction strategies, including common panels of biomarkers for the prediction of multiple diseases, and provides a guide for future studies on disease causes that might benefit the most from proteomic, genomic, or combined approaches.

accessibility make them attractive for the prediction, diagnosis, and prognosis of different diseases, with many examples established and in clinical use.[4] However, even for diseases with well-established biomarkers in clinical use, these proteins were derived from targeted studies, and systematic comparisons against or in combination with other omics biomarkers identified through broad-capture and hypothesis-free studies are lacking.

Affinity-based technologies can now capture thousands of proteins in a single experiment across the entire range of abundance in blood. Although they are expensive compared with mass spectrometry-based technologies that capture fewer proteins of higher abundance,[5] a large proportion of disease-relevant proteins, such as those involved in signalling or reflecting tissue damage, are expected to be found in the lower end of the spectrum of abundance.[6] Deep mass spectrometry-based workflows are only now emerging,[7,8] but are still challenging to implement at population scale.

Machine learning approaches have enabled systematic, data-driven investigation of broad-coverage proteomic platforms to identify novel biomarkers,[1] introducing the possibility to assess specificity or sharedness across diseases. Despite the large potential of these new proteomic technologies to improve prediction and prognosis of non-communicable and infectious diseases,[1,9] the absence of prospective studies investigating different diseases with linkage to electronic health records has so far limited progress in testing and translating the utility of these technologies.

See Online for appendix

Here, we integrate serum proteomic data with genomic data, hospital admission records, and cancer registry data to systematically and prospectively evaluate the potential of serum proteins to improve risk prediction over and above health-derived information and polygenic risk scores (PRSs) across a diverse set of 23 non-communicable diseases and all-cause premature mortality.

Methods

Study design and participants

The European Prospective Investigation into Cancer-Norfolk (EPIC-Norfolk) is a cohort study of 25 639 middle-aged individuals (40–79 years at baseline) recruited from the general population of Norfolk, a county in the east of England, between March, 1993, and December, 1997.[10] The study was approved by the Norfolk Research Ethics Committee (reference 05/Q0101/191) and all participants provided written informed consent. We designed multiple case-cohort studies, all nested within the EPIC-Norfolk study, among participants with available serum samples and genome-wide genotype data. Participants were of European ancestry. Individuals were excluded owing to failed genotyping or proteomic quality control, relatedness, or missing information on age, sex, BMI, or smoking status.

Outcomes

We studied the onset of 24 outcomes (appendix pp 13–15) from diverse clinical specialties. To enable prospective investigation of both less and more frequently occurring non-communicable diseases, we selected participants (n=989) who developed one of ten less common diseases within 10 years of follow-up: lung cancer, haemorrhagic stroke, Parkinson's disease, colon cancer, breast cancer, venous thrombosis, type 2 diabetes, vascular dementia, Alzheimer's disease, and acute pancreatitis. We

subsampled a randomly drawn control subcohort (n=1759), which also facilitated the investigation of 14 more common outcomes that occurred frequently (n>70) within this subcohort: renal disease, atrial fibrillation, heart failure, ischaemic heart disease, cerebral stroke, peripheral arterial disease, asthma, chronic obstructive pulmonary disease (COPD), non-melanoma skin cancer, prostate cancer, fractures, cataracts, glaucoma, and all-cause premature mortality (death before the age of 75 years; appendix pp 7, 13–16; figure 1). For each disease analysed, we excluded individuals who reported the disease at baseline or in whom the disease occurred within the first 6 months of follow-up. All other individuals from the randomly drawn subcohort who did not develop the disease of interest (including those individuals with other prevalent diseases) were defined as non-incident cases for analyses.

Exposures

Serum samples from the baseline assessment (1993–97) that had been stored in liquid nitrogen were used for proteomic profiling in two separate batches ($n_{set\ 1}$=1040, $n_{set\ 2}$=1708) using the Olink Explore 1536 and Olink Explore Expansion panels, targeting 2923 unique proteins by 2941 assays (appendix p 7). Assay details are described in the appendix (pp 2–3) and have been described in

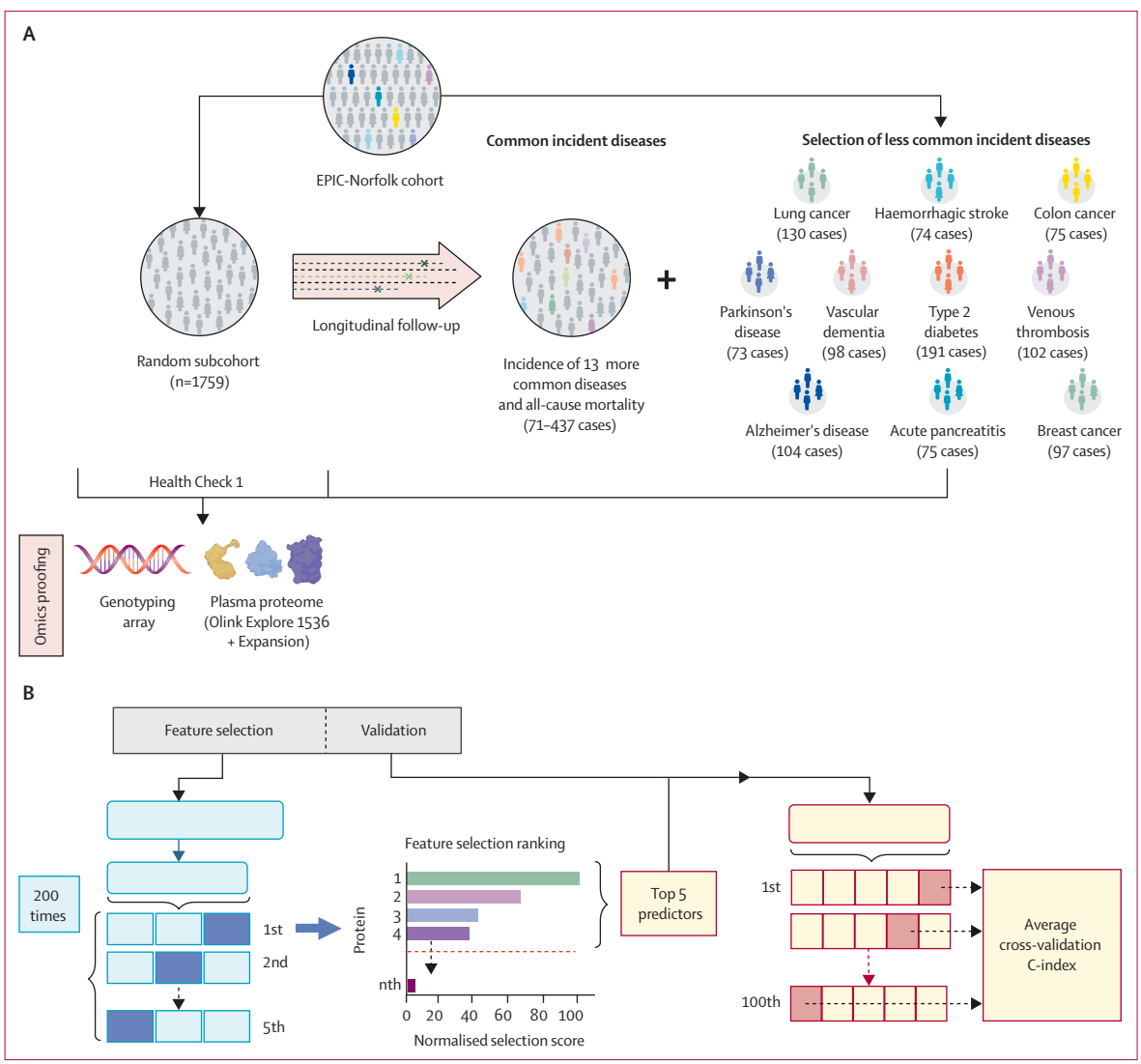

Figure 1: **Study design**

(A) Participants. We selected a random subcohort (n=1759) from the EPIC-Norfolk prospective cohort study to investigate the incidence of 13 common diseases and all-cause mortality. For ten less common diseases among the general population, we selected incident cases within 10 years of follow-up. Proteomic profiling was done in samples from the baseline visit that had been stored in liquid nitrogen. (B) General machine learning framework. We did feature selection by LASSO regression over bootstrap resampling or subsampling of the feature selection set to assign a selection score to each of the proteins. We took the top five predictor proteins for optimisation by five-fold cross-validation in a separate optimisation set, and we report average cross-validation performance metrics (C-index) over 100 iterations. Figure created with BioRender.com.

detail elsewhere.[11] After exclusions owing to quality control, each batch consisted of a randomly selected control subcohort ($n_{control\_1}$=749, $n_{control\_2}$=1010) and multiple case cohorts (n_{cases_1}=291, n_{cases_2}=698) to enrich for less common diseases (appendix pp 3, 16). Owing to the lack of bridge normalisation between batches, and the high concordance of participant characteristics between batches (appendix p 17), we excluded protein assays for which the difference in coefficients of variation between each batch in the control subcohort was greater than 0·5. These differences probably indicate protein targets with high measurement variation (appendix p 18), leaving a total of 2319 protein assays for downstream analyses.

Genome-wide genotyping was done by use of the Affymetrix UK Biobank Axiom Array (Thermo Fisher Scientific, Santa Clara, CA, USA), with imputation to the Haplotype Reference Consortium r1.0 reference panel and the combined UK10K plus 1000 Genomes phase 3 reference panel. We computed weighted genetic risk scores using genome-wide significant variants, as well as genome-wide PRSs using LDpred2[12] for diseases with publicly available summary statistics (appendix pp 40–41). Further details are provided in the appendix (pp 3–4). We compared the predictive performance of the genome-wide significant genetic risk score versus the genome-wide PRS and kept the best-performing score for all subsequent analysis (appendix p 8). For practical reasons, we refer to the best-performing score as PRS throughout the text.

Statistical analysis

We adapted a machine learning framework that we had previously developed to identify a sparse set of predictor proteins[13] for each of the 24 incident outcomes. This included a feature selection step, and model testing by cross-validation. We aimed to use the technical separation of our study into proteomic batch 1 and batch 2 to design independent feature selection and validation sets whenever possible (15 of the outcomes under study that were available in both batches). For the remaining nine diseases (all designed as case-cohorts and available in only one of the batches), we split the entire set into two separate feature selection (70%, n=452–821) and validation (30%, n=194–352) subsets (appendix p 9). We did feature selection using least absolute shrinkage and selection operator regression over 200 subsamples (appendix pp 4–5).

The five proteins with the highest final selection scores (appendix pp 4–5) were taken forward for validation, which was done in either batch 2 or in the remaining, independent 30% validation set from batch 1 or batch 2 for nine of the case-cohorts. We did regularised Cox regression by five-fold cross-validation over 100 iterations. The cross-validation concordance index (C-index)—ie, from the held-out folds—along with the lower and upper bounds (the cross-validation error) were averaged over the 100 iterations (appendix p 9). For the ten incident diseases, which we analysed in a case-cohort design, we used Prentice weighted regularised Cox models (appendix p 5).

We further evaluated the following: (1) a patient-derived information model that included age, sex (except for the sex-specific outcomes of ovarian, breast, endometrial, and prostate cancer), BMI, and smoking status; (2) a patient-derived information model plus the top five proteins; (3) a patient-derived information model plus the disease PRS; and (4) a patient-derived information model plus the top five proteins plus the disease PRS. We also evaluated the performance of a patient-derived information model plus all 2319 proteins using ridge regression. The performance of PRS-only models was tested by use of simple Cox proportional hazards models, with an analogous bootstrapping framework. The category-free risk difference-based net reclassification improvements from the addition of five proteins, disease PRSs, or five proteins plus disease PRSs to the patient-derived information-only model were estimated in the validation subset by use of the R package nricens (version 1.6),[14] with a 0·15 cutoff in risk difference to provide more conservative estimates. Similarly, we estimated integrated Brier scores in the validation set, using the R package pec.[15]

We further derived a binary outcome for multimorbidity, defined as at least two conditions out of the 14 incident outcomes studied in the control subcohort (appendix pp 13–15). Using batch 1, we did feature selection for this multimorbidity binary outcome by least absolute shrinkage and selection operator regression over 200 subsamples, excluding participants with a prevalent status for any of these 14 outcomes. We took forward the top ten proteins with the highest feature selection score for individual disease validation by five-fold cross-validation over 100 iterations using regularised Cox regression in batch 2. We adhered to TRIPOD guidelines,[16] and provide a completed checklist (appendix pp 42–43).

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Characteristics of participants included in the current study, including those in the randomly drawn subcohort (n=1759) and those who developed one of ten less common diseases within 10 years of follow-up (n=989), are provided in the appendix (pp 16–17). Participants in the random subcohort had a mean age at baseline of 58·79 years (SD 9·31); 990 (56·28%) of the 1759 participants were women and 769 (43·72%) were men. Participants who developed one of ten less common diseases had a mean age of 64·56 years (8·08); 482 (48·74%) of 989 were women and 507 (51·26%) were men.

We derived sparse protein models for 24 different outcomes (appendix pp 44–191), including as few as

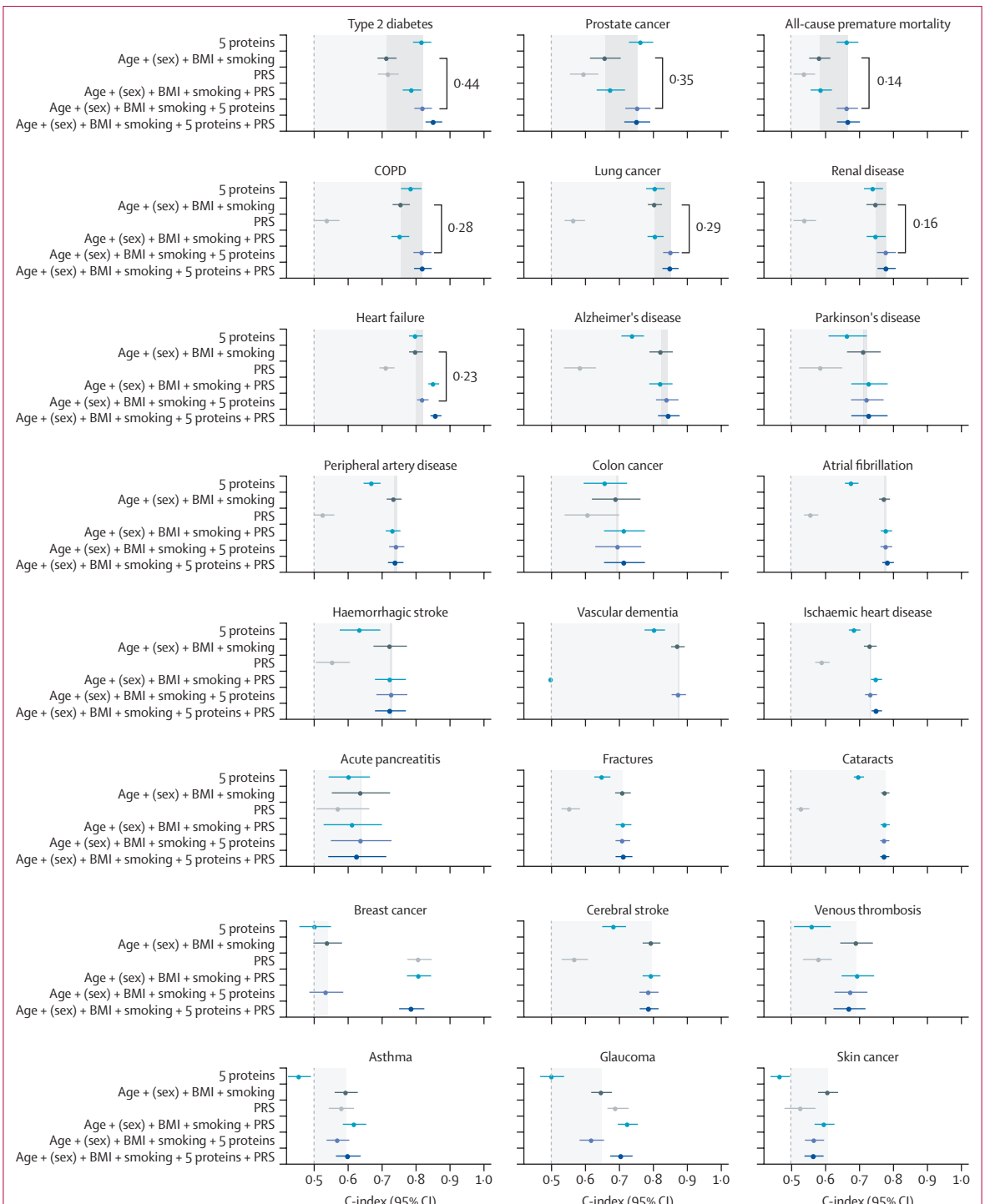

Figure 2: **Cross-validated predictive performance of protein biomarkers for 23 diseases and all-cause premature mortality**
The concordance index (C-index) of the top five predictor proteins for each disease was compared with those of basic patient-derived information models that used age, sex, BMI, and smoking status or disease polygenic risk scores. The C-indices achieved by adding the top five proteins or top five proteins plus disease PRSs onto the patient information models are also shown. The dark grey area represents the change in the average cross-validation C-index provided by adding the top five proteins on top of the patient information model. Diseases are ordered according to the improvement in C-index provided by the five proteins on top of the patient information model. For diseases that were significantly improved by proteins, annotated numbers represent the category-free net reclassification improvement of the model that included five proteins on top of the patient information model. COPD=chronic obstructive pulmonary disease. PRS=polygenic risk score.

five proteins, which achieved a median C-index of 0·67 (IQR 0·62–0·75; figure 2). The top five proteins did better than models trained using all 2319 proteins for most diseases, achieving a median C-index that was 0·04 higher (IQR 0·01–0·06; appendix p 192). For eleven of the outcomes under study, protein-only models (median C-index 0·74 [IQR 0·66–0·80]) did as well as or outperformed basic patient-derived information models that included risk factors (median C-index 0·71 [0·65–0·75]; figure 2). Proteins alone further outperformed PRSs containing up to 722 108 genetic variants for 17 diseases. The median difference in C-index between the five-protein and PRS models was 0·13 (IQR 0·10–0·17). Most of the selected predictor proteins were positively associated with disease risk, with few examples of inverse associations (appendix p 10). We note that for some diseases that were poorly predicted, there was substantial effect heterogeneity between the two proteomic batches (eg, for N-terminal pro-B-type natriuretic peptide and atrial fibrillation), indicating further potential to improve assay performance and generalisability (appendix p 10).

Adding the top five proteins to these patient-derived information models improved the predictive performance for seven outcomes: type 2 diabetes, prostate cancer, all-cause premature mortality, COPD, lung cancer, renal disease, and heart failure (range of C-index improvements 0·02–0·11; figure 2). The largest improvements were seen for type 2 diabetes (C-index improvement 0·11 [+/– cross-validation error 0·08–0·13]), prostate cancer (0·10 [0·06–0·13]), and all-cause premature mortality (0·08 [0·05–0·11]; appendix pp 193–202). Proteins also improved the performance of models with already strong baseline predictors, such as smoking status for respiratory diseases such as COPD (0·06 [0·04–0·09]) and lung cancer (0·05 [0·02–0·07]). Across these seven outcomes, the median C-index was 0·82 (IQR 0·77–0·82). The median net reclassification improvement (NRI) was 0·28 (IQR 0·19–0·37), mostly attributable to correct reclassification of cases (median P[Up|Case]; i.e. the probability of correct reclassification of cases 0·30 [IQR 0·24–0·39]; P[Down|Control]; ie, the probability of correct reclassification of controls 0·01 [IQR 0·002–0·02]; appendix pp 203–05). PRSs improved prediction over patient-derived information models for five diseases (range of C-index improvements 0·02–0·27), including breast cancer, type 2 diabetes, glaucoma, heart failure, and ischaemic heart disease (appendix pp 193–94). The median NRI was 0·19 (IQR 0·17–0·26), with a greater contribution from correct reclassification of controls (median P[Up|Case] 0·22 [0·18–0·35]; P[Down|Control] 0·04 [0·01–0·08]) compared with proteomic prediction (appendix pp 203–05). Synergistic improvements from adding disease PRSs plus five proteins on top of the patient information models were only achieved for type 2 diabetes (improvement in C-index compared with the patient information model 0·14 [+/– cross-validation error 0·11–0·16]; figure 2). Integrated Brier scores showed superior calibration for most models that included proteins or PRSs compared with patient information models (appendix p 206).

We next sought to establish whether we could derive a single common sparse proteomic signature for prediction of multiple diseases at once, which would provide a cost-effective strategy with improved potential for clinical translation. The top ten multimorbidity proteins achieved a median C-index of 0·72 (IQR 0·64–0·76) across 21 individual incident diseases. This was higher, on average, than the performance of the disease-specific protein signatures, which might point to shared disease mechanisms. These ten proteins improved the predictive performance for six diseases and all-cause premature mortality over the patient information model (range of change in C-index 0·02–0·06; median C-index 0·81 [IQR 0·80–0·82]; figure 3; appendix pp 193–94). Notably, type 2 diabetes and vascular dementia were not included in the definition of the composite multimorbidity outcome, but their prediction was still improved by the ten-protein signature. The median NRI across these seven diseases was 0·28 (IQR 0·18–0·31). As with single-disease predictive signatures, this was mainly attributable to correct reclassification of cases (median P[Up|Case] 0·30 [IQR 0·23–0·34]; P[Down|Control] 0·01 [0·006–0·02]; appendix pp 207–08).

For most disease-specific protein signatures that provided an improvement over patient information models, there were a few strong candidates, beyond which there was a marked decrease in selection scores, suggesting potentially little value in generating more comprehensive proteomic signatures (figure 4). We note that allowing selection of a variable number of proteins as predictors on the basis of normalised selection scores above a fixed threshold (ranging from one to 16 proteins; appendix p 209) resulted in C-indices similar to those for the selection of five proteins only (Pearson's r=0·99).

Among the top predictors were established clinical biomarkers, but also strongly predictive proteins that have been rarely reported in the literature so far, including C-X-C motif chemokine ligand 17 (CXCL17) for lung cancer and COPD, and leiomodin 1 (LMOD1) for renal disease (appendix p 210).

We did not observe an enrichment of proteins in any of the 384-plex pre-grouped and commercially available specific panels (ie, cardiometabolic, inflammatory, oncology, or neurology) for a specific group of related diseases among the selected predictor proteins (appendix p 12). This is an important consideration for the design of explorative studies.

Overall, across the top 20 proteins from disease-specific signatures that performed at least as well as or improved on patient information models, 26 proteins were shared between two or more diseases (figure 4). Although these results indicate less overlap compared with other omics layers, such as metabolomics,[17] they highlight that some

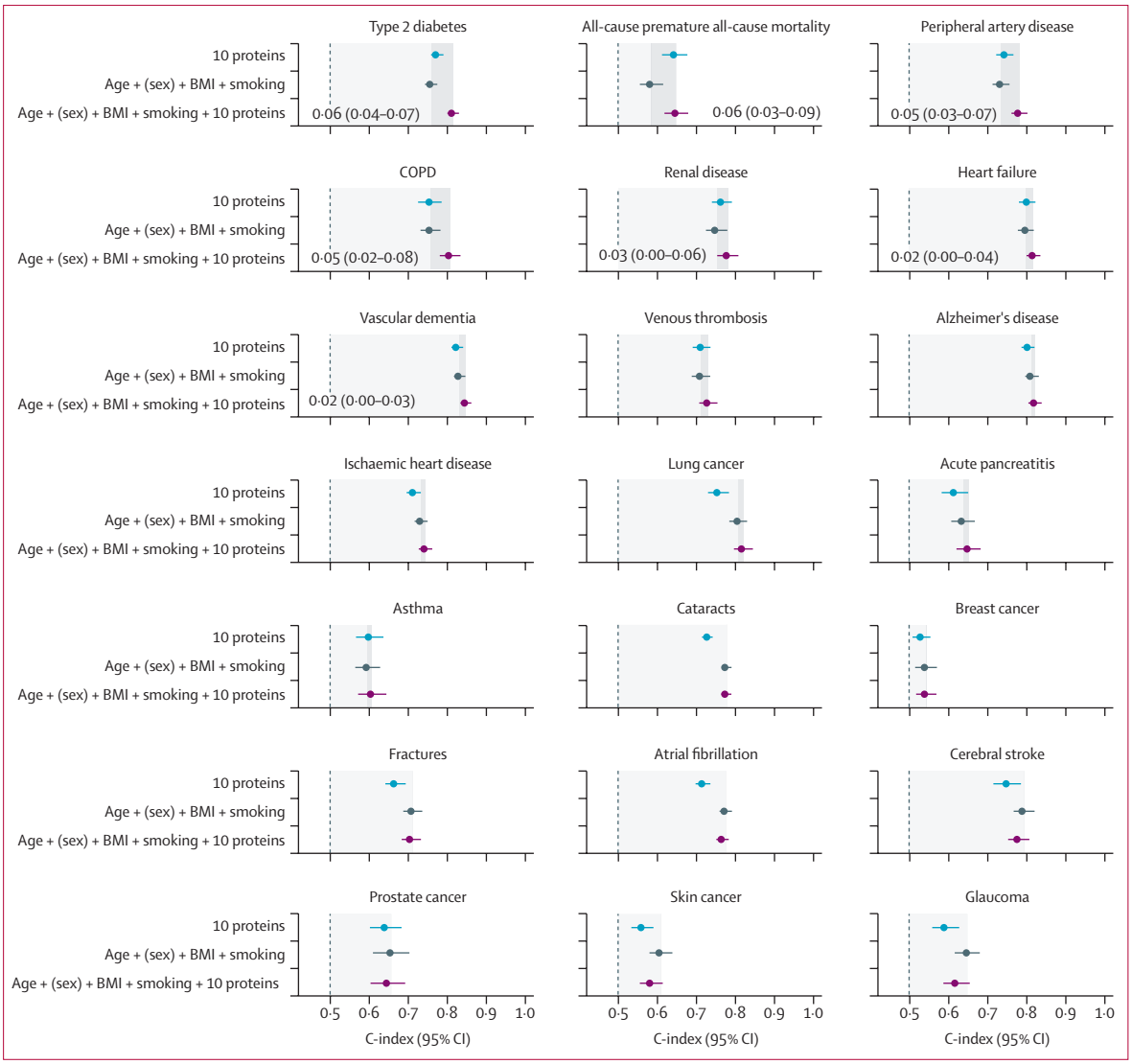

***Figure 3*: Cross-validated predictive performance of ten multimorbidity proteins for 20 diseases and all-cause premature mortality**
The concordance index (C-index) achieved by the top ten multimorbidity proteins for each disease was compared with those of basic patient-derived information models that used age, sex, BMI, and smoking status. The C-indices achieved by adding the top ten multimorbidity proteins onto the patient information models are also shown. The dark grey area represents the change in the average cross-validation C-index provided by adding the top ten proteins on top of the patient information model. Diseases are ordered according to the improvement in C-index provided by the ten proteins on top of the patient-information model. COPD=chronic obstructive pulmonary disease.

proteins might indicate mechanisms shared by multiple diseases and hence provide markers for clusters of multimorbidity. This possibility was highlighted by the good predictive performance, on average, of the ten multimorbidity protein signatures across individual diseases tested. This signature included shared markers from single-disease models such as growth-differentiation factor 15, CUB domain containing protein 1, ectodysplasin A2 receptor, neurofascin, or matrix metalloproteinase 12 (figure 4).

We systematically tested whether any of the selected predictive proteins might be causally involved in the pathogenesis of the associated disease or related entities, but found no matching evidence from a comprehensive genetic colocalisation study.[18]

Discussion

Here, we provide proof-of-concept that machine learning-guided proteomic biomarker discovery studies have the potential to improve prediction models for selected diseases. We showed that a single common sparse multimorbidity signature, including as few as ten proteins out of almost 3000 targets tested, improved the predictive performance over patient-derived risk factors for six diseases spanning different clinical specialties and all-cause premature mortality. Furthermore, we showed

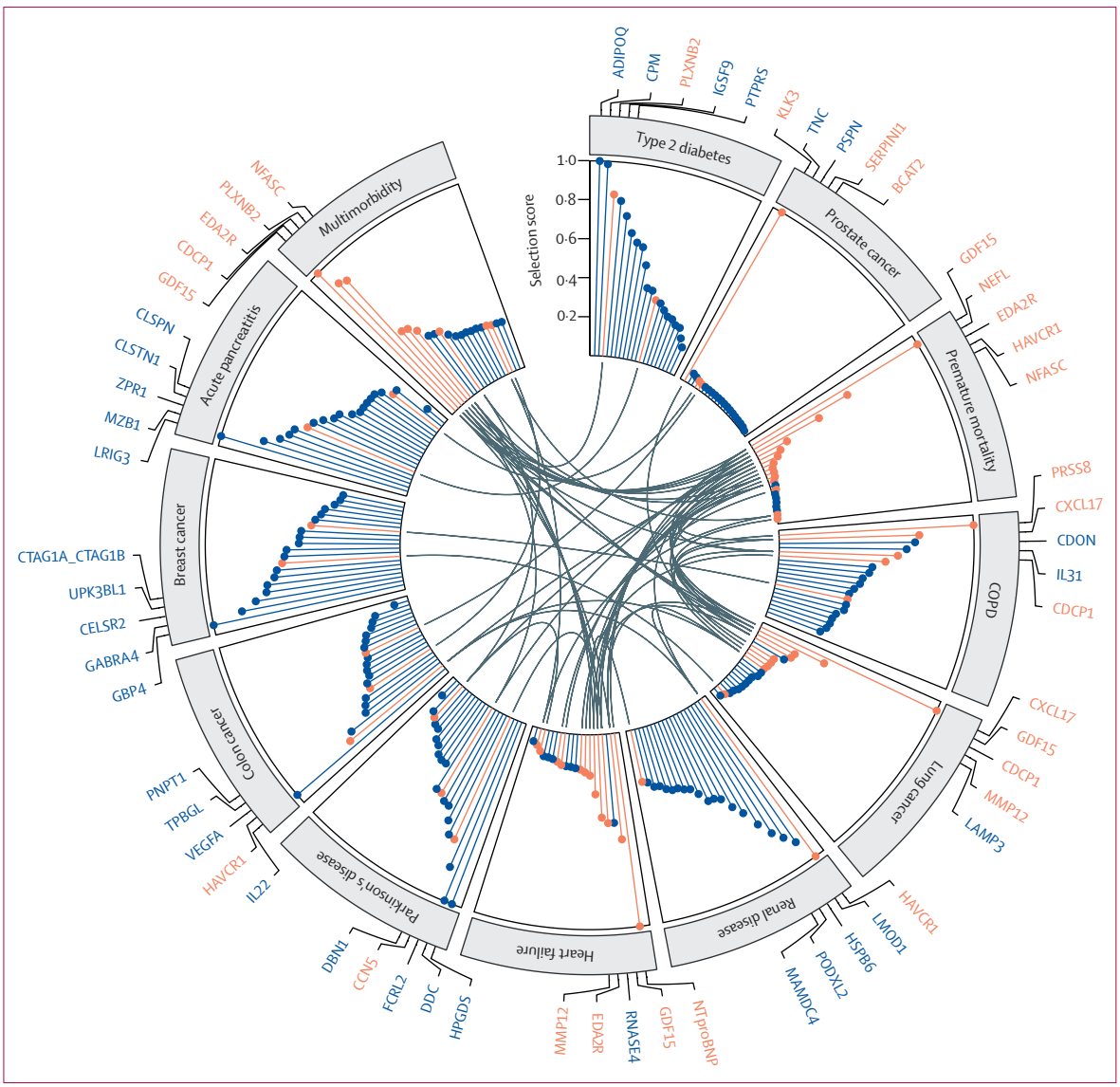

Figure 4: **Normalised feature selection scores for the top 20 proteins for the diseases in which proteins at least equalled or improved the performance of the patient information model**

Selection scores are shown for the top 20 proteins, normalised to the protein with the highest selection for interpretability. The links in the inner track represent proteins that were selected among the top 20 predictors for more than one disease (orange points and labels). Blue points and labels represent predictors that were disease specific (ie, not selected among the top 20 for any other of these 11 outcomes). ADIPOQ=adiponectin. BCAT2=branched-chain-amino-acid aminotransferase, *mitochondrial*. CCN5=CCN family member 5. CDCP1=CUB domain containing protein 1. CDON=cell adhesion molecule-related/down-regulated by oncogenes. CELSR2=cadherin EGF LAG seven-pass G-type receptor 2. CLSPN=claspin. CLSTN1=calsyntenin-1. CPM=carboxypeptidase M. CTAG1A_CTAG1B=cancer/testis antigen 1. CXCL17=C-X-C motif chemokine ligand 17. DBN1=drebrin. DDC=aromatic-L-amino-acid decarboxylase. EDA2R=ectodysplasin A2 receptor. FCRL2=Fc receptor-like protein 2. GABRA4=GABA receptor subunit alpha-4. GBP4=guanylate-binding protein 4. GDF15=growth-differentiation factor 15. HAVCR1=hepatitis A virus cellular receptor 1. HPGDS=haematopoietic prostaglandin D synthase. HSPB6=heat shock protein beta-6. IGSF9=protein turtle homolog A. IL22=interleukin-22. IL31=interleukin-31. KLK3=prostate specific antigen. LAMP3=lysosome-associated membrane glycoprotein 3. LMOD1=leiomodin 1. LRIG3=leucine-rich repeats and immunoglobulin-like domains protein 3. MAMDC4=apical endosomal glycoprotein. MMP12=matrix metalloproteinase 12. MZB1=marginal zone B- and B1-cell-specific protein. NEFL=neurofilament light polypeptide. NFASC=neurofascin. NTproBNP=N-terminal pro-B-type natriuretic peptide. PLXNB2=plexin-B2. PNPT1=polyribonucleotide nucleotidyltransferase 1, mitochondrial. PODXL2=podocalyxin-like protein 2. PRSS8=prostasin. PSPN=persephin. PTPRS=receptor-type tyrosine-protein phosphatase S. RNASE4=ribonuclease 4. SERPINI1=neuroserpin. TNC=tenascin. TPBGL=trophoblast glycoprotein-like. UK3BL1=uroplakin-3b-like protein 1. VEGFA=vascular endothelial growth factor A. ZPR1=zinc finger protein ZPR1.

that as few as five disease-specific proteins achieved similar predictive performances or improved patient-derived information models for selected diseases, including lung cancer and COPD as prominent examples.

Predicting the risk of future disease development could enable early intervention and targeting of preventive strategies to high-risk groups and individuals. Much effort over the past few years has focused on genetic and

polygenic prediction[2,3]—an appealing concept, since it could represent a single, inexpensive test across many diseases done at any stage of life. We showed that proteomic models often have superior performance over static germline-based models. This might reflect the potential of circulating proteins to capture current health status[1,19] and act as early disease detectors, which might be sensitive to pathological processes even before the development of overt symptoms (albeit not in a tissue-specific manner). In contrast, PRSs are static and do not capture stages of disease processes in response to environmental and lifestyle risk factors. This might explain why even very sparse protein-only models outperformed PRSs for most of the diseases in this study.

Our systematic investigation enabled the identification of diseases for which proteins, PRSs, or a combination provided added predictive utility over basic patient-derived risk factors. To the best of our knowledge, this comparison has not been previously done systematically. We provide early insights suggesting the potential to identify groups of diseases for which genetic or proteomic screening might be better suited, guiding future avenues of research.

We provide proof-of-concept of the potential to identify a sparse core set of proteins that might improve the prediction of multiple diseases at once. This approach could represent a cost-effective strategy to improve, conceptually, the utility of proteomic-based models in clinical settings (although we acknowledge the practical challenges). Among these proteins were known markers of mortality and morbidity such as growth-differentiation factor 15 (a strong correlate of age), but also biomarkers shared across diseases at different anatomical sites. For example, matrix metalloproteinase 12, a predictor of lung cancer and heart failure, and included in the multimorbidity signature, is involved in extracellular matrix breakdown and remodelling and has been previously associated with lung function.[20] Matrix metalloproteinase 12 might therefore point to the molecular mechanisms linking poor lung function and an increased risk of heart failure.[21] Conversely, neurofilament light chain (forming intermediate filaments for neurons) and neurofascin (involved in neurite outgrowth and stabilisation of axon initial segments), which were also predictors for all-cause premature mortality, might represent circulating markers of neuronal or glial ageing or damage. However, the underlying mechanisms that link these proteins to the onset of multiple diseases, and whether a wider range of diverse diseases would benefit from this cost-effective strategy, need to be established in future research.

Although we did not observe evidence that predictive proteins are causally linked to outcomes, some might convey information on early disease processes. For example, CXCL17, selected for lung cancer and COPD, is expressed in the epithelium of the lung airways and involved in innate immunity by attracting lung macrophages.[22] CXCL17 has been implicated in various cancers, including non-small-cell lung cancer,[23] without a clearly emerging mechanism of action,[24] as well as in idiopathic pulmonary fibrosis and influenza A (H1N1).[25] Altogether, this suggests that CXCL17 might be a general marker of lung inflammation, which possibly contributes to a tumorigenic environment in a chronic scenario. We further identified candidates that have been proposed as prognostic markers, or that have been shown to be associated with disease incidence, such as hepatitis A virus cellular receptor 1 for kidney injury.[26] Here, we show their added predictive value over clinical risk factors and the value of systematic feature selection strategies to recapitulate known and novel predictive biomarkers.

We extended the scope of earlier studies that explored signatures including hundreds of proteins for the prediction of selected diseases[1,19] to 24 diverse incident outcomes, and by deriving extremely sparse prediction models. This framework could, more feasibly, enable follow-up validation studies with standardised immunoassays to explore transferability of proteomic models across different ethnicities, sex differences, and in different time intervals before disease development. These aspects of disease prediction remain largely unexplored and are required to understand the potential for clinical translation.

Although our case-cohort design is an efficient approach to derive biomarkers for diseases of interest, the sample sizes for some outcomes are comparatively small and represent a limitation of our study. Larger studies might explore the predictive utility of protein signatures for more clinically useful timeframes (eg, 1-year or 5-year incidence) and estimate improvements by the use of clinically meaningful performance metrics, such as detection rates or predictive values, more accurately. Although emerging findings already provide independent support for selected diseases, including lung cancer,[27] our results require external validation in independent studies with gold-standard case ascertainment, and additional benchmarking against blood tests already used in clinical practice. Given that our study included only participants of European descent from the east of England, generalisability to other ethnically diverse populations must be assessed in future work. Furthermore, we did proteomic measurements in two batches owing to various practical restrictions, in a cohort with historical samples, which probably introduced unwanted technical variation. Although this might have masked true biological variation preceding disease onset, it reflects a more realistic scenario for model development, and for independent validation and potential application in different studies. Our findings might, therefore, point to only the more generalisable models for diseases with early strong effects on circulating candidate protein biomarkers that can be targeted more robustly, and highlight the need for further assay development to ensure reliable model transferability. Finally, proteomic technologies able to capture

post-translational modification might expand the biomarker discovery space.

We show the value of broad-capture proteomic platforms to enable systematic and hypothesis-free biomarker discovery strategies. Our study provides timely insights into the way in which improvements in disease prediction, over and above the use of common risk factors for selected diseases, can be achieved through the integration of proteomics, health record linkage, and machine learning, providing a guide for further advances in the context of an ever-increasing number of large-scale cohorts with proteomic profiling.

Contributors

JC-Z, MP, NJW, and CL designed the analysis and drafted the manuscript. JC-Z, MP, and MK analysed the data. JC-Z and NDK verified the raw data and did the quality control of proteomic data, and EW assisted and advised on genetic analyses. NJW is principal investigator of the EPIC-Norfolk study. All authors contributed to the interpretation of the results and critically reviewed the manuscript. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

EW is now an employee at AstraZeneca. The remaining authors declare no competing interests.

Data sharing

The EPIC-Norfolk data can be requested by bona fide researchers for specified scientific purposes via the study website (https://www.mrc-epid.cam.ac.uk/research/studies/epic-norfolk/). Data will either be shared through an institutional data sharing agreement or arrangements will be made for analyses to be done remotely without the need for data transfer. The code used for the machine learning framework has been deposited in the following GitHub repository: https://github.com/MRC-Epid/EPIC_Olink_protein_prediction.

Acknowledgments

The EPIC-Norfolk study has received funding from the Medical Research Council (numbers MR/N003284/1 and MC_UU_00006/1 to NJW) and Cancer Research UK (number C864/A14136 to NJW). We thank the EPIC-Norfolk investigators, the Study Co-ordination team, and the Epidemiology Field, Data and Laboratory teams. This work was supported in part by MRC Rapid Call (MC_PC_21036, to NJW and CL) and HDRUK Multi-Omics (G107794, to CL) grants and the UKRI–NIHR Strategic Priorities Award in Multimorbidity Research for the Multimorbidity Mechanism and Therapeutics Research Collaborative (MR/V033867/1, to CL). Proteomics measurements were supported by a collaboration agreement between the University of Cambridge and Olink. We thank Philippa Pettingill, Ida Grundberg, and Janet Kenyon for their support with quality control of the proteomic data. JC-Z is supported by a 4-year Wellcome Trust PhD Studentship and the Cambridge Trust. CL, EW, and NJW are funded by the Medical Research Council (MC_UU_00006/1).

References

1 Williams SA, Kivimaki M, Langenberg C, et al. Plasma protein patterns as comprehensive indicators of health. *Nat Med* 2019; **25:** 1851–57.

2 Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* 2018; **19:** 581–90.

3 Polygenic Risk Score Task Force of the International Common Disease Alliance. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat Med* 2021; **27:** 1876–84.

4 Anderson NL. The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. *Clin Chem* 2010; **56:** 177–85.

5 Zanini JC, Pietzner M, Langenberg C. Integrating genetics and the plasma proteome to predict the risk of type 2 diabetes. *Curr Diab Rep* 2020; **20:** 60.

6 Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* 2002; **1:** 845–67.

7 Wu CC, Tsantilas KA, Park J, et al. Mag-Net: Rapid enrichment of membrane-bound particles enables high coverage quantitative analysis of the plasma proteome. *bioRxiv* 2024; published online April 2. https://doi.org/10.1101/2023.06.10.544439 (preprint).

8 Ferdosi S, Tangeysh B, Brown TR, et al. Engineered nanoparticles enable deep proteomics studies at scale by leveraging tunable nano–bio interactions. *Proc Natl Acad Sci USA* 2022; **119:** e2106053119.

9 Messner CB, Demichev V, Wendisch D, et al. Ultra-high-throughput clinical proteomics reveals classifiers of COVID-19 infection. *Cell Syst* 2020; **11:** 11–24.e4.

10 Day N, Oakes S, Luben R, et al. EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *Br J Cancer* 1999; **80** (suppl 1)**:** 95–103.

11 Zhong W, Edfors F, Gummesson A, Bergström G, Fagerberg L, Uhlén M. Next generation plasma proteome profiling to monitor health and disease. *Nat Commun* 2021; **12:** 2493.

12 Prive F, Arbel J, Vilhjalmsson BJ. LDpred2: better, faster, stronger. *Bioinformatics* 2020; **36:** 5424–31.

13 Carrasco-Zanini J, Pietzner M, Lindbohm JV, et al. Proteomic signatures for identification of impaired glucose tolerance. *Nat Med* 2022; **28:** 2293–300.

14 Inoue E. nricens: NRI for risk prediction models with time to event and binary response data. R package version 1.6 ed; 2018. https://cran.r-project.org/web/packages/nricens/index.html (accessed April 1, 2024).

15 Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. *J Stat Softw* 2012; **50:** 1–23.

16 Walsh I, Fishman D, Garcia-Gasulla D, et al. DOME: recommendations for supervised machine learning validation in biology. *Nat Methods* 2021; **18:** 1122–27.

17 Pietzner M, Stewart ID, Raffler J, et al. Plasma metabolites to profile pathways in noncommunicable disease multimorbidity. *Nat Med* 2021; **27:** 471–79.

18 Koprulu M, Carrasco-Zanini J, Wheeler E, et al. Proteogenomic links to human metabolic diseases. *Nat Metab* 2023; **5:** 516–28.

19 Williams SA, Ostroff R, Hinterberg MA, et al. A proteomic surrogate for cardiovascular outcomes that is sensitive to multiple mechanisms of change in risk. *Sci Transl Med* 2022; **14:** eabj9625.

20 Hunninghake GM, Cho MH, Tesfaigzi Y, et al. MMP12, lung function, and COPD in high-risk populations. *N Engl J Med* 2009; **361:** 2599–608.

21 Silvestre OM, Nadruz W Jr, Querejeta Roca G, et al. Declining lung function and cardiovascular risk: the ARIC study. *J Am Coll Cardiol* 2018; **72:** 1109–22.

22 Burkhardt AM, Maravillas-Montero JL, Carnevale CD, et al. CXCL17 is a major chemotactic factor for lung macrophages. *J Immunol* 2014; **193:** 1468–74.

23 Djureinovic D, Pontén V, Landelius P, et al. Multiplex plasma protein profiling identifies novel markers to discriminate patients with adenocarcinoma of the lung. *BMC Cancer* 2019; **19:** 741.

24 Ohlsson L, Hammarström ML, Lindmark G, Hammarström S, Sitohy B. Ectopic expression of the chemokine CXCL17 in colon cancer cells. *Br J Cancer* 2016; **114:** 697–703.

25 Choreño-Parra JA, Jiménez-Álvarez LA, Ramírez-Martínez G, et al. CXCL17 Is a specific diagnostic biomarker for severe pandemic influenza A(H1N1) that predicts poor clinical outcome. *Front Immunol* 2021; **12:** 633297.

26 McCoy IE, Hsu JY, Bonventre JV, et al. Acute kidney injury associates with long-term increases in plasma TNFR1, TNFR2, and KIM-1: findings from the CRIC Study. *J Am Soc Nephrol* 2022; **33:** 1173–81.

27 Albanes D, Alcala K, Alcala N, et al. The blood proteome of imminent lung cancer diagnosis. *Nat Commun* 2023; **14:** 3042.