Plasma protein patterns as comprehensive indicators of health.

Stephen A Williams[1], Mika Kivimaki[2], Claudia Langenberg[3], Aroon D Hingorani[4,5,6,] JP Casas[7], Claude Bouchard[8], Christian Jonasson[9], Mark A Sarzynski[10], Martin J Shipley[2], Leigh Alexander[1], Jessica Ash[1], Tim Bauer[1], Jessica Chadwick[1], Gargi Datta[1], R Kirk DeLisle[1], Yolanda Hagar[1], Michael Hinterberg[1], Rachel Ostroff[1], Sophie Weiss[1], Peter Ganz[11], Nicholas Wareham[3]

Institutions

[1] SomaLogic Inc, Boulder, USA.
[2] Department of Epidemiology and Public Health, University College London, UK.
[3] Department of MRC Epidemiology, University of Cambridge School of Clinical Medicine, UK.
[4] Institute of Cardiovascular Science, University College London, UK.
[5] University College London, British Heart Foundation Research Accelerator, UK
[6] Health Data Research UK
[7] Massachusetts Veterans Epidemiology and Research Information Center, Veterans Affairs Boston Healthcare System, USA
[8] Pennington Biomedical Research Laboratory, Louisiana State University, USA
[9] Norwegian University of Technology and Sciences, HUNT Research Center, NO
[10] Department of Exercise Science Arnold School of Public Health University of South Carolina, USA
[11] Division of Cardiology, Center of Excellence in Vascular Research, Zuckerberg San Francisco General Hospital, University of California San Francisco, USA

Equal contributions statement
Drs S Williams, P Ganz and N Wareham as co-senior authors were equally responsible for the overall program design, the interpretation of the findings, writing and editing of the manuscript and for the responses to reviewers' comments.

Corresponding author statement
Dr S Williams as the corresponding author certifies that that data, materials, and code comply with transparency and reproducibility standards of the field and journal; that original data/materials/code upon which the submission is based are preserved following best practices in the field so that they are retrievable for reanalysis; he confirms that data/materials/code presentation accurately reflects the original; he commits to minimizing obstacles to the sharing of data/materials/code described in the work and he has ensured that all authors have certified the author list and author contributions

Abstract

Proteins are effector molecules that mediate the functions of genes[1-2] and modulate co-morbidities [3-10], behaviors and drug treatments.[11] They represent an enormous potential resource for personalised, systemic and data driven diagnosis, prevention, monitoring and treatment. However, the concept of using plasma proteins for individualized health assessment across many health conditions simultaneously has not been tested. Here, we show that plasma protein expression patterns strongly encode for multiple different health states, future disease risks and lifestyle behaviors. We developed and validated protein-phenotype models for 11 different health indicators: liver fat, kidney filtration, percent body fat, visceral fat mass, lean body mass, cardio-pulmonary fitness, physical activity, alcohol consumption, cigarette smoking, diabetes risk and primary cardiovascular event risk. The analyses were prospectively planned, documented and executed at scale on archived samples and clinical data, with a total of ~ 85 million protein measurements in 16,894 participants. Our proof-of-concept study demonstrates that protein expression patterns reliably encode for many different health issues, and that large-scale protein scanning[12-16] coupled with machine learning is viable for the development and future simultaneous delivery of multiple measures of health. We anticipate that with further validation and the addition of more protein-phenotype models this approach could enable a single-source, individualized so-called liquid health-check.

Letter

As populations worldwide are increasingly affected by multimorbidity and avoidable chronic health conditions, the need to prevent illness is increasing.[17] In response, health care providers have instituted preventative medicine programmes. For example, the British National Health Service has implemented a triple prevention strategy[18] with initiatives such as Health Check[19] Healthier You[20] and the National Diabetes Prevention Programme.[20] The advantages of such approaches are that they are inexpensive, cost effective and scalable.[20] However, the key tools to make them useful could be improved beyond taking medical history, a limited number of laboratory tests and group participation in health coaching. While the low-cost tests and assessments of lifestyle are prognostic on a population level, long-term adherence is difficult to sustain[21] and a process that is not individualised cannot be optimal for everyone.

Applications of big data and systems medicine have been suggested to provide additional information to transform healthcare[22,23] but these claims depend on the degree to which the information sought is encoded within the data source and whether it can be easily extracted. There is some evidence of reduced healthcare utilisation associated with information-rich physiologic health measurements[24] but scalability is limited by the high cost of generating these data. This study evaluates whether protein scanning can fill the gap between contemporary demands for practicality and low cost and the future promise of the impact of personalised, systemic and data driven medicine.

Proteins regulate biological processes and can integrate the effects of genes with those of the environment, age, comorbidities, behaviours and drugs[2]. There are about 19,000 human genes coding for approximately 30,000 proteins.[24] Of these, up to 2,200 proteins enter the blood stream by purposeful secretion to orchestrate biological processes in health or in disease, including hormones, cytokines, chemokines, adipokines and growth factors.[26] Other proteins enter plasma through leakage from cell damage and cell death. Both secreted and leakage proteins can inform health status and disease risk. We therefore hypothesised that protein scanning could deliver comprehensive individualized health assessments – but with single-source convenience and greater usability in typical medical practice. While this approach using modified aptamers has gained provenance for discovering and understanding gene-protein interactions[1], drug pharmacology[11], biological control systems[2], biomarkers in individual diseases and risks[3-8], aging[9] and obesity[10] it has not been evaluated previously as a potentially holistic, quantitative health assessment to evaluate multiple health issues simultaneously.

In this proof-of-concept study based on five observational cohorts in 16,894 participants, we evaluated the ability of scanning ~5000 proteins in each plasma sample to simultaneously capture the individualised imprints of current health status, the impact of modifiable behaviors and incident risk of cardiometabolic diseases (diabetes, coronary heart disease, stroke or heart failure).

Models were developed for 11 of 13 pre-defined health measures; their performance metrics are shown in **Table 1** and graphically in **Figure 1.** Success was defined as at least equivalent performance of a validated model to the best available comparator (CV risk and incident diabetes risk, measured by c-statistic and/or net reclassification index[27,28]). Where there was no

comparator, success was a high degree of correlation with a truth standard (spearman correlation coefficients > 0.6 (i.e. $r^2 > 0.36$) or for binary measures an AUC > 0.7).

For current health states, protein-phenotype model performance metrics in the validation datasets are as follows: predicting presence/absence of liver fat by ultrasound: AUC = 0.83 for proteins, AUC = 0.64 for the best clinical model using age, sex, alcohol, statins and pre-diabetes status; predicting kidney function, eGFR above/below 60ml/min: AUC = 0.94; predicting % body fat by DEXA: r^2= 0.92 for proteins and 0.74 for the best clinical model using sex, height and weight; predicting kg. visceral fat by DEXA: r^2= 0.70; predicting kg. lean body mass by DEXA: r^2= 0.82 for proteins and 0.74 for the best clinical model using age, sex and height; predicting cardiopulmonary fitness, VO_2max ml/min/kg: r^2= 0.71.

For modifiable behaviours, model validation performance metrics are as follows: predicting average daily physical activity energy expenditure in kJ/kg/day from individually calibrated heart rate and movement sensing: r^2= 0.38; predicting alcohol consumption on self-reported questionnaires above or below UK guidelines of 14 units/wk., separate models for men and women: AUC = 0.86 for women and 0.82 for men; predicting current cigarette smoking on self-reported questionnaires: AUC = 0.82.

For future cardiometabolic risks, model validation performance metrics are as follows: predicting incident diabetes in pre-diabetics within 10 years: accuracy 67% vs. 61% for the best oral glucose tolerance model trained in the same participants using combined fasting and peak glucose levels; predicting primary CV events (MI, stroke, hospitalization for heart failure or CV death) within 5 years: C-statistic of 0.66 and Net Reclassification Index of +0.21 vs. the reference 2013 ACC/AHA ASCVD risk score, which had a C-statistic of 0.65.

There were two unsuccessful model attempts: we found no significant proteins that predicted future body weight 5 years after the blood sample when evaluated in the incident diabetes subset of Whitehall II; and preliminary model correlations within the Fenland study predicting macronutrient intake by questionnaire (dietary fat, carbohydrate and protein intake) only had r^2 values of ~ 0.1 each.

Overall, each successful model incorporated between 13 and 375 protein measurements, with a total of 891 unique human proteins incorporated across all models. The top 3 proteins with the largest mathematical contribution to each model, along with their biological relevance to the phenotype are shown in **Table 2** and complete protein lists for all the models can be found in **Supplemental Table 1**. The proportionate degree of protein overlap across phenotype models is shown in **Figure 2**. Overall, the degree that proteins in one model were represented in another was modest, with a mean of 12% shared. The most frequently selected individual protein was leptin, which was important for percent total body fat, visceral fat, physical activity and cardiorespiratory fitness. Within the 110 possible cross-model comparisons in **Figure 2**, only 12 had more than a 25% overlap in proteins shared across models. The highest combined overlap was between visceral fat and liver fat (38% of visceral fat proteins were represented in the liver fat model and [coincidentally] 38% of the liver fat proteins were represented in the visceral fat model). Of the 96 proteins in the model for visceral fat, 29%, 29% and 38% were shared with incident diabetes, lean body mass and liver fat, respectively. Of the 115 proteins in the protein-

phenotype model for lean body mass, 29%, 26% and 26% were shared with the visceral fat, physical activity and VO_2 max. models, respectively.

Discussion

To our knowledge, this is the largest proteomic study of any kind published to date, representing a set of prospectively defined analyses of retrospective, archived samples and data from five well-characterised cohorts Approximately 5000 proteins were measured in nearly 17,000 participants, resulting in ~85 million individual protein measurements. The results were analysed rigorously by pre-defined statistical plans that relied on several state-of-the-art supervised machine learning approaches.

The intent of this proof-of-concept study was to evaluate the potential of protein scanning to become a sole information source, capable of characterising multiple elements of an individual's current health state, modifiable behaviors and future cardiometabolic health risks from a single blood sample. Capturing health information in each of these domains would be a prerequisite for an idea of a future so-called liquid health check.

The objectives were largely fulfilled. Patterns of scanned plasma proteins were validated for 6 current health states, 3 behaviors and 2 key future disease risks. The validation of these protein-phenotype models, each consisting of 13 to 375 protein measurements, involving a total of 891 human proteins, provides proof of concept for a scalable, individualised and holistic proteomic health assessment that might be delivered from plasma proteins alone.

The models we developed predicted results from some of the best clinical or physiological measures relevant to preventative health.[29-34] Acquiring the same information using standard techniques would require physician examination, laboratory testing, exercise stress testing and imaging assessments, with up to 9 different patient appointments and potentially thousands of pounds in costs per patient as shown in **Supplemental Table 2**. While some of the models demonstrated high performance (e.g. the r^2 of 0.91 for percent body fat), others had only modest prognostic power (e.g. the C-statistic of 0.66-0.69 for cardiovascular events); however, this was still modestly better than traditional risk factors and could also add value in overcoming the incomplete utilization of risk calculation in primary care.

An important feature of our study is the use of a sole information source (i.e. a single blood draw) for protein phenotype models. This was a key objective of our health-check proof of concept and therefore we did not include demographic or known risk factors in the models - unless absolutely necessary to achieve desired performance. This approach enabled the machine learning algorithms to include proteins that represented the biology of clinical and demographic factors where useful. For the same reason we also did not test whether the models could be further enhanced by the addition of other features (history, physical signs, laboratory tests or genetic information). It is possible that these multi-source models could improve absolute models' performance, although their inclusion has potential implications for increasing costs and loss of convenience.

Another nonconforming feature of this study is its separation from biological analysis. We did not use any biological plausibility or causality information from the literature for feature

selection because most proteins scanned have never been measured at scale, and because some of the proteins in our models are leakage proteins, which might inform cell injury rather than biological causality. A full biological analysis of proteins in the models is ongoing; however, this is made complex by the algorithms' biases for correlated features and their selection of proteins for normalising adjustments not related to the target physiology. Nevertheless, as a simplified alternative, we present the biological functions of the top three proteins that make the greatest mathematical contribution to each of the 11 successful models in **Table 2**. All proteins included in the 11 successful models can be perused in **Supplemental Table 1** and all proteins measured in **Supplemental Table 3**. The degree of sharing of proteins across phenotype models shown in **Figure 2** was modest, averaging 12% (range 0%-38%). The individual proteins' functions and the sharing of proteins between models were largely physiologically plausible. The individual protein with highest impact in multiple models was the appetite and metabolism regulator leptin, which was included in percent body fat, visceral fat, physical activity and cardiopulmonary fitness models. The highest overlap across models was the coincident inclusion of proteins in the liver fat model in the visceral fat model and incident diabetes models.

One limitation of our study is the nature of the truth standards we used for model training. In some cases, other good techniques exist – for example liver biopsy or magnetic resonance imaging as alternatives to ultrasound for detection of liver fat, but in all cases the chosen reference measures we used have widespread use in medicine. In other cases, self-reported measures such as alcohol and smoking are subject to individuals' truthfulness, in which case we depended on the careful evaluations made across the cohort studies that can now be applied to individuals.

Another limitation of our study is that the populations' characteristics may limit the potential generalisability of the results; in particular, a Caucasian bias in some of our cohorts will demand calibration testing in different populations. Similarly, there is a bias in model development thus far towards metabolic health that limits claims of comprehensiveness. An obvious omission here is cancer, to which earlier versions of the SomaSscan modified-aptamer assay have been applied[35,36], but these cancer findings have not yet been translated to the current more advanced platform. Finally, the greatest potential value for such assessments is likely to come from their sensitivity to longitudinal change in health status or risks; future studies will have to investigate this question.

In conclusion, this proof-of-concept study shows that scanned protein expression patterns encode for several markedly different types of health information. It is thus conceivable that with further validation and the potential for expansion of the number of tests, a comprehensive, holistic health evaluation using a battery of protein models derived from a single blood sample could be performed. The next step is to test the applicability of the protein models that we have derived and validated in observational cohorts under research conditions in real-world health care systems.

Acknowledgements

The Whitehall II study is supported by the UK Medical Research Council UK (MR/R024227/1 to MK), the US National Institutes on Aging (NIH, US R01AG056477, R01AG062553 to MK, and the British Heart Foundation (RG/16/11/32334 to MJS). ADH is an NIHR Senior Investigator and was also supported, in part, by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. FENLAND (the Fenland study (10.22025/2017.10.101.00001) is funded by UK Medical Research Council (MC_UU_12015/1) and NW is an NIHR senior investigator, and we also thank the Fenland Study Investigators, Fenland Study Co-ordination team and the Epidemiology Field, Data and Laboratory teams. HUNT3 is funded by the Norwegian Ministry of Health, Norwegian University of Science and Technology and Norwegian Research Council, Central Norway Regional Health Authority, the Nord-Trondelag County Council and the Norwegian Institute of Public Health. The HERITAGE Family Study was funded by the US National Heart, Lung and Blood Institute grants (NIH/NHLBI) R01HL146462 (MAS) and HL45670 (HERITAGE, CB). All authors are grateful to all the volunteers/participants in all of the cohorts and to the General Practitioners, other physicans and practice staff for assistance with recruitment. SomaScan assays and the Covance study were funded by SomaLogic, Inc. The authors also thank Ashley Lowell the leader of the SomaLogic assay team, Darryl Perry for the bioinformatics of quality control, Jessica Williams for the agreements with the study institutions and Jordan Zach for clinical data organization and management.

Author contributions

In an academic-industry partnership, SomaLogic and the academic collaborators worked together on study design, interpretation of the data and preparation of the manuscript. S.A.W., P.G. and N.W. were responsible for designing, writing and final editing of the manuscript and responses to reviewer comments. In addition to all authors being generally involved in the program, specific contributions were as follows: M.K and M.J.S were accountable for the data from the Whitehall II study; C.L. and N.W. were accountable for the data from the Fenland study and advising on diabetes risk and behavioral models; C.B. and M.A.S were accountable for the data from the Heritage Family study; C.J. was accountable for the data from the HUNT3 study; R.O. was accountable for the data from the Covance study; L.A., G.D., R. K. D., Y. H., M. H., and S.W. designed and executed the machine-learning tactics and developed the models; R.O., J.A., T.B., J.C and S.A.W were responsible for the design and integration of the program across studies and R.D.H and J.P.C were particularly involved in the design, execution and intepretation of the cardiovascular risk evaluations.

Competing interests

The SomaLogic coauthors (Stephen Williams, Leigh Alexander, Jessica Ash, Tim Bauer, Jessica Chadwick, Gargi Datta, R Kirk DeLisle, Yolanda Hagar, Michael Hinterberg, Rachel Ostroff and Sophie Weiss) were/are all employees of SomaLogic Inc. which has a commercial interest in the

results. Drs. Wareham and Langenberg declared that SomaLogic has given a grant to the University of Cambridge. Dr Ganz is a member of the SomaLogic Medical Advisory board, for which he receives no remuneration of any kind. The remaining authors (Mika Kivimaki, Aroon Hingorani, JP Casas, Claude Bouchard, Christian Jonasson, Mark Sarzynski, and Martin Shipley) have no competing interests.

References

1. Sun BB, Maranville JC, Peters JE, et al. Genomic atlas of the human plasma proteome. *Nature* 2018; **558**(7708): 73-9.
2. Emilsson V, Ilkov M, Lamb JR, et al. Co-regulatory networks of human serum proteins link genetics to disease. *Science* 2018; **361**(6404): 769-73.
3. Tasaki S, Suzuki K, Kassai Y, et al. Multi-omics monitoring of drug response in rheumatoid arthritis in pursuit of molecular remission. *Nat Commun* 2018; **9**(1): 2755.
4. O'Dwyer DN, Norman KC, Xia M, et al. The peripheral blood proteome signature of idiopathic pulmonary fibrosis is distinct from normal and is associated with novel immunological processes. *Sci Rep* 2017; **7**: 46560.
5. Christensson A, Ash JA, DeLisle RK, et al. The Impact of the Glomerular Filtration Rate on the Human Plasma Proteome. *Proteomics Clin Appl* 2018; **12**(3): e1700067..
6. Ganz P, Heidecker B, Hveem K, et al. Development and Validation of a Protein-Based Risk Score for Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease. *JAMA* 2016; **315**(23): 2532-41.
7. Ngo D, Sinha S, Shen D et al. Aptamer-Based Proteomic Profiling Reveals Novel Candidate Biomarkers and Pathways in Cardiovascular Disease. Wood GC, Chu X, Argyropoulos G, et al. A multi-component classifier for nonalcoholic fatty liver disease (NAFLD) based on genomic, proteomic, and phenomic data domains. *Sci Rep* 2017; **7**: 43238.
8. Han Z, Xiao Z, Kalantar-Zadeh K, et al. Validation of a Novel Modified Aptamer-Based Array Proteomic Platform in Patients with End-Stage Renal Disease. *Diagnostics (Basel)* 2018; **8**(4).
9. Menni C, Kiddle S, Mangino M et al. Circulating Proteomic Signatures of Chronolological Age, *J Gerontol A Biol Sci Med Sci* 2014; **70**(7): 809-8161
10. Thrush A, Antoun G, Nikpay M et al. Diet-resistant obesity is characterized by a distinct plasma proteomic signature and impaired muscle fiber metabolism. *International Journal of Obesity* 2018 **42**, 353-362..
11. Williams SA, Murthy AC, DeLisle RK, et al. Improving Assessment of Drug Safety Through Proteomics: Early Detection and Mechanistic Characterization of the Unforeseen Harmful Effects of Torcetrapib. *Circulation* 2018; **137**(10): 999-1010.
12. Rohloff JC, Gelinas AD, Jarvis TC, et al. Nucleic Acid Ligands With Protein-like Side Chains: Modified Aptamers and Their Use as Diagnostic and Therapeutic Agents. *Molecular therapy Nucleic acids* 2014; **3**: e201.
13. Gold L, Ayers D, Bertino J, et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoSOne* 2010; **5**(12): e15004.
14. Brody E, Gold L, Mehan M, et al. Life's simple measures: unlocking the proteome. *JMolBiol* 2012; **422**(5): 595-606.
15. Kim CH, Tworoger SS, Stampfer MJ, et al. Stability and reproducibility of proteomic profiles measured with an aptamer-based platform. *Sci Rep* 2018; **8**(1): 8382.
16. Candia J, Cheung F, Kotliarov Y, et al. Assessment of Variability in the SOMAscan Assay. *Sci Rep* 2017; **7**(1): 14248.
17. Collaborators GBDRF, Forouzanfar MH, Alexander L, et al. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 2015; **386**(10010): 2287-323.
18. Maruthappu M. Delivering triple prevention: a health system responsibility. *Lancet Diabetes Endocrinol* 2016; **4**(4): 299-301.
19. Robson J, Dostal I, Sheikh A, et al. The NHS Health Check in England: an evaluation of the first 4 years. BMJ Open 2016;6:e008840.

20. Jonathan Valabhji, Nicholas Wareham, Kamlesh Khunti, Bridget Turner, Jenifer Smith. Response: Efficacy and effectiveness of screen and treat policies in prevention of type 2 diabetes: systematic review and meta-analysis of screening tests and interventions. *British Medical Journal* 2017; **356**(16538).
21. Middleton KR, Anton SD, Perri MG. Long-Term Adherence to Health Behavior Change. *Am J Lifestyle Med* 2013; **7**(6): 395-404.
22. Dimitrov DV. Medical Internet of Things and Big Data in Healthcare. *Healthc Inform Res* 2016; **22**(3): 156-63.
23. Flores M, Glusman G, Brogaard K, Price ND, Hood L. P4 medicine: how systems medicine will transform the healthcare sector and society. *Per Med* 2013; **10**(6): 565-76.
24. Musich S, Wang S, Hawkins K, Klemes A. The Impact of Personalized Preventive Care on Health Care Quality, Utilization, and Expenditures. *Popul Health Manag* 2016; **19**(6): 389-97.
25. Ezkurdia I, Juan D, Rodriguez JM, et al. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum Mol Genet* 2014; **23**(22): 5866-78.
26. Lin H, Lee E, Hestir K, et al. Discovery of a cytokine and its receptor by functional screening of the extracellular proteome. *Science* 2008; **320**(5877): 807-11.
27. Harrell Jr, Frank E. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer, 2015.
28. Pencina, Michael J., et al. "Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond." Statistics in medicine 27.2 (2008): 157-172.
29. Fielding CM, Angulo P. Hepatic steatosis and steatohepatitis: Are they really two distinct entities? *Curr Hepatol Rep* 2014; **13**(2): 151-8.
30. Yki-Jarvinen H. Non-alcoholic fatty liver disease as a cause and a consequence of metabolic syndrome. *Lancet Diabetes Endocrinol* 2014; **2**(11): 901-10.
31. Shuster A, Patlas M, Pinthus JH, Mourtzakis M. The clinical importance of visceral adiposity: a critical review of methods for visceral adipose tissue analysis. *Br J Radiol* 2012; **85**(1009): 1-10.
32. Ross R, Blair SN, Arena R, et al. Importance of Assessing Cardiorespiratory Fitness in Clinical Practice: A Case for Fitness as a Clinical Vital Sign: A Scientific Statement From the American Heart Association. *Circulation* 2016; **134**(24): e653-e99.
33. de Souza de Silva CG, Kokkinos P, Doom R, et al. Association between cardiorespiratory fitness, obesity, and health care costs: The Veterans Exercise Testing Study. *Int J Obes (Lond)* 2018.
34. Hobbs FD, Jukema JW, Da Silva PM, McCormack T, Catapano AL. Barriers to cardiovascular disease risk scoring and primary prevention in Europe. *QJM* 2010; **103**(10): 727-39.
35. Ostroff RM, Bigbee WL, Franklin W, et al. Unlocking biomarker discovery: large scale application of aptamer proteomic technology for early detection of lung cancer. *PLoSOne* 2010; **5**(12): e15003.
36. Ostroff RM, Mehan MR, Stewart A, et al. Early detection of malignant pleural mesothelioma in asbestos-exposed individuals with a noninvasive proteomics-based surveillance tool. *PLoSOne* 2012; **7**(10): e46091.

Figure 1: Model outputs compared with the truth standards against which they were derived. All panels show the data from the validation sets except for the diabetes survival model where the Kaplan-Meier curves are shown for the much larger discovery datasets for clarity. Box plots are broken down into quantiles: Minimum, 25%, Median, 75% and Maximum. Scatter plots include a linear line of fit (red solid line). Dashed lines represent upper and lower 95% confidence intervals.

Figure 2: **Proportion (%) of proteins in any one model which overlap with other models**. The model labels in the left column identify a group of proteins within that model; each box within a row to the right of that label contains the % of proteins in that model which are shared with the other models. For each model, the single highest degree of overlap is shaded green along with any other overlap above 25%, which are additionally emboldened. The percentages for the separate alcohol models for males and females are averaged for clarity. The total number of unique proteins used by all the models was 891; the actual number of protein features used in any one model is shown in parentheses in the 100% boxes.

	Model output (truth standard)	Action	Source	Number of participants	Metric	Result
Current health state	Liver fat: presence/absence (ultrasound)	Derivation: best subject characteristics	Fenland (70%)	7054	AUC	0.64
		Derivation: 94 protein model	Fenland (70%)	7054	AUC	0.85
		Validation: 94 protein model	Fenland (15%)	1512	AUC	0.83
	Kidney function: eGFR below 60 ml/min (CKD-EPI equation)	Derivation: 55 protein model	HUNT3 (80%)	2013	AUC	0.94
		Validation: 55 protein model	Covance (100%)	1029	AUC	0.94
	Body fat: % (DEXA)	Derivation: best subject characteristics	Fenland (70%)	8030	r^2	0.74
		Derivation: 219 protein model	Fenland (70%)	8030	r^2	0.92
		Validation: 219 protein model	Fenland (15%)	1721	r^2	0.92
	Lean body mass: kg. (DEXA)	Derivation: best subject characteristics	Fenland (70%)	8030	r^2	0.74
		Derivation: 115 protein model	Fenland (70%)	8030	r^2	0.83
		Validation: 115 protein model	Fenland (15%)	1721	r^2	0.82
	Visceral fat: kg. (DEXA)	Derivation: 96 protein model	Fenland (70%)	8016	r^2	0.71
		Validation: 96 protein model	Fenland (15%)	1718	r^2	0.70
	Cardio-pulmonary fitness: ml/kg/min (V02 max.)	Derivation: 115 protein model	Heritage (80%)	523	r^2	0.80
		Validation: 115 protein model	Heritage (10%)	62	r^2	0.71
Modifiable behavioral factors	Alcohol consumption: above/below 14 units (self-report)	Derivation: Women, 30 protein model	Fenland (70%)	3396	AUC	0.89
		Validation: Women, 30 protein model	Fenland (15%)	728	AUC	0.86
		Derivation: Men, 33 protein model	Fenland (70%)	3362	AUC	0.83
		Validation: Men, 33 protein model	Fenland (15%)	720	AUC	0.82
	Weekly physical activity: kJ/kg/day (actigraphy and individually calibrated heart rate)	Derivation: 65 protein model	Fenland (70%)	8187	r^2	0.36
		Validation: 65 protein model	Fenland (15%)	1754	r^2	0.38

	Cigarette smoking: current y/n (self-report)	Derivation: 145 protein model	Covance (80%)	820	AUC	0.97
		Validation: 145 protein model	Covance (20%)	205	AUC	0.82
Future metabolic health risks	Conversion from pre-diabetes to diabetes within 10 years, above or below 3x risk	Derivation: OGTT fasting and peak 2h glucose	Whitehall II (80%)	330	Accuracy	61%
		Derivation: 375 protein model	Whitehall II (80%)	330	Sensitivity improvement over OGTT	+30%
		Validation: 375 protein model	Whitehall II (20%)	83	Accuracy	67%
					Sensitivity improvement over OGTT	+6%
	Relative probability of a first cardiovascular event within 5 years (1x to 6x)	Derivation: ACC/AHA risk factors	HUNT 3	2464	C-statistic	0.66
		Validation: ACC/AHA risk factors	Whitehall II	265	C-statistic	0.65
		Derivation: 13 proteins & age interactions	HUNT 3	2464	C-statistic	0.69
		Validation: 13 proteins & age interactions	Whitehall II	265	C-statistic	0.66
					Event NRI	+0.13
					No-Event NRI	+0.07
					Total NRI	+0.21

Table 1: Performance metrics for each protein model and comparators/references in Derivation and validation datasets; AUC = Area Under Curve for receiver operating characteristic; NRI = Net Reclassification Index [vs. reference ACC/AHA risk score]; ACC = American College of Cardiology; AHA = American Heart Association; ASCVD = Atherosclerotic Cardiovascular disease; OGTT = Oral Glucose Tolerance Test; Spearman's correlation coefficient is used for values of r^2; best subject characteristics models were developed individually for some measures of current state, and included the highest performing combination of demographics such as age, sex, BMI and diabetes status. For diabetes prediction, accuracy was used rather than AUC because the latter was artificially inflated for the reference (risk factor) model because of late censoring of the non-diabetic group.

Issue	Model	Top 3 Proteins	Potential Role in Target Biology
Current health state	Liver fat: presence/absence (ultrasound)	SEZ6L Seizure 6-like protein	Genetic marker for cardiometabolic conditions and associated with an unhealthy lifestyle score (BMI status, physical activity, smoking and alcohol habits)
		FABPA Fatty acid binding protein (adipocyte)	Expressed in adipocytes; strongly linked to metabolic and inflammatory pathways; Increased hepatic expression and circulating levels of A-FABP (FABP-4) have been observed in patients with non-alcoholic fatty liver disease.
		IGFBP-1 Insulin-like growth factor-binding protein 1	Synthesized in the liver and plays a role in metabolism regulation and insulin resistance
	Kidney function: eGFR below 60 ml/min	TMEDA Transmembrane emp24 domain-containing protein 10	Involved in kidney development
		Apo A-IV Apolipoprotein A-IV	Lipid binding protein but also a known association with kidney disease
		b2-Microglobulin	Well known clinical measure of kidney filtration
	Body fat: %	Leptin	Produced in adipose tissue, and present in higher amounts in subjects with high BMI and % BF. Previously shown to enhance the accuracy of BMI estimates of % BF when DEXA was unavailable
		FABP Fatty acid binding protein	Involved in active fatty acid metabolism and correlated with fatty liver, diabetic nephropathy and metabolic syndrome
		SFRP4 Secreted frizzled related protein 4	Elevated in obesity and involved in obese adipose tissue pathophysiology
	Lean body mass: kg	SEZ6L Seizure 6-like protein	Genetic marker for cardiometabolic conditions and associated with an unhealthy lifestyle score (BMI status, physical activity, smoking and alcohol habits)
		SLIK4 SLIT and NTRK protein-like 4	Involved in synaptogenesis and neurite growth; no clear connection to target biology.
		WISP-2	Secreted adipokine increased in obesity and insulin resistance in the subcutaneous adipose tissue
	Visceral fat: kg	Leptin	Produced in adipose tissue, and present in higher amounts in subjects with high BMI and % BF. Leptin has also been shown to enhance the accuracy of BMI estimates of % BF when DEXA was unavailable
		FABPA	Strongly linked to metabolic and inflammatory pathways; found in adipocytes.
		INHBC	Inhibins have been shown to play a role in body composition and energy expenditure
	Cardio-pulmonary fitness: ml/kg/min	Leptin	Produced in adipose tissue, and present in higher amounts in subjects with high BMI and % BF. Leptin has also been shown to enhance the accuracy of BMI estimates of % BF when DEXA was unavailable.
		C1QR1 Complement component C1Q receptor	Part of the innate immune system
		GGH Gamma glutaryl hydrolase	Regulates intracellular folate; Energy production and the rebuilding and repair of muscle tissue by physical activity requires folate.

Modifiable behavioral factors	Alcohol consumption: above/below 14 units (FEMALE)	SCUB1 Signal peptide, CUB and EGF-like domain-containing protein 1	Promotes platelet–platelet interaction and is a biomarker of platelet activation in acute thrombotic diseases; may relate to impact of alcohol on platelets.
		SERC Phosphoserine aminotransferase	No known/clear relation with target physiology
		SCF Kit ligand	Mainly involved in hematopoiesis in adults; may relate to alcohol effects on hematopoiesis
	Alcohol consumption: above/below 14 units (MALE)	SCUB1 Signal peptide, CUB and EGF-like domain-containing protein 1	Promotes platelet–platelet interaction and is biomarker of platelet activation in acute thrombotic events; may relate to impact of alcohol on platelets.
		PTPRJ Receptor-type tyrosine-protein phosphatase eta	Modulator of cell signaling. No known/clear relation with target physiology
		Apo F Apolipoprotein F	Important role in lipid metabolism; biomarker for cirrhosis in Hepatitis C patients, associated with advancing fibrosis in fatty liver disease; may relate to alcohol effect on liver
	Weekly physical activity: kcal/day	Leptin	Produced in adipose tissue, and present in higher amounts in subjects with high BMI and % BF
		IGLO5 IgLON family member 5	Immunoglobulin adhesion molecule. No known/clear relation with target physiology
		ATF6A Cyclic AMP-dependent transcription factor ATF-6 alpha	Transcription activator that initiates the unfolded protein response during endoplasmic reticulum stress
	Cigarette smoking: current Y/N	SLIK3 SLIK and NTR protein-like 3	Involved in neurite growth; no known/clear relation with target physiology
		Secretoglobin family 3A member 1	Associated with chronic obstructive airways disease and prognosis in non–small-cell lung cancer
		TM108 Transmembrane protein 108	GWAS associations with successful smoking cessation
Future metabolic health risks	Conversion from pre-diabetes to diabetes within 10 years, above or below 3x risk	LPH Lactase-phlorizin hydrolase	No known/clear relation with target physiology
		Quinone reductase 2	Flavoprotein that catalyzes metabolic reduction of quinones; Quinone reductase may play a role in insulin secretion
		Prolylcarboxypeptidase	Key enzyme that degrades α-MSH to an inactive form unable to inhibit food intake; viewed as a therapeutic target for the treatment of metabolic disorders, such as obesity and diabetes
	Relative probability of a first cardiovascular event within 5 years (1x to 6x)	Gelsolin	Regulates dynamic actin filament organization, cell morphology, differentiation, movement, and apoptosis; overexpression has been shown to induce cardiac hypertrophy
		Antithrombin III	inactivates several enzymes of the coagulation system; may play a role in role in the progression of atherosclerosis and in the pathogenesis of acute coronary syndromes

		sTREM-1	Amplifies neutrophil and monocyte-mediated inflammatory responses; levels rise significantly in all kinds of cardiovascular disease and associated organ dysfunction

Table 2: Biological plausibility of key proteins in relation to the target physiology for each model; the top 3 mathematical contributors to each model's output are shown. A full listing of all 891 proteins in all models is shown in **Supplemental Table 1** and all proteins that were measured in **Supplemental Table 3**.

Methods

Study design

We pre-specified 13 distinct measures of current health, modifiable behaviors and incident disease risks that are recognized by health experts as useful and/or commonly used for preventative health[28-33]. They have been well-characterized in at least one of 5 independent cohort studies, as the truth-standards for deriving and validating proteomic model predictions: the British Whitehall II and Fenland, the Norwegian HUNT3 and the US Covance and HERITAGE Family studies. EDTA plasma samples had been collected from all these studies and the samples were centrifuged and frozen typically 2-10 hours after collection, a timeframe which is representative of how blood is handled in typical medical practice. Aliquots of these samples were assayed on the proteomic platform without further processing after transport and thawing.

The study designs and sample selections were from whole cohorts or case-cohort fractions throughout, intended to reduce selection and spectrum biases.[37,38]. The multi-cohort study approach was needed as no single cohort has all the specified clinical measures or outcomes. Protein model outputs were deliberately simplified with primary care practitioners and patients in mind as the key target users. The flowchart for the proteomic programme, including the source of the samples, data, model training and replication is shown in **Extended Data Figure 1. Extended Data Figure 2** shows details of the 5 parent cohort studies and **Extended Data Figures 3-6** the participant characteristics for each model endpoint.

Proteomic platform

The modified aptamer binding reagents[12], SomaScan assay[13]and its performance characteristics[15,16] have been previously described. The annotated menu for all ~5000 modified-aptamer binding reagents is shown in **Supplemental Table 3** and a diagram of how the assay works in **Supplemental Figure 1**. Median intra- and inter-assay coefficients of variation are approximately 5%[16] and assay sensitivity is comparable to that of typical immunoassays, with a median lower limit of detection in the femtomolar range.

Specificity of the modified aptamer reagents has been established in several ways. The binding affinity of 1612 reagents has been tested against structurally related proteins as described by the manufacturer in the succeeding paragraphs in this section. Because many proteins share structural and functional features, it is possible that the structural epitope to which a reagent binds is present on proteins other than the one initially used to select the reagent. Indeed, we have observed that a minority of reagents are able to bind with some degree of affinity to highly similar proteins, presumably through such a shared structural epitope, although not always with the same high affinity. Because the assay is performed in a complex biological sample containing thousands of different proteins, experimentally determining which reagents may also target other proteins to some degree can be extremely valuable in interpreting biomarker discovery data.

We first analyzed publicly available databases of known human protein sequences using sequence alignment tools (e. g., BLAST) to identify those “relevant relative” proteins that share

significant homology with proteins used to select the modified aptamer reagents. Proteins with significant homology to the target protein (i.e., proteins with greater than 40% amino acid sequence identity with the target protein) were tested experimentally, if available in the inventory or commercially available as full-length proteins from reliable vendors.

Available related proteins were analyzed with affinity capture experiments similar to immunoprecipitation protocols. Modified aptamer reagents were immobilized on streptavidin coated beads and then incubated with either the target protein or the identified related protein. The reagent-protein complexes were then washed, and the proteins labeled with a fluorophore. The complexes were then eluted and the recovery of bound protein vs. input protein was analyzed by SDS-PAGE and fluorescent imaging. When any reagent binding to proteins other than the SELEX target was observed, we performed solution affinity measurements to determine whether the reagent has similar or different affinities for the target protein and related protein. If the solution KD was within 10-fold of that for the SELEX target, the reagent was reported to bind the SELEX target and other proteins with "similar affinity". If the measured affinity differed by greater than 10-fold, we reported that the reagent binds to the protein(s) other than the SELEX target with "at least 10-fold weaker affinity." Although this is a broad statement regarding specific affinity, we do not report exact KD values because of the high variability observed in both the quality and the reported concentrations of commercially obtained purified proteins.

For 73% of cases in which proteins related to the SELEX target were available for testing, we observed binding of the reagent to the specific SELEX target and not to any of the related proteins. For example, a reagent selected to bind the protein tissue inhibitor of metalloproteinase-1 (TIMP-1), was also tested against the related proteins TIMP-2 (60% identical), TIMP-3 (31% identical), and TIMP-4 (40% identical). When this same TIMP-1 SOMAmer reagent was used in affinity enrichment from human plasma, four unique peptides corresponding to endogenous TIMP-1 were identified by LC-MS/MS in the enriched sample and no peptides corresponding to any other member of the TIMP protein family were identified. Additionally, no peptides corresponding to TIMP-1 were identified in any other plasma pulldown samples performed using 142 different SOMAmer reagents, including a TIMP-2-specific reagent. In another representative example of highly specific binding, a reagent specific for matrix metalloproteinase-10 (MMP-10) does not bind MMP-12 (61% identical), MMP-13 (57% identical), MMP-3 (80% identical), MMP-1 (61% identical), or MMP-8 (50% identical).

Whenever we observed any binding to proteins other than the SELEX target (27% of the reagents tested) in initial pulldown tests, we followed up with measurements of solution affinity. We typically measure the association of radiolabeled reagent with protein and then capture the complex using a protein-affinity chromatography medium. Saturation binding curves are then generated by titrating increasing amounts of protein in the presence of a constant, limiting amount of reagent. The KD is determined to be the protein concentration at which the half maximal binding is observed. In one typical example, initial pulldown tests indicated that one reagent binds not only to its original SELEX target (pyrophosphatase 1 (PPA1)), but also the related protein PPA2, which shares 68% amino acid sequence identity. However, solution affinity measurements determined that the affinity was greater than 10-fold stronger for PPA1 than for PPA2.

We observed that 13% of the reagents tested bound to members of a protein family with similar affinities. As previously noted, this recognition most often occurs when proteins share extensive sequence identity. Presumably, the structural epitope to which the reagent was selected is highly conserved and biochemically indistinguishable by solution equilibrium binding affinities. In fact, of the reagents that could bind a related target, ~6% (i.e. almost half of the 13%) were products of the same gene with a common epitope (e.g., splice variants like vascular endothelial growth factor (VEGF) 121 and 165 isoforms) or shared subunits in a multi-subunit complex (e.g., cyclin-dependent kinase 1/cyclin B1 complex, in which the reagent binds to the cyclin B1 subunit). The remaining ~7% appear to bind to epitopes shared amongst highly related families of proteins. For example, a reagent that binds to its SELEX target calcium/calmodulin dependent protein kinase II delta (CAMK2D) also binds the closely related proteins CAMK2A (91% identical) and CAMK2B (87% identical). Solution affinity comparisons determined that this reagent has a similar binding affinity, of approximately 2nM, for all three proteins. As expected, the amino acid sequence identity tended to be greater for those pairs that exhibited cross-reactivity: 48% mean for pairs that exhibited no cross-reactivity (no positive pull-down results), 62% for pairs with >10-fold lower affinity, but positive pull-down results, and 70% for pairs with similar affinity.

In summary, we have tested binding to related proteins for 1612 modified reagents to date. We were unable to detect binding to any related proteins for 73% of those tested. When binding to related proteins was detected, about half of these reagents exhibited binding to at least one related protein with similar affinity while the other half bound to related proteins, but with at least 10-fold weaker affinity. Specific target enrichment by pull downs from human plasma has been confirmed for 123 of the SOMAmer reagents.

In orthogonal tests of specificity, the effect of cis genetic variants on protein expression in the assay has been published for 552[1] and 1046[2] variants and orthogonal validation by mass spectrometry has been performed for ~1000 reagents[2].

In addition to mitigations from reagent specificity and affinity, the impact of non-specific binding is further reduced through a kinetic challenge during the assay. During a series of wash-steps, excess unlabeled polyanion is added (aptamers are also polyanions) which successfully competes with modified aptamers associated with highly abundant plasma proteins with low affinity, non-specific binding, and capitalises on the slow off-rates (disassociation rates) of aptamers from their intended targets. A diagrammatic overview of how the assay works is shown in **Supplemental Figure 1.**

Derivation and validation of protein-phenotype models

a) Models of current health state

Liver fat (predicting liver ultrasound result of no fat or excess fat [excess defined as the combined mild/moderate/severe grades of fat]). Within the Fenland study, 10,077 participants had liver ultrasound; 75% had no fat and 25% had mild, moderate or severe fat. An elastic net model was trained, refined and validated in 70%/15%/15% of the entire sample set, respectively.

Kidney filtration (predicting normal or impaired eGFR [≥ or < 60ml/min]). Within the 2,515 HUNT3 participants in the CV events program, 87% had eGFR ≥ 60 ml/min/1.73m^2 and 13% < 60 ml/min/1.73m^2 using the creatinine-based CKD EPI equation.[39] An elastic net model was derived and refined on 80%/20% of these participants respectively. Validation was performed using Covance, an independent sample set with 1,029 participants, of whom 94% and 6% had eGFR of ≥ or < 60 ml/min/1.73m^2, respectively.

Body composition (predicting dual-energy X-ray absorption [DEXA] components). Within the Fenland study, 11,471 participants had DEXA scans to assess body fat (%), lean body mass (kg.) and visceral fat (kg.), although the latter was not measurable in 20 subjects. An elastic net linear regression model with continuous output on the same scales as the original measurements was derived, refined and validated on 70%/15%/15% of the total population.

Cardiopulmonary fitness (predicting maximal oxygen uptake on a treadmill [VO_2 max], ml/kg/min). Within the HERITAGE family study, 648 participants completed maximal exercise tests and had blood samples and measures of $V0_2$max. at baseline and after a 20-week exercise-regimen. An unpaired cross-over sampling method (with 50% of samples from participants at baseline and 50% from participants post-exercise) was used to avoid correlation from pairs and increase the observed range of fitness values in the dataset. An elastic net linear regression model was derived/refined/validated on 80%/10%/10% of participants.

b) Modifiable behavioral factors

Alcohol consumption (predicting self-reported consumption above or below UK guidelines [14 units/week for men and women]). Within the Fenland study there were 4,851 women, of whom 11% reported consumption above UK guidelines and 4,803 men of whom 31% reported consumption above guidelines. Elastic net regression models were derived, refined and validated using the same 70%/15%/15% sample distributions; separate models were created for men and women to account for residual error differences associated with participants' sex.

Physical activity (predicting average daily physical activity energy expenditure estimated from combined heart-rate and movement sensing for 1 week [J/day or kcal/day]). This was calculated for the 11,695 participants within the Fenland study with this measure available, using the same 70%/15%/15% fractions for derivation/refinement/validation as for body composition. An elastic net linear regression model was validated with a kcal/day output.

Current cigarette smoking (predicting self-reported questionnaire results). Of the 1,025 Covance participants 15% self-reported as current smokers and 85% prior or never smokers. An elastic net regression model was derived and validated in 80% and 20% of the participants, respectively.

c) Future cardiometabolic health risks

Incident diabetes (predicting future diagnosis in people with pre-diabetes). There were 413 participants within the Whitehall II study at baseline who had pre-diabetic fasting glucose (5.5-6.9mmol/l) or elevated 2h glucose (7.8-11.0 mmol/l) during an oral glucose tolerance test, of whom 23% became diabetic within 10 years. An elastic-net Cox proportional hazards model was

derived on 80% of this pre-diabetic fraction and then validated on a 20% blinded holdout fraction. A decision threshold of ≥3-fold risk (in reference to the average risk score in all Whitehall participants in our study, not just the pre-diabetic fraction) was defined and applied to the pre-diabetic participants.

Incident CV events (predicting any type of first event or cardiovascular death within 5 years). A fully parametric Accelerated Failure Time (AFT) survival model was derived from HUNT3 using a case-cohort design. There were 1,050 cases with an incident "hard" CV event (CV death, myocardial infarction (MI), stroke or hospitalisation for heart failure (HF)) and a random fraction of 1,414 participants selected from the overall cohort, for a total of 2,464 participants. The model was derived and refined on 80% and 20% of HUNT3, respectively. It was validated in Whitehall II using samples from all 101 cases with an incident CV event within 5 years and a random fraction of the cohort (164 participants) without an incident CV event within 5 years. The model is capable of relative risk stratification ranging from ≤1 fold to ≥6 fold compared to low risk individuals at an absolute event rate of <2.5% in 5 years.

Quality control and data normalization

All the samples from all the studies were run on the SomaScan assay and standard SomaLogic normalisation, calibration and data quality control processes were applied as described in detail below.

Quality control over the first year of production for the SomaScan V4 Assay was performed on an average of 2000 samples per week using 24 assay runs which include eleven control replicates from three control lots and a maximum of 85 samples per run. Reference standards, expected values for each protein control replicate lot for each SOMAmer reagent, are derived during assay qualification. Five calibrator replicates per run are used with a reference standard to control for batch effects. Three quality control replicates per run are used with a reference standard to evaluate the accuracy of the assay after data standardization. Standard assay run acceptance criteria require that 85% of the content are accurate to within 20% of the reference; in practice, an average of 96% of the content meets the acceptance standard. The lifetime median precision of the assay over approximately 3000 plasma quality control replicates and 5,207 SOMAmer reagents to protein targets is 6.2% (5^{th} percentile, 3.4%; 95^{th} percentile 19.1%). In addition to standard acceptance criteria, alternate assay summary metrics including overall run signal bias from the reference, calibration scale factor percent outside of 0.6-1.4, quality control replicate five plate running precision, and buffer background or estimated lower limit of detection (eLOD) are monitored for failures or trends over time on a daily basis by Production Bioinformatics and Quality Assurance.

In order to correct for assay-intrinsic variation such as that due to minor variation in sample dilutions by the pipetting robot, we have generally used (in prior studies) typical median normalization – scaling the total fluorescence from a given sample to the median on the same 96-well assay plate. This has two limitations: first, the scaling of any one sample can be impacted by the other samples on the plate which establish the median; second, that there are assay-extrinsic sources of variation in the sample that can affect overall fluorescence, such as sample quality (where plasma from samples with lysed cells is "brighter" because of the leakage of intracellular proteins) and kidney function (where lower filtration rates lead to the elevation of a large proportion of the proteome and again "brighter" samples). In this study, both these limitations

were overcome; the former by using an external reference for the median, rather than the other samples on the same plate, and the latter by restricting the analytes used for normalization to those not impacted by sample quality or disease. This was accomplished by comparing each analyte in a new sample to its counterpart in a reference well-collected "healthy" population (the Covance study described in this manuscript). The subset of analytes in the test sample that were within the expected population distribution of fluorescence in the reference sample were used for calculation of the normalization scale factors.

Statistical analysis and machine learning

Statistical analysis plans for each model were prospectively documented and filed to an auditable software regulatory document vault (Veeva Vault [Veeva Inc.]) prior to analysis, such that the studies became "virtual prospective trials" on retrospectively assayed, archived samples. Sample-size calculations were not carried out prospectively as the likely effect-sizes were hitherto unknown.

Supervised machine learning is the process whereby a computer uses an algorithm applied to data to "train" a model – to derive a fixed equation relating the features chosen to a pre-designated truth standard. The algorithm makes predictions on the training data, the error between predicted and actual values of the truth standard is assessed, and the algorithm is applied iteratively with small changes in parameters to reduce the predictive error. Learning can stop when the algorithm achieves its highest level of performance assessed after cross-validation (multiple iterations of model assessment on different splits in the training data). In this study, the features in a model are the protein measurements and the truth-standards are the health outcomes or measures of behaviour.

When developing predictive models using machine learning techniques, in order to avoid over-fitting, it is common practice to use multiple data sets or fractions of datasets to identify and test or validate the model that has the most reliable predictive capabilities. To this end, we applied the following tactics for splitting data. If the dataset is large (thousands, e.g., Fenland), the data is split into three sets: a derivation set, used for identifying top models through cross-validation [typically a 70% fraction and 5 repeats of 10-fold cross-validation], a refinement set (a second derivation set that allows us to tune the parameters of the top models, typically 15%), and a validation set (a hold-out set that is only used to assess the final model and is not used for model development, typically 15%). If the dataset is smaller (hundreds, e.g., Covance), the data is split into two sets: a derivation set of 80% that again uses cross-validation (typically 10-fold 90%/10% derivation/refinement splits within that 80% fraction) and a validation set of 20% not used for model development. If the dataset contains pairs of samples from the same subjects (e.g., Heritage), the data is split into two sets: a derivation set (70%-80%) and a validation set (10%-20%). Within the derivation set, the model is derived on time point 1 from half the participants and on time point 2 from the other half of the participants (avoiding pairs of samples from the same participants). The model is verified on samples with the opposite time points in the same participants and then validated in the holdout test set data not used for derivation.

Because of the intent to test the extent to which proteins could be a sole information source, demographic features or other laboratory test results were deliberately excluded from the feature

selection process with two exceptions: 1) if the predefined minimum performance could not be reached, the most impactful demographic factor could be added; 2) if the residual errors within a model were related to a demographic feature. In practice, these exceptions were triggered only twice; to include age interactions in the cardiovascular model in order to exceed the performance of the 2013 ACC/AHA atherosclerotic cardiovascular (ASCVD) risk score and to use sex to create separate alcohol models for men and women in order to overcome a sex-related residual error distribution.

The sequence of events for model development was initiated with the definition and documentation of the analysis plan, the truth standard (the variable against which the model is trained) and minimum acceptable performance standard for a model. This was followed by normalization and calibration of the proteins measured in the datasets, the assessment of sample quality, the exclusion of any measured proteins failing to meet the quality control measures described above from model development and the division of the available datasets into training, refinement and validation as shown in **Extended Data Figure 1**.

This was followed by univariate ranking and filtering of proteins' statistical association with the truth standard within the training data and automated application to the training data of several different types of machine learning algorithms with different methodological approaches[40,41]

A semi-automated approach to univariate testing and machine learning analyses was designed to efficiently understand if there is any evidence of signal for the endpoint of interest and to identify the model type that is the best match for the data. The derivation data set was used for univariate tests and preliminary machine learning models.

For continuous measurements (lean body mass, percent body fat, alcohol consumption, energy expenditure from physical activity, visceral adipose tissue, cardiopulmonary fitness VO2 max, weight trajectory and OGTT) we used regression methods. The associations between each analyte and endpoint (lean body mass or percent body fat) on a univariate level, were assessed using the univariate tests for coefficients/importance metrics from linear and robust regression models, Spearman's correlation coefficient and random forest (importance scores calculated). Following the univariate analyses, candidate features were ranked based on false discovery rate (FDR) corrected p-values. At this stage, fairly lenient FDR-corrected p-values of 0.1 or even 0.2 were used to enrich the lists because the truly multivariate models would not depend on univariate significance, but nonetheless there is a need to perform some reduction in dimensionality. Using this subset of features, the following types of models were fit: elastic net linear models (which combines lasso and ridge penalties for feature reduction), support vector machines (which are more robust to outliers) and random forests (a non-linear, tree-based approach)

For dichotomous measurements (liver steatosis, current cigarette smoking and kidney filtration) we used classification methods. The associations between each analyte and endpoint (liver steatosis, cigarette smoking or kidney filtration) on a univariate level, were assessed using t-tests, Mann-Whitney, logistic regression and random Forest (importance scores calculated). The same approach to using univariate FDR-corrected p-value ranking to aid dimensionality reduction was used as for continuous measurements. For the preliminary multivariate models, 5 repeats of 10-fold cross-validation was used in derivation. The following multivariate, machine learning models were then run: elastic net logistic regression model (which combines lasso and ridge penalties for

feature reduction), linear discriminant analysis (similar to Naïve Bayes, but handles correlated features better) and random forest (a non-linear, tree-based approach).

For survival data (diabetes diagnosis within 10 years and cardiovascular primary event risk), we used survival models. The association between each aptamer and the rate of diagnosis (binary outcome and time to event or censoring) on a univariate level was assessed using accelerated failure time (AFT) survival models and cox proportional hazards (PH) models. Again, FDR corrected p-values were used to reduce the number of candidate features to 200. This reduction was done so that the AFT and Cox PH algorithms converged. For the preliminary multivariate models, 5 repeats of 10-fold cross-validation were used. The following multivariate, machine learning models were run: elastic net AFT models (which combines the ridge and LASSO penalties) and cox proportional hazards elastic net models.

Given that the elastic nets routine consistently gave the best result and was ultimately selected for each model, we describe here the processes specific to that algorithm. There are two penalization parameters (variables that add a penalty to each new feature added to a model). The first parameter is associated with penalizing specifically any correlated features and the second is associated with penalizing the overall number of features in the model. Without such penalization, some algorithms would include all the measured proteins. Readers more familiar with the Lasso algorithm may be interested to know that it is equivalent to setting the elastic nets correlated feature penalty to its maximum setting so that these are eliminated[40]. In contrast, elastic nets allows the inclusion of more correlated features as that penalty is reduced. The optimal values of these parameters are determined during the cross-validation phase during which each of the two parameters are varied at fixed increments, and model performance is assessed for each combination of settings. The parameter values associated with the model that has the best predictive performance are then selected as the final values.

During model refinement and prior to validation, advanced feature selection techniques were applied to the features that pass the FDR cutoff, such as forward selection, backward selection, and stability selection. Ensemble methods and approaches were employed to develop the optimal model. In the cross-validation stage, models were optimized based on AUC, sensitivity, and specificity for classification and survival models and adjusted r^2 values for continuous endpoint models. For survival models, the C-Index, Brier score, and net reclassification index[38] were also examined. These predictive metrics were confirmed in analyzing the test or hold out data sets. The number of features within a model was determined simply by the algorithmic selection of the optimal number (e.g. by Elastic Net or Lasso).

The best derived models from the prior step were then examined in more detail. Each of the best models were assessed to determine whether the predefined performance standard could be met without the addition of non-protein features. Additionally, unwanted associations of errors with sex or sample quality were evaluated, and decision-thresholds (or risk-bins) defined to stratify the populations in a simple but informative way.

Validation was performed by applying the final model from derivation and refinement, with its predefined decision thresholds, to the validation data set. Ideally this would be a truly independent replication dataset (such as with the cardiovascular, kidney models). But where such

a matching dataset was not available at this time, a random fraction of the same study (10%-20% depending on study size) with data not used in training was used for testing the predictive accuracy of the model.

The restriction to people with pre-diabetes for the incident diabetes prediction model reflected the intended-use population for the first clinical application and the assumption that a diabetes prognostic model would be highly impacted by pre-diabetes status. Further results of the diabetes, kidney and cardiovascular models are described **in Supplemental Tables 4-6**. All other models were derived in the general study populations (**Extended Data Figures 2-6)** with performance in the participants with pre-diabetes (typically >30%) confirmed when possible.

Data availability statement

Pre-existing data access policies for each of the five parent cohort studies specify that research data requests can be submitted to each steering committee; these will be promptly reviewed for confidentiality or intellectual property restrictions and will not unreasonably be refused. Individual level patient or protein data may further be restricted by consent, confidentiality or privacy laws/considerations. These policies apply to both clinical and proteomic data.

References

37. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. *BMJ* 2016; **353**: i3139.
38. Ganna A, Reilly M, de Faire U, Pedersen N, Magnusson P, Ingelsson E. Risk prediction measures for case-cohort and nested case-control designs: an application to cardiovascular disease. *Am J Epidemiol* 2012; **175**(7): 715-24.
39. Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. *Annals of internal medicine* 2009; **150**(9): 604-12.
40. Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." Journal of the royal statistical society: series B (statistical methodology) 67.2 (2005): 301-320.
41. Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society: Series B (Methodological) 58.1 (1996): 267-288.