nature medicine

LETTERS

https://doi.org/10.1038/s41591-019-0665-2

Plasma protein patterns as comprehensive indicators of health

Stephen A. Williams[1,12]*, Mika Kivimaki[2], Claudia Langenberg[3], Aroon D. Hingorani[4,5,6], J. P. 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], Robert Kirk DeLisle[1], Yolanda Hagar[1], Michael Hinterberg[1], Rachel Ostroff[1], Sophie Weiss[1], Peter Ganz[11,12] and Nicholas J. Wareham[3,12]

Proteins are effector molecules that mediate the functions of genes[1,2] and modulate comorbidities[3–10], behaviors and drug treatments[11]. They represent an enormous potential resource for personalized, 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, percentage body fat, visceral fat mass, lean body mass, cardiopulmonary 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.

As populations worldwide are increasingly affected by multimorbidity and avoidable chronic health conditions, the need to prevent illness is increasing[17]. In response, healthcare providers have instituted preventative medicine programs. For example, the UK 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 tools key to making 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 individualized 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 for reduced healthcare utilization 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 personalized, systemic and data-driven medicine.

Proteins regulate biological processes and can integrate the effects of genes with those of the environment, age, comorbidities, behaviors and drugs[2]. There are about 19,000 human genes coding for approximately 30,000 proteins[25]. Of these, up to 2,200 proteins enter the bloodstream 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 hypothesized 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 for simultaneous evaluation of multiple health issues.

In this proof-of-concept study based on five observational cohorts in 16,894 participants, we evaluated the ability of the scanning of ~5,000 proteins in each plasma sample to simultaneously

[1]SomaLogic, Inc., Boulder, CO, USA. [2]Department of Epidemiology and Public Health, University College London, London, UK. [3]MRC Epidemiology Unit, University of Cambridge, Cambridge, UK. [4]Institute of Cardiovascular Science, University College London, London, UK. [5]University College London, British Heart Foundation Research Accelerator, London, UK. [6]Health Data Research UK, London, UK. [7]Massachusetts Veterans Epidemiology and Research Information Center, Veterans Affairs Boston Healthcare System, Boston, MA, USA. [8]Pennington Biomedical Research Center, Louisiana State University, Baton Rouge, LA, USA. [9]HUNT Research Center and K. G. Jebsen Center for Genetic Epidemiology, Faculty of Medicine and Health Sciences, NTNU–Norwegian University of Science and Technology, Trondheim, Norway. [10]Department of Exercise Science, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA. [11]Division of Cardiology, Center of Excellence in Vascular Research, Zuckerberg San Francisco General Hospital, University of California San Francisco, San Francisco, CA, USA. [12]These authors contributed equally: Stephen A. Williams, Peter Ganz, Nicholas J. Wareham. *e-mail: swilliams@somalogic.com

Table 1 | Performance metrics for each protein model and comparators/references in derivation and validation datasets

Model output (truth standard)	Action	Source	Participants (*n*)	Metric	Result
Current health state					
Liver fat: presence/absence (ultrasound)	Derivation: best subject characteristics	Fenland (70%)	7,054	AUC	0.64
	Derivation: 94-protein model	Fenland (70%)	7,054	AUC	0.85
	Validation: 94-protein model	Fenland (15%)	1,512	AUC	0.83
Kidney function: eGFR <60 ml min^{-1} (CKD–EPI equation)	Derivation: 55-protein model	HUNT3 (80%)	2,013	AUC	0.94
	Validation: 55-protein model	Covance (100%)	1,029	AUC	0.94
Body fat: % (DEXA)	Derivation: best subject characteristics	Fenland (70%)	8,030	r^2	0.74
	Derivation: 219-protein model	Fenland (70%)	8,030	r^2	0.92
	Validation: 219-protein model	Fenland (15%)	1,721	r^2	0.92
Lean body mass: kg (DEXA)	Derivation: best subject characteristics	Fenland (70%)	8,030	r^2	0.74
	Derivation: 115-protein model	Fenland (70%)	8,030	r^2	0.83
	Validation: 115-protein model	Fenland (15%)	1,721	r^2	0.82
Visceral fat: kg (DEXA)	Derivation: 96-protein model	Fenland (70%)	8,016	r^2	0.71
	Validation: 96-protein model	Fenland (15%)	1,718	r^2	0.70
Cardiopulmonary fitness: ml kg^{-1} min^{-1} (VO_2 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%)	3,396	AUC	0.89
	Validation: women, 30-protein model	Fenland (15%)	728	AUC	0.86
	Derivation: men, 33-protein model	Fenland (70%)	3,362	AUC	0.83
	Validation: men, 33-protein model	Fenland (15%)	720	AUC	0.82
Weekly physical activity: kJ kg^{-1} d^{-1} (actigraphy and individually calibrated heart rate)	Derivation: 65-protein model	Fenland (70%)	8,187	r^2	0.36
	Validation: 65-protein model	Fenland (15%)	1,754	r^2	0.38
Cigarette smoking: current yes/no (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 3× risk	Derivation: OGTT fasting and 2-h 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 CV event within 5 years (1–6×)	Derivation: ACC/AHA risk factors	HUNT3	2464	*C*-statistic	0.66
	Validation: ACC/AHA risk factors	Whitehall II	265	*C*-statistic	0.65
	Derivation: 13 proteins & age interactions	HUNT3	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

CKD–EPI, Chronic Kidney Disease–Epidemiology Collaboration; DEXA, dual-energy X-ray absorption; eGFR, estimated glomerular filtration rate; OGTT, oral glucose tolerance test. Spearman's correlation coefficient is used for values of r^2. Best subject characteristics models were developed individually for certain measures of current state, and included the highest performing combination of demographics such as age, sex, body mass index (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 nondiabetic group.

capture the individualized 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 predefined health measures; their performance metrics are shown in Table 1 and graphically in Fig. 1. Success was defined as at least equivalent performance of a validated model to the best available comparator (cardiovascular (CV) risk and incident diabetes risk, measured by *C*-statistic and/or net reclassification index (NRI[27,28]) (versus reference American College of Cardiology (ACC)/American Heart Association (AHA) risk score)). Where there was no comparator, success was a high degree of correlation with a truth standard (Spearman correlation

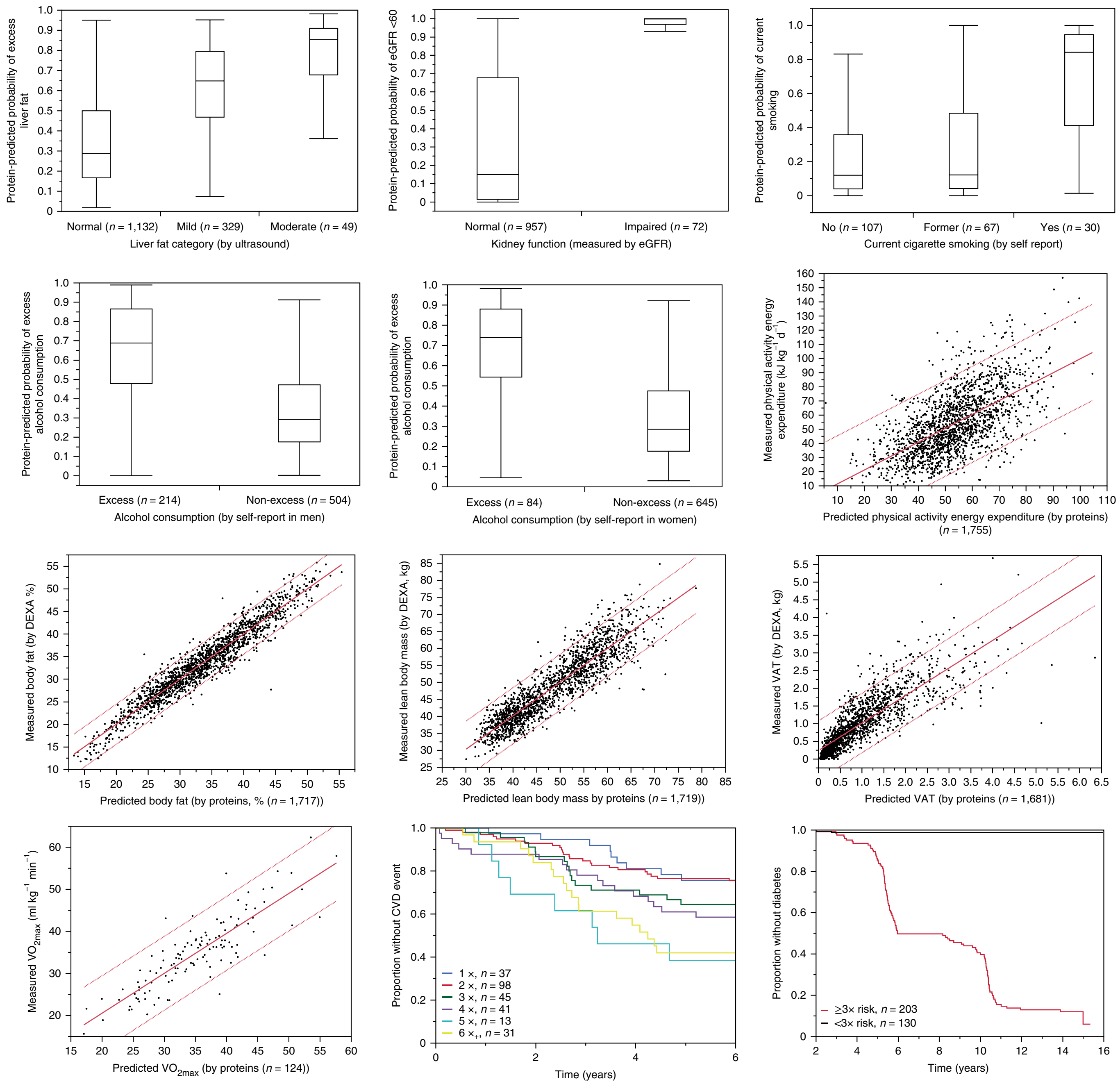


Fig. 1 | Model outputs compared to the truth standards against which they were derived. All panels show the data from the validation sets, except for the diabetes survival model where, for clarity, the Kaplan–Meier curves are shown for the much larger discovery datasets. Box plots are broken down into 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. VAT, visceral adipose tissue. CVD, cardiovascular disease.

coefficients >0.6 (that is, $r^2 > 0.36$) or, for binary measures, an area under the curve for receiver operating characteristic (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, estimated glomerular filtration rate (eGFR) above/below 60 ml min^{-1}: AUC = 0.94; predicting percentage body fat (kg) by DEXA: $r^2 = 0.92$ for proteins and 0.74 for the best clinical model using sex, height and weight; predicting visceral fat (kg) by DEXA: $r^2 = 0.70$; predicting lean body mass (kg) 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_2 max ml min^{-1} kg^{-1}): $r^2 = 0.71$.

For modifiable behaviors, model validation performance metrics are as follows: predicting average daily physical activity energy expenditure (kJ kg^{-1} d^{-1}) 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 per week, 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

Table 2 | Biological plausibility of key proteins in relation to the target physiology for each model; the top three mathematical contributors to each model's output are shown

Issue	Model	Top three 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 nonalcoholic 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 <60 ml min^{-1}	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
		Beta-2 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 percentage body fat; previously shown to enhance the accuracy of BMI estimates of percentage body fat 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	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 subcutaneous adipose tissue
	Visceral fat: kg	Leptin	Produced in adipose tissue and present in higher amounts in subjects with high BMI and percentage body fat; has also been shown to enhance the accuracy of BMI estimates of percentage body fat when DEXA 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
	Cardiopulmonary fitness V0$_2$ max: ml kg^{-1} min^{-1}	Leptin	Produced in adipose tissue and present in higher amounts in subjects with high BMI and percentage body fat; has also been shown to enhance the accuracy of BMI estimates of percentage body fat when DEXA unavailable
		C1QR1 (complement component C1Q receptor)	Part of the innate immune system
		GGH (gamma glutaryl hydrolase)	Regulates intracellular folate; energy production and rebuilding and repair of muscle tissue by physical activity require folate
Modifiable behavioral factors	Alcohol consumption: above/below 14 units (women)	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 (stem cell factor) kit ligand	Mainly involved in hematopoiesis in adults; may relate to alcohol effects on hematopoiesis
	Alcohol consumption: above/below 14 units (men)	SCUB1	Promotes platelet–platelet interaction and is a 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 patients with hepatitis C; associated with advancing fibrosis in fatty liver disease; may relate to alcohol effect on liver
	Weekly physical activity: kcal d^{-1}	Leptin	Produced in adipose tissue and present in higher amounts in subjects with high BMI and percentage body fat
		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 yes/no	SLIK3 (SLIK and NTR protein-like 3)	Involved in neurite growth; no known/clear relation with target physiology
		Secretoglobin family 3 A member 1	Associated with chronic obstructive airways disease and prognosis in non-small-cell lung cancer
		TM108 (transmembrane protein 108)	Genome-wide Association Study linked this with successful smoking cessation

Continued

Table 2 | Biological plausibility of key proteins in relation to the target physiology for each model; the top three mathematical contributors to each model's output are shown (continued)

Issue	Model	Top three proteins	Potential role in target biology
Future metabolic health risks	Conversion from pre-diabetes to diabetes within 10 years, above or below 3× 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-melanocyte-stimulating hormone 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 CV event within 5 years (1–6×)	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 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 types of CVD and associated organ dysfunction

A full listing of all 891 proteins in all models is shown in Supplementary Table 1, and all proteins that were measured are listed in Supplementary Table 3.

within 10 years: accuracy 67% versus 61% for the best oral glucose tolerance model trained in the same participants using combined fasting and peak glucose levels; predicting primary CV events (myocardial infarction, stroke, hospitalization for heart failure or CV death) within 5 years: *C*-statistic of 0.66 and NRI = +0.21 versus the reference 2013 ACC/AHA atherosclerotic CV (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 blood sampling 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) had r^2 values of only ~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 three 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 Supplementary Table 1. The proportionate degree of protein overlap across phenotype models is shown in Table 3. Overall, the degree to which 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 percentage total body fat, visceral fat, physical activity and cardiorespiratory fitness. Within the 110 possible cross-model comparisons in Table 3, only 12 had >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

In this large proteomic study representing a set of prospectively defined analyses of retrospective, archived samples and data from five well-characterized cohorts, approximately 5,000 proteins were measured in nearly 17,000 participants, resulting in ~85 million individual protein measurements. The results were analyzed rigorously by predefined 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 in becoming a sole information source capable of characterizing 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 six current health states, three behaviors and two key future disease risks. The validation of these protein-phenotype models, each consisting of 13–375 protein measurements involving a total of 891 human proteins, provides proof of concept for a scalable, individualized 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 nine different patient appointments and potentially thousands of pounds in costs per patient, as shown in Supplementary Table 2. While some of the models demonstrated high performance (for example, the r^2 of 0.91 for percentage body fat), others had only modest prognostic power (for example, the *C*-statistic of 0.66–0.69 for CV 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 (that is, 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

Table 3 | Proportion (%) of proteins in any one model that overlap with other models

Model	Cigarette smoking	Incident diabetes	Kidney filtration	Primary CV risk	Lean body mass	Alcohol	Percentage body fat	Liver fat	VO$_2$ max	Physical activity	Visceral fat
Cigarette smoking	100 (145)	19	3	2	10	2	17	7	6	3	8
Incident diabetes	7	100 (375)	1	1	5	2	9	7	4	3	7
Kidney filtration	7	5	100 (55)	2	18	0	15	4	9	9	5
Primary CV risk	23	15	8	100 (13)	15	0	**30**	15	15	0	15
Lean body mass	13	16	9	2	100 (115)	4	**32**	17	17	15	24
Alcohol (M/F)	8	19	0	0	11	100 (33/30)	10	19	9	8	20
Percentage body fat	11	16	4	2	17	17	100 (219)	7	16	10	12
Liver fat	11	**29**	2	2	20	6	17	100 (94)	13	13	**38**
VO$_2$ max	7	12	4	2	**26**	3	**30**	10	100 (115)	14	10
Physical activity	6	18	8	0	**26**	4	**32**	23	23	100 (65)	18
Visceral fat	11	**29**	3	2	**29**	7	**27**	**38**	13	13	100 (96)

The model labels in the first column identify a group of proteins within that model; each box within a row to the right of that label contains the percentage of proteins in that model shared with the other models. Overlaps >25% are shown in bold for clarity. The percentages for the separate alcohol models for males (M) and females (F) 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.

and because some of the proteins in our models are leakage proteins that 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 normalizing 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 Supplementary Table 1, and all proteins measured in Supplementary Table 3. The degree of sharing of proteins across phenotype models shown in Table 3 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 percentage body fat, visceral fat, physical activity and cardiopulmonary fitness models. The highest overlap across models was the coincident inclusion of proteins in the models of liver fat, visceral fat and incident diabetes.

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 generalizability 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 SomaScan 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 healthcare systems.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-019-0665-2.

Received: 20 June 2019; Accepted: 23 October 2019;
Published online: 2 December 2019

References

1. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
2. Emilsson, V. et al. Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769–773. (2018).
3. Tasaki, S. et al. Multi-omics monitoring of drug response in rheumatoid arthritis in pursuit of molecular remission. *Nat. Commun.* **9**, 2755 (2018).
4. O'Dwyer, D. N. et al. The peripheral blood proteome signature of idiopathic pulmonary fibrosis is distinct from normal and is associated with novel immunological processes. *Sci. Rep.* **7**, 46560 (2017).
5. Christensson, A. et al. The impact of the glomerular filtration rate on the human plasma proteome. *Proteom. Clin. Appl.* **12**, e1700067 (2018).
6. Ganz, P. et al. Development and validation of a protein-based risk score for cardiovascular outcomes among patients with stable coronary heart disease. *J. Am. Med. Assoc.* **315**, 2532–2541 (2016).
7. Wood, G. C., 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.* **7**, 43238 (2017).
8. Han, Z. et al. Validation of a novel modified aptamer-based array proteomic platform in patients with end-stage renal disease. *Diagnostics (Basel)* **8**, 71 (2018).
9. Menni, C. et al. Circulating proteomic signatures of chronolological age. *J. Gerontol. A* **70**, 809–816 (2014).

10. Thrush, A. et al. Diet-resistant obesity is characterized by a distinct plasma proteomic signature and impaired muscle fiber metabolism. *Int. J. Obes.* **42**, 353–362 (2018).
11. Williams, S. A. et al. Improving assessment of drug safety through proteomics: early detection and mechanistic characterization of the unforeseen harmful effects of torcetrapib. *Circulation* **137**, 999–1010 (2018).
12. Rohloff, J. C. et al. Nucleic acid ligands with protein-like side chains: modified aptamers and their use as diagnostic and therapeutic agents. *Mol. Ther. Nucleic Acids* **3**, e201 (2014).
13. Gold, L. et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS ONE* **5**, e15004 (2010).
14. Brody, E. et al. Life's simple measures: unlocking the proteome. *J. Mol. Biol.* **422**, 595–606 (2012).
15. Kim, C. H. et al. Stability and reproducibility of proteomic profiles measured with an aptamer-based platform. *Sci. Rep.* **8**, 8382 (2018).
16. Candia, J. et al. Assessment of variability in the SOMAscan assay. *Sci. Rep.* **7**, 14248 (2017).
17. Collaborators GBDRF, Forouzanfar, M. H. 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* **386**, 2287–2323 (2015).
18. Maruthappu, M. Delivering triple prevention: a health system responsibility. *Lancet Diabetes Endocrinol.* **4**, 299–301 (2016).
19. Robson, J. et al. The NHS Health Check in England: an evaluation of the first 4 years. *BMJ Open* **6**, e008840 (2016).
20. Valabhji, J. et al. Efficacy and effectiveness of screen and treat policies in prevention of type 2 diabetes: systematic review and meta-analysis of screening tests and interventions. *BMJ* **356**, i6538 (2017).
21. Middleton, K. R., Anton, S. D. & Perri, M. G. Long-term adherence to health behavior change. *Am. J. Lifestyle Med.* **7**, 395–404 (2013).
22. Dimitrov, D. V. Medical internet of things and big data in healthcare. *Health Inf. Res.* **22**, 156–163 (2016).
23. Flores, M., Glusman, G., Brogaard, K., Price, N. D. & Hood, L. P4 medicine: how systems medicine will transform the healthcare sector and society. *Per. Med.* **10**, 565–576 (2013).
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.* **19**, 389–397. (2016).
25. Ezkurdia, I. et al. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum. Mol. Genet.* **23**, 5866–5878 (2014).
26. Lin, H. et al. Discovery of a cytokine and its receptor by functional screening of the extracellular proteome. *Science* **320**, 807–811 (2008).
27. Harrell, F. E. Jr. *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. *Stat. Med.* **27**, 157–172 (2008).
29. Fielding, C. M. & Angulo, P. Hepatic steatosis and steatohepatitis: are they really two distinct entities? *Curr. Hepatol. Rep.* **13**, 151–158 (2014).
30. Yki-Jarvinen, H. Non-alcoholic fatty liver disease as a cause and a consequence of metabolic syndrome. *Lancet Diabetes Endocrinol.* **2**, 901–910. (2014).
31. Shuster, A., Patlas, M., Pinthus, J. H. & Mourtzakis, M. The clinical importance of visceral adiposity: a critical review of methods for visceral adipose tissue analysis. *Br. J. Radiol.* **85**, 1–10 (2012).
32. Ross, 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* **134**, e653–e699 (2016).
33. de Souza de Silva, C. G. et al. Association between cardiorespiratory fitness, obesity, and health care costs: The Veterans Exercise Testing Study. *Int. J. Obes. (Lond.)* https://doi.org/10.1038/s41366-018-0257-0 (2018).
34. Hobbs, F. D., Jukema, J. W., Da Silva, P. M., McCormack, T. & Catapano, A. L. Barriers to cardiovascular disease risk scoring and primary prevention in Europe. *QJM* **103**, 727–739 (2010).
35. Ostroff, R. M. et al. Unlocking biomarker discovery: large scale application of aptamer proteomic technology for early detection of lung cancer. *PLoS ONE* **5**, e15003 (2010).
36. Ostroff, R. M. et al. Early detection of malignant pleural mesothelioma in asbestos-exposed individuals with a noninvasive proteomics-based surveillance tool. *PLoS ONE* **7**, e46091 (2012).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Study design. We prespecified 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[29–33]. These have been well characterized in at least one of five independent cohort studies as the truth standards for deriving and validating proteomic model predictions: the UK 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 h after collection, a timeframe that 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 program, including the source of the samples, data, model training and replication, is shown in Extended Data Fig. 1. Extended Data Fig. 2 shows details of the five parent cohort studies, and Extended Data Figs. 3–6 the participant characteristics for each model endpoint. The Nature Research Reporting Summary for this study is available as part of the online publication.

Proteomic platform. The modified aptamer binding reagents[12], and SomaScan assay[13] and its performance characteristics[15,16], have previously been described. The annotated menu for all ~5,000 modified-aptamer binding reagents is shown in Supplementary Table 3.

The SomaScan Assay begins in each well of a 96-well plate, as a mix of thousands of slow off-rate modified aptamers (SOMAmer reagents). These are labeled with a 5′ fluorophore, photocleavable linker and biotin and immobilized on streptavidin-coated beads through biotin–streptavidin interaction. A plasma sample from each participant is diluted and added to each well.

Cognate and nonspecific SOMAmer–protein complexes form on the beads. After washing away unbound proteins, captured proteins are labeled with biotin. SOMAmer–protein complexes are released from the beads by photocleavage of the linker with ultraviolet light and incubated in a buffer containing an unlabeled polyanionic competitor. This competes with the nonspecific binding of the ‘incorrect’ protein to any SOMAmer reagents that dissociate rapidly owing to the fast off-rate of such interactions, whereas the cognate (intended) SOMAmer–protein interaction has a much slower off-rate (this is part of the original reagent selection process). This differential in kinetics, coupled with polyanionic competition, represents a second element of specificity (the first being the high affinity, enhanced by modifications to the aptamers), analogous to the effect of adding a second antibody in a conventional immunoassay.

SOMAmer–protein complexes are recaptured on a second set of streptavidin-coated beads through biotin-labeled proteins, followed by additional washing steps that facilitate further removal of nonspecifically bound SOMAmer reagents. SOMAmer reagents are then released from the complex in a denaturing buffer.

For readout, SOMAmer reagents are hybridized to complementary sequences on a DNA microarray chip and quantified by fluorescence. Fluorescence intensity in the SomaScan assay for each reagent is related to the relative availability of the three-dimensional shape-charge epitope on each protein (the binding site of the SOMAmer reagent) in the original sample. This is a reflection of each protein's abundance (concentration), the shape of the protein itself (which may be impacted by a genetic variant or by modification) or by a circulating competitor (physiologic or a therapeutic antibody).

Median intra- and interassay coefficients of variation are ~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 1,612 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 (for example, 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 (that is, proteins with >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 versus 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 dissociation constant (K_d) was within tenfold 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 tenfold, we reported that the reagent binds to the protein(s) other than the SELEX target with ‘at least tenfold weaker affinity’. Although this is a broad statement regarding specific affinity, we do not report exact K_d values because of the high variability observed in both the quality and 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 liquid chromatography–tandem mass spectrometry 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 K_d is determined to be the protein concentration at which 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 to the related protein PPA2, which shares 68% amino acid sequence identity. However, solution-affinity measurements determined that the affinity was greater than tenfold 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% (that is, almost half of the 13%), were products of the same gene with a common epitope (for example, splice variants such as vascular endothelial growth factor 121 and 165 isoforms) or shared subunits in a multi-subunit complex (for example, cyclin-dependent kinase 1/cyclin B1 complex, in which the reagent binds to the cyclin B1 subunit). The remaining ~7% appears 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 ~2 nM, 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 pulldown results), 62% for pairs with greater than tenfold lower affinity but positive pulldown results, and 70% for pairs with similar affinity.

In summary, we have tested binding to related proteins for 1,612 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 tenfold weaker affinity. Specific target enrichment by pulldowns 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 (ref. [1]) and 1,046 (ref. [2]) variants, and orthogonal validation by mass spectrometry has been performed for ~1,000 reagents[2].

In addition to mitigations arising from reagent specificity and affinity, the impact of nonspecific 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, nonspecific binding, and capitalizes on the slow off-rates (disassociation rates) of aptamers from their intended targets.

Derivation and validation of protein-phenotype models. *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 underwent liver ultrasound; 75% had no fat and 25% had mild, moderate or severe fat. An elastic net model was derived, refined and validated in 70, 15 and 15% of the entire sample set, respectively.

Kidney filtration (predicting normal or impaired eGFR (≥ or $<60\,\mathrm{ml\,min^{-1}}$)). Within the 2,515 HUNT3 participants in the CV events program, 87% had eGFR $\geq 60\,\mathrm{ml\,min^{-1}\,1.73\,m^{-2}}$ and 13% $<60\,\mathrm{ml\,min^{-1}\,1.73\,m^{-2}}$ using the creatinine-based CKD–EPI equation[39]. An elastic net model was derived and refined on 80 and 20% of these participants, respectively. Validation was performed using Covance, an independent sample set with 1,029 participants, of whom 93 and 7% had eGFR of ≥ or $<60\,\mathrm{ml\,min^{-1}\,1.73\,m^{-2}}$, respectively.

Body composition (predicting dual-energy X-ray absorption (DEXA) components). Within the Fenland study, 11,471 participants had DEXA scans to assess percentage body fat, lean body mass (kg) and visceral fat (kg), although the last of these 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 and 15%, respectively, of the total population.

Cardiopulmonary fitness (predicting maximal oxygen uptake on a treadmill ($\mathrm{VO_2}$ max), $\mathrm{ml\,kg^{-1}\,min^{-1}}$). Within the HERITAGE Family study, 648 participants completed maximal exercise tests and had blood samples and measures of $\mathrm{VO_2}$ 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 to increase the observed range of fitness values in the dataset. An elastic net linear regression model was derived, refined and validated on 80, 10 and 10% of participants, respectively.

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 distribution; 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 ($\mathrm{kJ\,kg^{-1}\,d^{-1}}$ or $\mathrm{kcal\,d^{-1}}$)). 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 $\mathrm{kcal\,d^{-1}}$ output.

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

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 ($5.5–6.9\,\mathrm{mmol\,l^{-1}}$) or elevated 2-h glucose ($7.8–11.0\,\mathrm{mmol\,l^{-1}}$) 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 risk threshold of greater than threefold (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 CV 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, stroke or hospitalization for heart failure) 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 ≤one- to ≥sixfold compared to low-risk individuals at an absolute event rate of <2.5% in 5 years.

Quality control and data normalization. All samples from all studies were run on the SomaScan assay, and standard SomaLogic normalization, 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 2,000 samples per week using 24 assay runs, which include 11 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 is 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 ~3,000 plasma quality control replicates and 5,207 SOMAmer reagents to protein targets is 6.2% (fifth percentile, 3.4%; 95th percentile, 19.1%). In addition to standard acceptance criteria, alternate assay summary metrics—including overall run signal bias from the reference, calibration scale factor percentage outside of 0.6–1.4, quality control replicate five-plate running precision and buffer background or estimated lower limit of detection—are monitored for failures or trends over time on a daily basis by production bioinformatics and quality assurance.

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 previous 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 that establish the median; second, 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.)) before analysis, such that the studies became 'virtual prospective trials' on retrospectively assayed, archived samples. Sample-size calculations were not carried out prospectively as the probable 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 predesignated 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 behavior.

When developing predictive models using machine learning techniques, to avoid over-fitting it is common practice to use multiple datasets 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, for example, Fenland), the data are split into three sets: a derivation set, used for identifying top models through cross-validation (typically a 70% fraction and five repeats of tenfold 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 holdout set that is used only to assess the final model and is not used for model development, typically 15%). If the dataset is smaller (hundreds, for example, Covance), the data are split into two sets: a derivation set of 80% that again uses cross-validation (typically tenfold 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 (for example, Heritage), the data are split into two sets: a derivation set (80–90%) 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 (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; and (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 CV model to exceed the performance of the 2013 ACC/AHA ASCVD risk score, and to use sex to create separate alcohol models for men and women 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 Fig. 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 understand efficiently whether 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 dataset was used for univariate tests and preliminary machine learning models.

For continuous measurements (lean body mass, percentage body fat, alcohol consumption, energy expenditure from physical activity, visceral adipose tissue, cardiopulmonary fitness VO_2 max, weight trajectory and OGTT) we used regression methods. The associations between each analyte and endpoint (lean body mass or percentage 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 combine LASSO and ridge penalties for feature reduction), support vector machines (which are more robust to outliers) and random forests (a nonlinear, tree-based approach).

For dichotomous measurements (liver fat, 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 utilizing univariate FDR-corrected *P* value ranking to aid dimensionality reduction was used as for continuous measurements. For the preliminary multivariate models, five repeats of tenfold cross-validation were 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 nonlinear, tree-based approach).

For survival data (diabetes diagnosis within 10 years and CV 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 AFT survival models and Cox proportional hazards 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 proportional hazards algorithms converged. For the preliminary multivariate models, five repeats of tenfold cross-validation were used. The following multivariate, machine learning models were run: elastic net AFT models (which combine the ridge and LASSO penalties) and 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 specific penalization of any correlated features, and the second is associated with penalization of 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 before validation, advanced feature selection techniques were applied to the features that passed 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 NRI[38] were also examined. These predictive metrics were confirmed in analyzing the test or holdout datasets. The number of features within a model was determined simply by the algorithmic selection of the optimal number (for example, by elastic net or LASSO).

The best derived models from the previous step were then examined in more detail. Each of the best models was assessed to determine whether the predefined performance standard could be met without the addition of nonprotein 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 dataset. Ideally this would be a truly independent replication dataset (such as with the CV, kidney models). However, 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 CV models are described in Supplementary Tables 4–6. All other models were derived in the general study populations (Extended Data Figs. 2–6), with performance in the participants with pre-diabetes (typically >30%) confirmed when possible.

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

Data availability

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, J. A., Sharp, S. J. & Griffin, S. J. The spectrum effect in tests for risk prediction, screening, and diagnosis. *BMJ* **353**, i3139 (2016).
38. Ganna, A. et al. Risk prediction measures for case-cohort and nested case-control designs: an application to cardiovascular disease. *Am. J. Epidemiol.* **175**, 715–724 (2012).
39. Levey, A. S. et al. A new equation to estimate glomerular filtration rate. *Ann. Intern. Med.* **150**, 604–612 (2009).
40. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* **67**, 301–320 (2005).
41. Tibshirani, R. Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. B* **58**, 267–288 (1996).

Acknowledgements

The Whitehall II study is supported by the UK Medical Research Council UK (no. MR/R024227/1, to M.K.), the US National Institutes on Aging (NIH, nos. US R01AG056477, R01AG062553) to M.K. and the British Heart Foundation (no. RG/16/11/32334) to M.J.S. A.D.H. is a NIHR Senior Investigator and was also supported, in part, by the National Institute for Health Research University College London Hospitals Biomedical Research Centre and the UCL BHF Research Accelerator (AA/18/6/34223). FENLAND (the Fenland study, no. 10.22025/2017.10.101.00001) is funded by UK Medical Research Council (no. MC_UU_12015/1), and N.W. is a NIHR senior investigator. 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, no. R01HL146462 to M.A.S.) and no. HL45670 (HERITAGE, to C.B.). All authors are grateful to all 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 A. Lowell (leader of the SomaLogic assay team), D. Perry for the bioinformatics of quality control, J. Williams for the agreements with the study institutions and J. Zach for clinical data organization and management.

Author contributions

In an academic–industry partnership, SomaLogic, Inc. 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 and advised on the study design for the CV and diabetes models. 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. A.D.H. and J.P.C. were particularly involved in the design, execution and interpretation of the CV risk evaluations.

Competing interests

The SomaLogic co-authors (S.W., L.A., J.A., T.B., J.C., G.D., R.K.D., Y.H., M.H., R.O. and S.W.) were/are all employees of SomaLogic, Inc., which has a commercial interest in the results. N.W. and C.L. declared that SomaLogic, Inc. has given a grant to the University of Cambridge. P.G. is a member of the SomaLogic Medical Advisory board, for which he receives no remuneration of any kind. The remaining authors (M.K., A.H., J.P.C., C.B., C.J., M.S. and M.S.) have no competing interests.

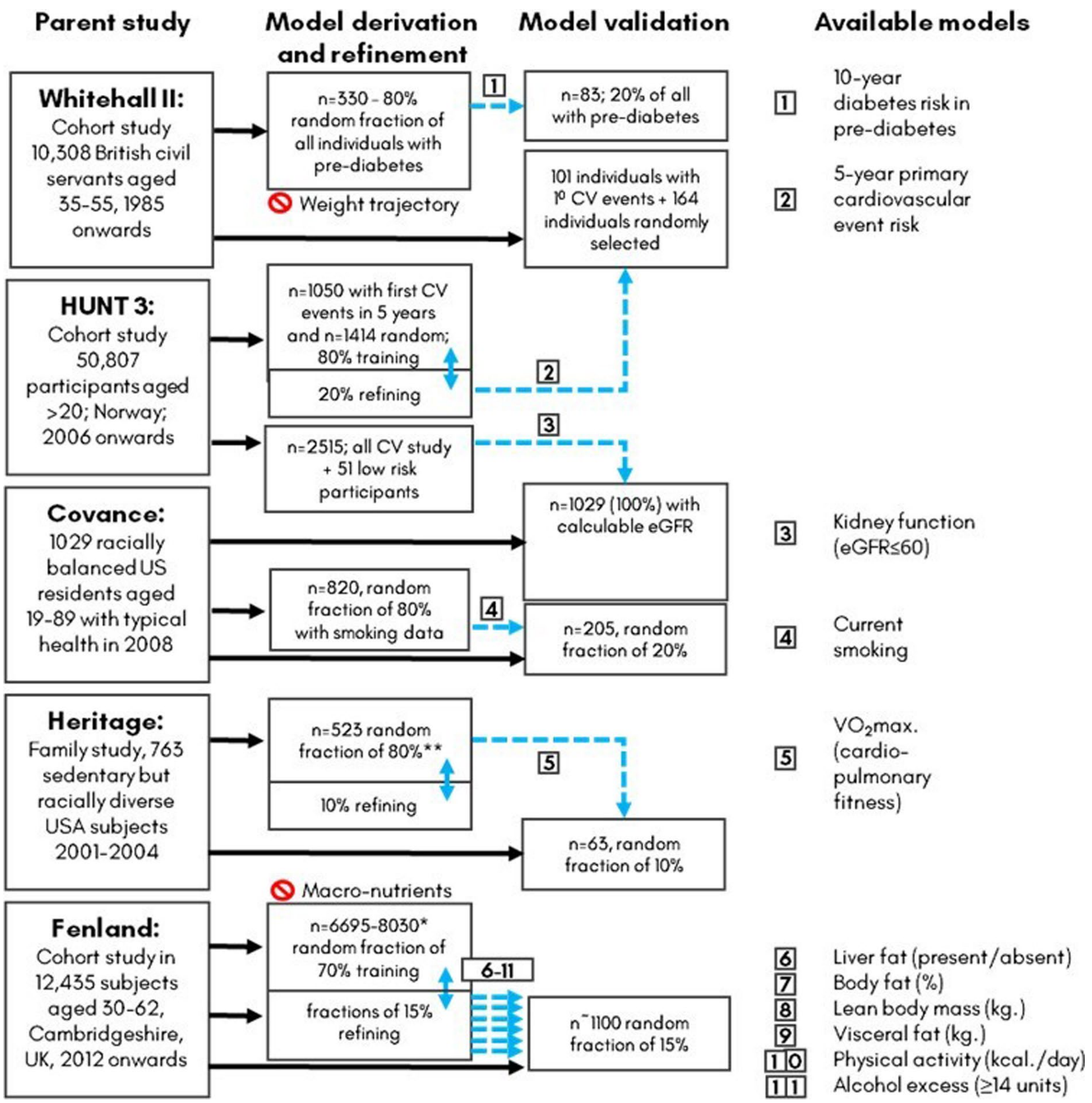
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0665-2.

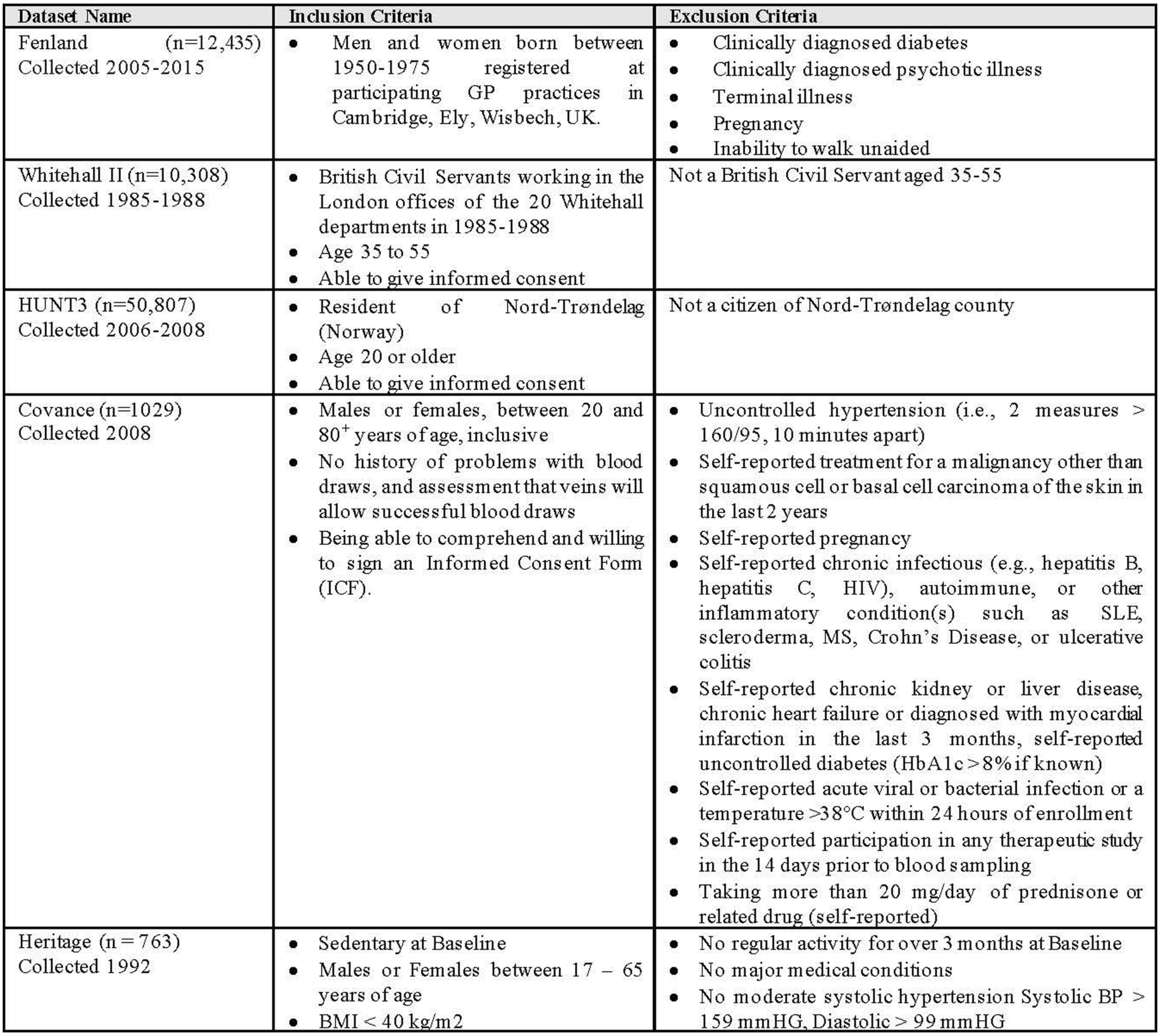
Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0665-2.

Correspondence and requests for materials should be addressed to S.A.W.

Peer review information Jennifer Sargent was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

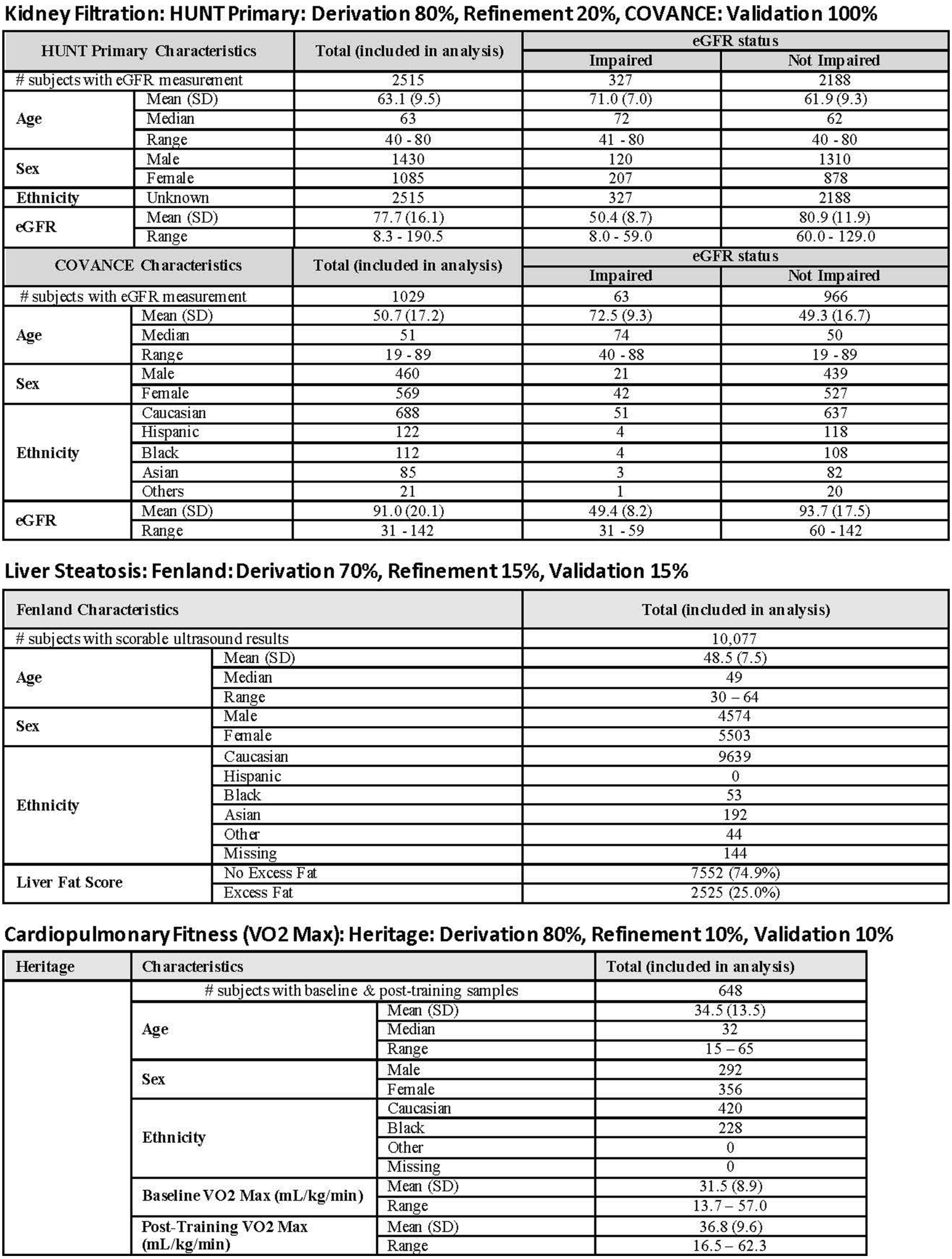

Parent study
Model derivation and refinement
Model validation
Available models
Whitehall II: Cohort study 10,308 British civil servants aged 35-55, 1985 onwards
n=330 - 80% random fraction of all individuals with pre-diabetes
Weight trajectory
1
n=83; 20% of all with pre-diabetes
101 individuals with 1º CV events + 164 individuals randomly selected
1 10-year diabetes risk in pre-diabetes
2 5-year primary cardiovascular event risk
HUNT 3: Cohort study 50,807 participants aged >20; Norway; 2006 onwards
n=1050 with first CV events in 5 years and n=1414 random; 80% training
20% refining
2
n=2515; all CV study + 51 low risk participants
3
Covance: 1029 racially balanced US residents aged 19-89 with typical health in 2008
n=1029 (100%) with calculable eGFR
3 Kidney function (eGFR≤60)
n=820, random fraction of 80% with smoking data
4
n=205, random fraction of 20%
4 Current smoking
Heritage: Family study, 763 sedentary but racially diverse USA subjects 2001-2004
n=523 random fraction of 80%**
10% refining
5
n=63, random fraction of 10%
5 VO2max. (cardio-pulmonary fitness)
Macro-nutrients
Fenland: Cohort study in 12,435 subjects aged 30-62, Cambridgeshire, UK, 2012 onwards
n=6695-8030* random fraction of 70% training
fractions of 15% refining
6-11
n~1100 random fraction of 15%
6 Liver fat (present/absent)
7 Body fat (%)
8 Lean body mass (kg.)
9 Visceral fat (kg.)
10 Physical activity (kcal./day)
11 Alcohol excess (≥14 units)

Extended Data Fig. 1 | Descriptors of parent studies and fractions used for model derivation and validation. Solid black arrows designate how fractions of samples and clinical data were utilized independently; blue dashed arrows designate the validation of finalized models either in new fractions of the same dataset or in independent datasets. eGFR = estimated glomerular filtration rate; VO$_2$max. = maximum rate of oxygen consumption; kg. = kilograms. *For Fenland, the precise numbers available for 70%/15%/15% fractions depended on the numbers of participants with data for each endpoint as follows: n=9654 for self-reported alcohol units, n = 11,471 with DEXA scans for body composition, n=10,077 with ultrasound for liver fat, n=11,695 with individually calibrated heart rate and movement sensing for caloric expenditure due to physical activity. **For HERITAGE the model was trained on the pre-training time point from half the 523 participants and the post training time point from the other half of the participants. The model was tested on samples with the opposite time points in the same participants and finally replicated in the 10% fraction not used for training.



Dataset Name	Inclusion Criteria	Exclusion Criteria
Fenland (n=12,435) Collected 2005-2015	• Men and women born between 1950-1975 registered at participating GP practices in Cambridge, Ely, Wisbech, UK.	• Clinically diagnosed diabetes • Clinically diagnosed psychotic illness • Terminal illness • Pregnancy • Inability to walk unaided
Whitehall II (n=10,308) Collected 1985-1988	• British Civil Servants working in the London offices of the 20 Whitehall departments in 1985-1988 • Age 35 to 55 • Able to give informed consent	Not a British Civil Servant aged 35-55
HUNT3 (n=50,807) Collected 2006-2008	• Resident of Nord-Trøndelag (Norway) • Age 20 or older • Able to give informed consent	Not a citizen of Nord-Trøndelag county
Covance (n=1029) Collected 2008	• Males or females, between 20 and 80^{+} years of age, inclusive • No history of problems with blood draws, and assessment that veins will allow successful blood draws • Being able to comprehend and willing to sign an Informed Consent Form (ICF).	• Uncontrolled hypertension (i.e., 2 measures > 160/95, 10 minutes apart) • Self-reported treatment for a malignancy other than squamous cell or basal cell carcinoma of the skin in the last 2 years • Self-reported pregnancy • Self-reported chronic infectious (e.g., hepatitis B, hepatitis C, HIV), autoimmune, or other inflammatory condition(s) such as SLE, scleroderma, MS, Crohn's Disease, or ulcerative colitis • Self-reported chronic kidney or liver disease, chronic heart failure or diagnosed with myocardial infarction in the last 3 months, self-reported uncontrolled diabetes (HbA1c > 8% if known) • Self-reported acute viral or bacterial infection or a temperature >38°C within 24 hours of enrollment • Self-reported participation in any therapeutic study in the 14 days prior to blood sampling • Taking more than 20 mg/day of prednisone or related drug (self-reported)
Heritage (n = 763) Collected 1992	• Sedentary at Baseline • Males or Females between 17 – 65 years of age • BMI < 40 kg/m2	• No regular activity for over 3 months at Baseline • No major medical conditions • No moderate systolic hypertension Systolic BP > 159 mmHG, Diastolic > 99 mmHG

Extended Data Fig. 2 | Details of the 5 parent cohort studies.



Kidney Filtration: HUNT Primary: Derivation 80%, Refinement 20%, COVANCE: Validation 100%

HUNT Primary Characteristics		Total (included in analysis)	eGFR status	
			Impaired	Not Impaired
# subjects with eGFR measurement		2515	327	2188
Age	Mean (SD)	63.1 (9.5)	71.0 (7.0)	61.9 (9.3)
	Median	63	72	62
	Range	40 - 80	41 - 80	40 - 80
Sex	Male	1430	120	1310
	Female	1085	207	878
Ethnicity	Unknown	2515	327	2188
eGFR	Mean (SD)	77.7 (16.1)	50.4 (8.7)	80.9 (11.9)
	Range	8.3 - 190.5	8.0 - 59.0	60.0 - 129.0

COVANCE Characteristics		Total (included in analysis)	eGFR status	
			Impaired	Not Impaired
# subjects with eGFR measurement		1029	63	966
Age	Mean (SD)	50.7 (17.2)	72.5 (9.3)	49.3 (16.7)
	Median	51	74	50
	Range	19 - 89	40 - 88	19 - 89
Sex	Male	460	21	439
	Female	569	42	527
Ethnicity	Caucasian	688	51	637
	Hispanic	122	4	118
	Black	112	4	108
	Asian	85	3	82
	Others	21	1	20
eGFR	Mean (SD)	91.0 (20.1)	49.4 (8.2)	93.7 (17.5)
	Range	31 - 142	31 - 59	60 - 142

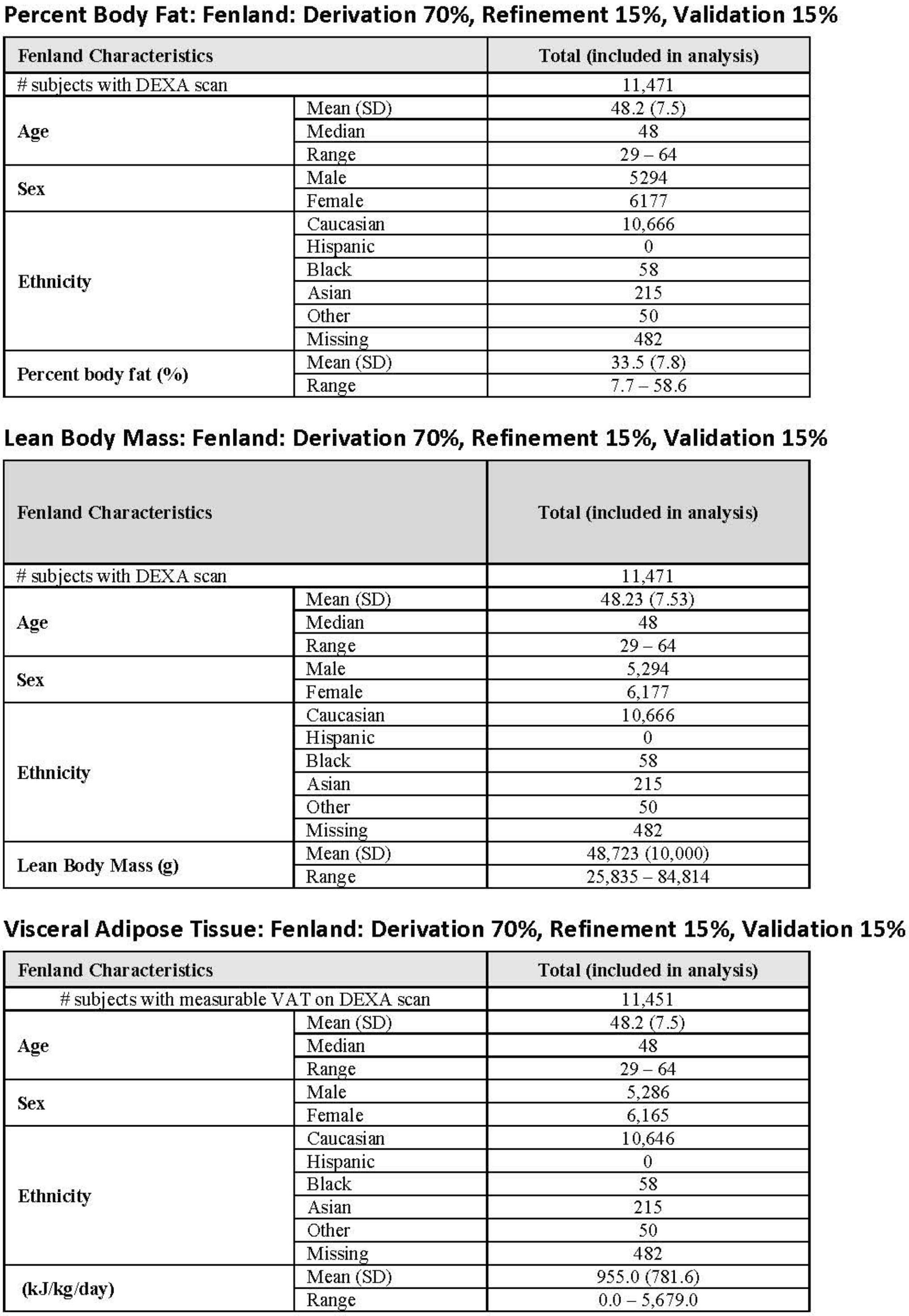
Liver Steatosis: Fenland: Derivation 70%, Refinement 15%, Validation 15%

Fenland Characteristics		Total (included in analysis)
# subjects with scorable ultrasound results		10,077
Age	Mean (SD)	48.5 (7.5)
	Median	49
	Range	30 – 64
Sex	Male	4574
	Female	5503
Ethnicity	Caucasian	9639
	Hispanic	0
	Black	53
	Asian	192
	Other	44
	Missing	144
Liver Fat Score	No Excess Fat	7552 (74.9%)
	Excess Fat	2525 (25.0%)

Cardiopulmonary Fitness (VO2 Max): Heritage: Derivation 80%, Refinement 10%, Validation 10%

Heritage	Characteristics		Total (included in analysis)
	# subjects with baseline & post-training samples		648
	Age	Mean (SD)	34.5 (13.5)
		Median	32
		Range	15 – 65
	Sex	Male	292
		Female	356
	Ethnicity	Caucasian	420
		Black	228
		Other	0
		Missing	0
	Baseline VO2 Max (mL/kg/min)	Mean (SD)	31.5 (8.9)
		Range	13.7 – 57.0
	Post-Training VO2 Max (mL/kg/min)	Mean (SD)	36.8 (9.6)
		Range	16.5 – 62.3

Extended Data Fig. 3 | Participant characteristics for current health state models.



Percent Body Fat: Fenland: Derivation 70%, Refinement 15%, Validation 15%

Fenland Characteristics		Total (included in analysis)
# subjects with DEXA scan		11,471
Age	Mean (SD)	48.2 (7.5)
	Median	48
	Range	29 – 64
Sex	Male	5294
	Female	6177
Ethnicity	Caucasian	10,666
	Hispanic	0
	Black	58
	Asian	215
	Other	50
	Missing	482
Percent body fat (%)	Mean (SD)	33.5 (7.8)
	Range	7.7 – 58.6

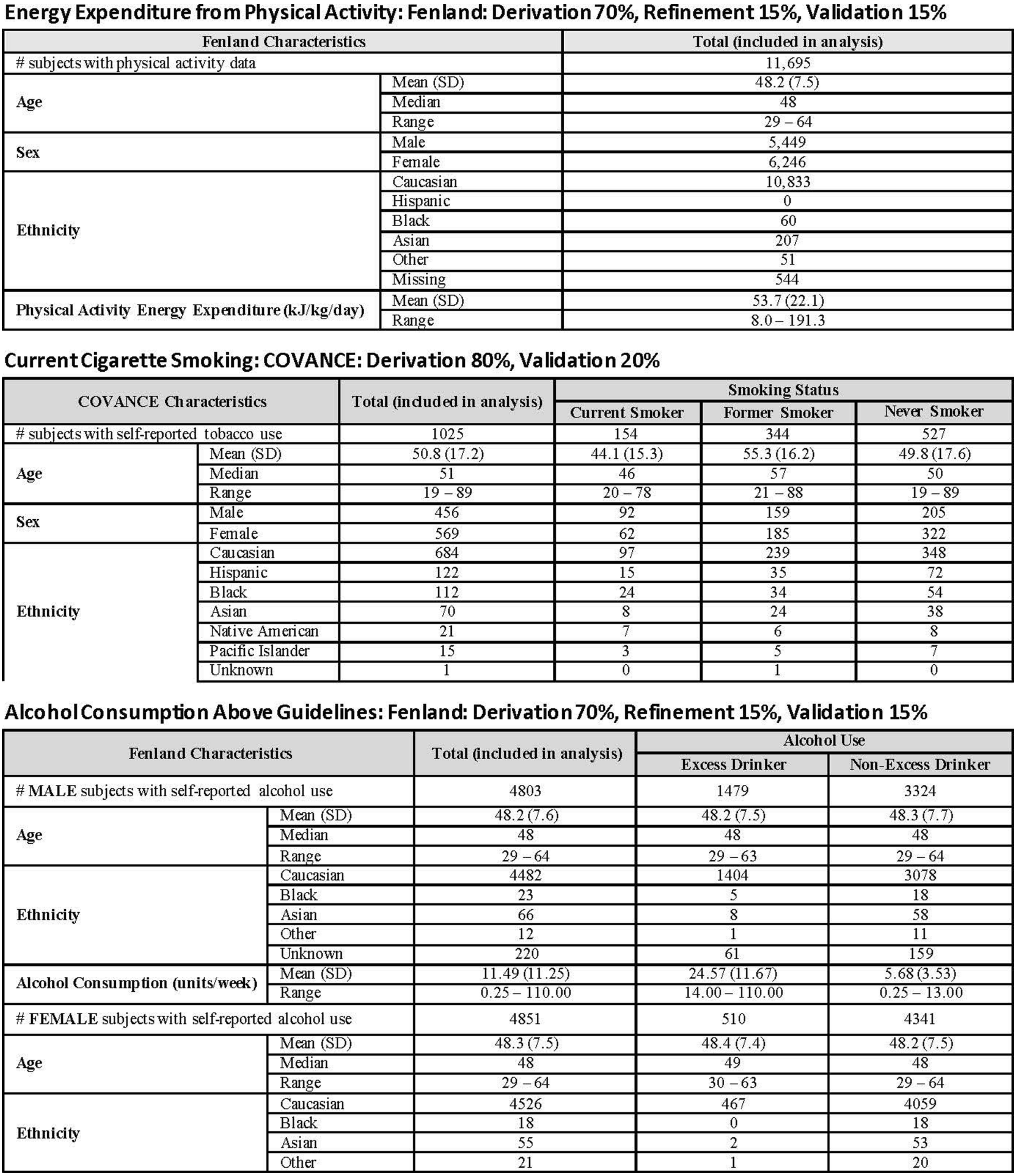
Lean Body Mass: Fenland: Derivation 70%, Refinement 15%, Validation 15%

Fenland Characteristics		Total (included in analysis)
# subjects with DEXA scan		11,471
Age	Mean (SD)	48.23 (7.53)
	Median	48
	Range	29 – 64
Sex	Male	5,294
	Female	6,177
Ethnicity	Caucasian	10,666
	Hispanic	0
	Black	58
	Asian	215
	Other	50
	Missing	482
Lean Body Mass (g)	Mean (SD)	48,723 (10,000)
	Range	25,835 – 84,814

Visceral Adipose Tissue: Fenland: Derivation 70%, Refinement 15%, Validation 15%

Fenland Characteristics		Total (included in analysis)
# subjects with measurable VAT on DEXA scan		11,451
Age	Mean (SD)	48.2 (7.5)
	Median	48
	Range	29 – 64
Sex	Male	5,286
	Female	6,165
Ethnicity	Caucasian	10,646
	Hispanic	0
	Black	58
	Asian	215
	Other	50
	Missing	482
(kJ/kg/day)	Mean (SD)	955.0 (781.6)
	Range	0.0 – 5,679.0

Extended Data Fig. 4 | Participant characteristics for current state body composition models.



Energy Expenditure from Physical Activity: Fenland: Derivation 70%, Refinement 15%, Validation 15%

Fenland Characteristics		Total (included in analysis)
# subjects with physical activity data		11,695
Age	Mean (SD)	48.2 (7.5)
	Median	48
	Range	29 – 64
Sex	Male	5,449
	Female	6,246
Ethnicity	Caucasian	10,833
	Hispanic	0
	Black	60
	Asian	207
	Other	51
	Missing	544
Physical Activity Energy Expenditure (kJ/kg/day)	Mean (SD)	53.7 (22.1)
	Range	8.0 – 191.3

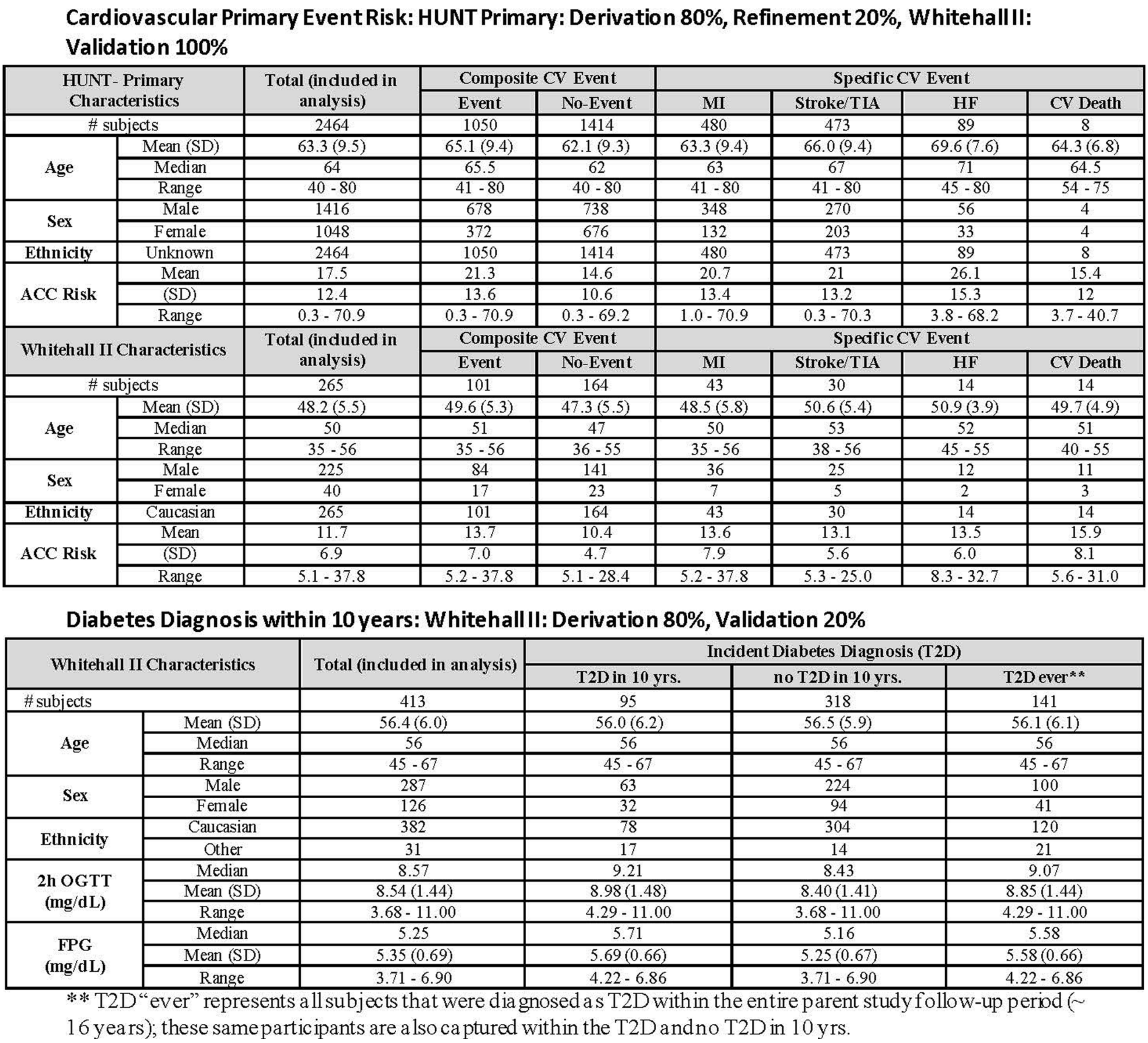
Current Cigarette Smoking: COVANCE: Derivation 80%, Validation 20%

COVANCE Characteristics		Total (included in analysis)	Smoking Status		
			Current Smoker	Former Smoker	Never Smoker
# subjects with self-reported tobacco use		1025	154	344	527
Age	Mean (SD)	50.8 (17.2)	44.1 (15.3)	55.3 (16.2)	49.8 (17.6)
	Median	51	46	57	50
	Range	19 – 89	20 – 78	21 – 88	19 – 89
Sex	Male	456	92	159	205
	Female	569	62	185	322
Ethnicity	Caucasian	684	97	239	348
	Hispanic	122	15	35	72
	Black	112	24	34	54
	Asian	70	8	24	38
	Native American	21	7	6	8
	Pacific Islander	15	3	5	7
	Unknown	1	0	1	0

Alcohol Consumption Above Guidelines: Fenland: Derivation 70%, Refinement 15%, Validation 15%

Fenland Characteristics		Total (included in analysis)	Alcohol Use	
			Excess Drinker	Non-Excess Drinker
# **MALE** subjects with self-reported alcohol use		4803	1479	3324
Age	Mean (SD)	48.2 (7.6)	48.2 (7.5)	48.3 (7.7)
	Median	48	48	48
	Range	29 – 64	29 – 63	29 – 64
Ethnicity	Caucasian	4482	1404	3078
	Black	23	5	18
	Asian	66	8	58
	Other	12	1	11
	Unknown	220	61	159
Alcohol Consumption (units/week)	Mean (SD)	11.49 (11.25)	24.57 (11.67)	5.68 (3.53)
	Range	0.25 – 110.00	14.00 – 110.00	0.25 – 13.00
# **FEMALE** subjects with self-reported alcohol use		4851	510	4341
Age	Mean (SD)	48.3 (7.5)	48.4 (7.4)	48.2 (7.5)
	Median	48	49	48
	Range	29 – 64	30 – 63	29 – 64
Ethnicity	Caucasian	4526	467	4059
	Black	18	0	18
	Asian	55	2	53
	Other	21	1	20

Extended Data Fig. 5 | Participant characteristics for modifiable behavioral factors models.



Cardiovascular Primary Event Risk: HUNT Primary: Derivation 80%, Refinement 20%, Whitehall II: Validation 100%

HUNT- Primary Characteristics		Total (included in analysis)	Composite CV Event		Specific CV Event			
			Event	No-Event	MI	Stroke/TIA	HF	CV Death
# subjects		2464	1050	1414	480	473	89	8
Age	Mean (SD)	63.3 (9.5)	65.1 (9.4)	62.1 (9.3)	63.3 (9.4)	66.0 (9.4)	69.6 (7.6)	64.3 (6.8)
	Median	64	65.5	62	63	67	71	64.5
	Range	40 - 80	41 - 80	40 - 80	41 - 80	41 - 80	45 - 80	54 - 75
Sex	Male	1416	678	738	348	270	56	4
	Female	1048	372	676	132	203	33	4
Ethnicity	Unknown	2464	1050	1414	480	473	89	8
ACC Risk	Mean	17.5	21.3	14.6	20.7	21	26.1	15.4
	(SD)	12.4	13.6	10.6	13.4	13.2	15.3	12
	Range	0.3 - 70.9	0.3 - 70.9	0.3 - 69.2	1.0 - 70.9	0.3 - 70.3	3.8 - 68.2	3.7 - 40.7
Whitehall II Characteristics		Total (included in analysis)	Composite CV Event		Specific CV Event			
			Event	No-Event	MI	Stroke/TIA	HF	CV Death
# subjects		265	101	164	43	30	14	14
Age	Mean (SD)	48.2 (5.5)	49.6 (5.3)	47.3 (5.5)	48.5 (5.8)	50.6 (5.4)	50.9 (3.9)	49.7 (4.9)
	Median	50	51	47	50	53	52	51
	Range	35 - 56	35 - 56	36 - 55	35 - 56	38 - 56	45 - 55	40 - 55
Sex	Male	225	84	141	36	25	12	11
	Female	40	17	23	7	5	2	3
Ethnicity	Caucasian	265	101	164	43	30	14	14
ACC Risk	Mean	11.7	13.7	10.4	13.6	13.1	13.5	15.9
	(SD)	6.9	7.0	4.7	7.9	5.6	6.0	8.1
	Range	5.1 - 37.8	5.2 - 37.8	5.1 - 28.4	5.2 - 37.8	5.3 - 25.0	8.3 - 32.7	5.6 - 31.0

Diabetes Diagnosis within 10 years: Whitehall II: Derivation 80%, Validation 20%

Whitehall II Characteristics		Total (included in analysis)	Incident Diabetes Diagnosis (T2D)		
			T2D in 10 yrs.	no T2D in 10 yrs.	T2D ever**
# subjects		413	95	318	141
Age	Mean (SD)	56.4 (6.0)	56.0 (6.2)	56.5 (5.9)	56.1 (6.1)
	Median	56	56	56	56
	Range	45 - 67	45 - 67	45 - 67	45 - 67
Sex	Male	287	63	224	100
	Female	126	32	94	41
Ethnicity	Caucasian	382	78	304	120
	Other	31	17	14	21
2h OGTT (mg/dL)	Median	8.57	9.21	8.43	9.07
	Mean (SD)	8.54 (1.44)	8.98 (1.48)	8.40 (1.41)	8.85 (1.44)
	Range	3.68 - 11.00	4.29 - 11.00	3.68 - 11.00	4.29 - 11.00
FPG (mg/dL)	Median	5.25	5.71	5.16	5.58
	Mean (SD)	5.35 (0.69)	5.69 (0.66)	5.25 (0.67)	5.58 (0.66)
	Range	3.71 - 6.90	4.22 - 6.86	3.71 - 6.90	4.22 - 6.86

** T2D “ever” represents all subjects that were diagnosed as T2D within the entire parent study follow-up period (~ 16 years); these same participants are also captured within the T2D and no T2D in 10 yrs.

Extended Data Fig. 6 | Participant characteristics for future metabolic health risks models.

nature research

Corresponding author(s): Dr. Stephen A Williams

Last updated by author(s): 2019-10-01

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection in this study.
Data analysis	Analysis and machine learning was performed using the open source R programming language version 3.4

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

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.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	Formal sample sizing was not performed; even though the studies were clearly over-powered for univariate significance testing (with hundreds of proteins signigicant, even when corrected for false discovery rates) the effect size in practice is the output of the machine-learning derived models which has not been established or estimated in advance.
Data exclusions	Pre-established quality controls (out of range hybridization, median scale or normalization scale factors) were used to eliminate protein analytes with sample handling or normalization issues from consideration into model development.
Replication	The manuscript clearly describes the replication process. All models were initially developed through multi-fold cross-validation of 90% train and 10% test. Then a separate pre-defined fraction or independent dataset was used for validation. In the largest study, a third predefined verification fraction was also interpolated between the two. All attempts at replication were successful.
Randomization	This is mostly not relevant as whole cohorts were included, although the fractions identified for verification and validation were randomly and prospectively identified.
Blinding	The samples were all assayed and protein data acquired without any connection made to the clinical database.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
☒	□	Human research participants
☒	□	Clinical data

Methods

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