scientific reports

Check for updates

OPEN

Cross-sectional metabolic subgroups and 10-year follow-up of cardiometabolic multimorbidity in the UK Biobank

Anwar Mulugeta[1], Elina Hyppönen[1], Mika Ala-Korpela[2,3,4] & Ville-Petteri Mäkinen[1,2,5]✉

We assigned 329,908 UK Biobank participants into six subgroups based on a self-organizing map of 51 biochemical measures (blinded for clinical outcomes). The subgroup with the most favorable metabolic traits was chosen as the reference. Hazard ratios (HR) for incident disease were modeled by Cox regression. Enrichment ratios (ER) of incident multi-morbidity versus randomly expected co-occurrence were evaluated by permutation tests; ER is like HR but captures co-occurrence rather than event frequency. The subgroup with high urinary excretion without kidney stress (HR = 1.24) and the subgroup with the highest apolipoprotein B and blood pressure (HR = 1.52) were associated with ischemic heart disease (IHD). The subgroup with kidney stress, high adiposity and inflammation was associated with IHD (HR = 2.11), cancer (HR = 1.29), dementia (HR = 1.70) and mortality (HR = 2.12). The subgroup with high liver enzymes and triglycerides was at risk of diabetes (HR = 15.6). Multimorbidity was enriched in metabolically favorable subgroups (3.4 ≤ ER ≤ 4.0) despite lower disease burden overall; the relative risk of co-occurring disease was higher in the absence of obvious metabolic dysfunction. These results provide synergistic insight into metabolic health and its associations with cardiovascular disease in a large population sample.

Abbreviations

IHD	Ischemic heart disease
SOM	Self-organizing map
HR	Hazard ratio
WHO	World Health Organization
HES	Hospital episode statistics
ICD	International classification of diseases
BMI	Body mass index
SD	Standard deviation
HDL	High-density lipoprotein
OR	Odds ratio
MetS	Metabolic syndrome
NCEP ATP	National cholesterol education program adult treatment panel
VLDL	Very-low-density lipoprotein

The top 10 global causes for death included ischemic heart disease (IHD, 1st), stroke (2nd), dementias (5th), respiratory cancers (6th) and diabetes (7th) according to the Global Health Estimates 2016 report by the WHO. Much of this disease burden is attributed to obesity-associated metabolic dysfunction that increases the risk of cardiometabolic diseases[1], multiple cancers[2] and dementia[3] in ageing individuals. These associations are supported by experimental studies of ageing[4]. There is thus a causal rationale why population subgroups with poor metabolic health bear a higher aggregate burden of multiple chronic diseases later in life.

[1]Australian Centre for Precision Health, Unit of Clinical and Health Sciences, University of South Australia, Adelaide, Australia. [2]Computational Medicine, Faculty of Medicine, University of Oulu and Biocenter Oulu, Oulu, Finland. [3]Center for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu, Finland. [4]NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland. [5]Computational and Systems Biology Program, Precision Medicine Theme, South Australian Health and Medical Research Institute, Adelaide, Australia. ✉email: ville-petteri.makinen@sahmri.com

The predictive power of metabolic profiling has been demonstrated in human populations[5,6], yet the practical value may be limited for an individual patient[7]. In fact, risk factors for common diseases tend to have small individual impact and *vice versa*[8], and prediction models for cardiovascular disease have modest performance at individual level[9] despite clear statistical association at population level. We propose that creating subgroups of metabolically similar individuals may represent a goldilocks solution that combines the robustness of population-wide statistics while retaining easy-to-interpret analogy to observable personal metabolic profiles for more individualised insight compared to traditional epidemiological modelling[10,11].

In a typical scenario, people with similar profiles are grouped together and the aggregate rates of disease outcomes are compared between the subgroups. For example, we developed subgroups of diabetic complication burden in 2008[12] and validated them in 2018 with new previously unseen data on clinical outcomes[13]. A recent investigation of body mass and the burden of 400 common diseases in the UK Biobank found clusters with distinct diagnostic profiles, and the authors also provided a comprehensive review of the literature related to biomedical subgrouping[14]. These studies are highly valuable since they produce quantitative descriptors of population health (biomarker profiles) that contain clues on how to reduce adverse long-term outcomes (biological interpretation of the biomarker profiles).

The first aim of this study was to define biologically meaningful metabolic subgroups in a large representative sample of a human population. The second aim was to identify those subgroups that carry the greatest aggregate risk of cardiometabolic and other diseases. To achieve the aims, we used data from the UK Biobank that includes half a million participants, 51 anthropometric and biochemical variables and ten years of follow-up data[15]. We also introduced the self-organizing map (SOM) as a powerful technique to determine metabolic subtypes[11]. Our framework is unique since it combines multi-variate data with expert consensus to infer metabolic subgroups from biochemical profiles while being blinded to clinical diagnoses during model fitting (robust statistics). We interpret these subgroups as prototypical "individuals" that can be used as the basis for targeted public health initiatives, recruitment of representative samples for clinical trials and for identifying synergistic patterns of cardiometabolic risk factors.

Materials and methods

The UK Biobank is a prospective cohort study of over 500,000 participants aged 37–73 years recruited between 2006 and 2010[15]. UK Biobank has approval from the North West Multi-centre Research Ethics Committee (URL: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics). The participants are volunteers who have provided written informed consent. No personal details were used in this study. Data storage and analyses were conducted according to the material transfer agreement between South Australian Health and Medical Institute and the UK Biobank. This study was designed and implemented according to UK Biobank project plan #29890.

Participants provided baseline information, physical measures and blood and urine samples and information on disease outcomes was obtained through register linkage, including Hospital Episode Statistics (HES), cancer and national death registries. Biochemical measures are described online (URL: https://biobank.ndph.ox.ac.uk/crystal/crystal/docs/serum_biochemistry.pdf). The dataset included in this study comprised 153,731 men and 176,177 women of white British ancestry (Supplementary Figure S1).

The self-organizing map (SOM) is an artificial neural network approach that is designed to facilitate the detection of multi-variable patterns in complex datasets[16]. The result of the analysis is a two-dimensional layout where individuals with similar profiles are close together on the map and thus can be assigned to the same subgroup by visually observable proximity. In this respect, the SOM is a type of clustering analysis, however, in our framework the final step of assigning subgroup labels to individuals is done by human consensus (study authors) rather than by mathematical rules[11]. This is particularly important for population-based datasets such as the UK Biobank that do not have a strong clustered structure due to the broad spectrum of volunteers.

The SOM was trained according to anthropometric and biochemical data; the health outcomes were excluded from the training set to prevent overfitting. The authors were blinded to disease outcomes until after the SOM subgroups were defined. A module-based approach was adopted to avoid collinearity artefacts. First, Spearman correlations were calculated for all pairs of variables. Next, the pairs of variables that were considered collinear ($R^2 > 50\%$) were collected into a network topology. Lastly, we used an agglomerative network algorithm to define modules of collinear variables[17] and principal component analysis to collapse each module into a single data column.

The training set was adjusted for age and sex, centered by mean and scaled by standard deviation. The SOM was created with default settings except for smoothness = 2.0 for a more conservative fit. The quality control tests for the SOM are shown in Supplementary Figure S2 (Plots A–L). We verified that every district of the map was populated (sample density ≥ 1293 across the map, Plot A), the model fit was sufficient (residuals below 3 SDs, Plot B) and that the coverage of available data was high (≥ 92% across the map, Plot C). We tested if centering by mean for those under medication affected the map colorings, but we observed no substantial changes in the regional patterns. The map patterns were not confounded by statins (original vs. adjusted LDL, Plots D–F), by anti-hypertensives (systolic BP, Plots G–I) or by diabetic medications (glucose, Plots J–L). To assess the influence of geographical location, we grouped the assessment centers according to latitude into ≤ 51°, 52° and 53°, 54° and ≥ 55°. We did not observe substantial stratification by assessment center location (Supplement Figure S2).

Clinical diagnoses were based on three-character ICD-10 codes (International Classification of Diseases, version 10) from registers of primary care, hospital inpatients, deaths and self-reported medical conditions. Combinations of ICD-10 codes for cardiometabolic diseases are described in Supplementary Table S1. Rheumatoid arthritis, dementia and cancer were included as examples of non-cardiometabolic diseases. Cancer cases were identified using ICD-9 and ICD-10 codes from the cancer registry. The first occurrence of a disease at or before

baseline was considered prevalent, new cases after baseline considered incident. Vitality status was obtained from mortality registers censored to 26th April 2020.

Associations with prevalent outcomes were modelled by logistic regression and incident outcomes by Cox regression. Both model types were adjusted for age, sex and assessment center. One subgroup was chosen as the reference and the other subgroups were compared against the reference one-by-one. Cardiometabolic multimorbidity was defined as having at least two out of the four conditions (IHD, stroke, diabetes or hypertension).

Observed multimorbidity was evaluated against simulated null distributions of random co-occurrence of diseases. Firstly, a binary table was created where participants were organized as rows and diseases as columns. To obtain a random sample, the binary columns were randomly shuffled, the aggregate disease tallies were counted for each row and the proportion of rows with a disease tally greater than one was recorded. The process was repeated 10,000 times to create the null distribution. The *P*-value was estimated by comparing the non-shuffled proportion of multimorbidity against the null distribution. Confidence intervals were estimated similarly, except with bootstrapping instead of permutations applied to the binary table. Statistical analyses were conducted with Stata (version 16.0, College Station, TX, StataCorp LP) and R v3.5.0 (URL: https://www.R-project.org/) with the Numero library v1.4[11].

Ethics statement. UK Biobank has approval from the North West Multi-centre Research Ethics Committee (URL: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics).

Consent for publication. UK Biobank participants are volunteers who have provided written informed consent. No personal details were used in this study.

Results

Correlation structure between metabolic variables. The characteristics of the study population are listed in Supplementary Table S2. The mean age was 57 years (SD 8 years), most individuals were overweight (BMI mean 27.4 kg/m^2, SD 4.8 kg/m^2) and 20,094 (6.1%) individuals died during a mean follow-up of 10.8 years. We investigated 51 metabolic variables (34 biochemical, 15 anthropometric and two blood pressures) that were reduced to 33 SOM inputs based on collinearity (details in Methods, see also Supplementary Figure S3). The final correlation structure is shown in Fig. 1.

Primer on the self-organizing map. The concept of the SOM is illustrated in Fig. 2. Each participant is represented by their individual preprocessed metabolic profile (Fig. 2A, 33 input dimensions). The Kohonen algorithm[16] is applied to project the high-dimensional input data onto the vertical and horizontal coordinates (two-dimensional layout in Fig. 2B). On the scatter plot, proximity between two participants means that their full multivariable input data are similar as well (Fig. 2C). However, scatter plots are cumbersome for large datasets and difficult to interpret in the absence of distinct clusters. The SOM circumvents these challenges by dividing the plot area into districts. To show statistical patterns, each district is colored according to the average value of a single biomarker or, in the case of morbidity, the local prevalence or incidence of a disease (Fig. 2D, E). The connection between proximity on the canvas and similarity of full profile works the same way on the SOM as it does on a scatter plot. Therefore, selecting a region on the SOM is the same as selecting a subgroup of individuals with mutually similar profiles of input data (Fig. 2F).

The technical details of the SOM have been published previously. In particular, we highlight extensive supplementary documents in four earlier papers that introduce the basic mathematical concepts and discuss the differences between textbook examples of clustered data and the nature of clinical cohort data as the motivation behind the SOM framework[11,17–19]. We also recommend the vignette in the Numero R package (URL: https://cran.r-project.org/web/packages/Numero/vignettes/intro.html) as a practical guide on how to construct a SOM.

Metabolic subgroups. IHD is the most common global cause for death[20] and causally connected to lipoproteins[21]. For this reason, we used the patterns of the apolipoprotein B module, triglycerides and the HDL module as the starting point for subgrouping (Fig. 3A, G, M). We identified map regions that captured the characteristic combinations of features for individuals that had the highest apolipoprotein B score (Subgroup I, top-left part of Fig. 3A–F), elevated triglycerides (Subgroups II and III, bottom-left quadrant of Fig. 3G–L), and the highest HDL score (Subgroup IV, top part of Fig. 3M–P).

Subgroup I was characterized by the combination of high apolipoprotein B score (Fig. 3A), high systolic blood pressure (Fig. 3B), high rheumatoid factor (Fig. 3C) and adequate glycemic control (Fig. 3D). Biomarkers of kidney disease were not elevated (Fig. 3E, F). The second and third subgroups featured elevated triglycerides (Fig. 3G) and high body fat score (Fig. 3H), however, Subgroup II was characterized by high liver enzymes (Fig. 3I–K) whereas Subgroup III had higher C-reactive protein (Fig. 3L). The highest HDL module scores (Subgroup IV) were observed together with the highest vitamin D (Fig. 3N) and bilirubin (Fig. 3O) and low estradiol (Fig. 3P, V). These individuals were the leanest (Fig. 3H).

The highest estradiol values were observed on the left side (Subgroup V, Fig. 3P, V) and Subgroup V also showed the highest testosterone in men (Fig. 3W) and sex-hormone binding globulin for both sexes (Fig. 3R). Sex dimorphism was pronounced; estradiol was fivefold higher in women, and testosterone was tenfold higher in men and we verified that the relative SOM patterns for women under and over the age of 51[22] were not disrupted by menopause (Supplementary Figure S4). The map area at the bottom (Subgroup VI) was characterized by high urinary excretion biomarkers without albuminuria (Fig. 3E, S, T) and these individuals had higher insulin-like growth factor Z-scores compared to the neighboring Subgroups III and V (Fig. 3U).

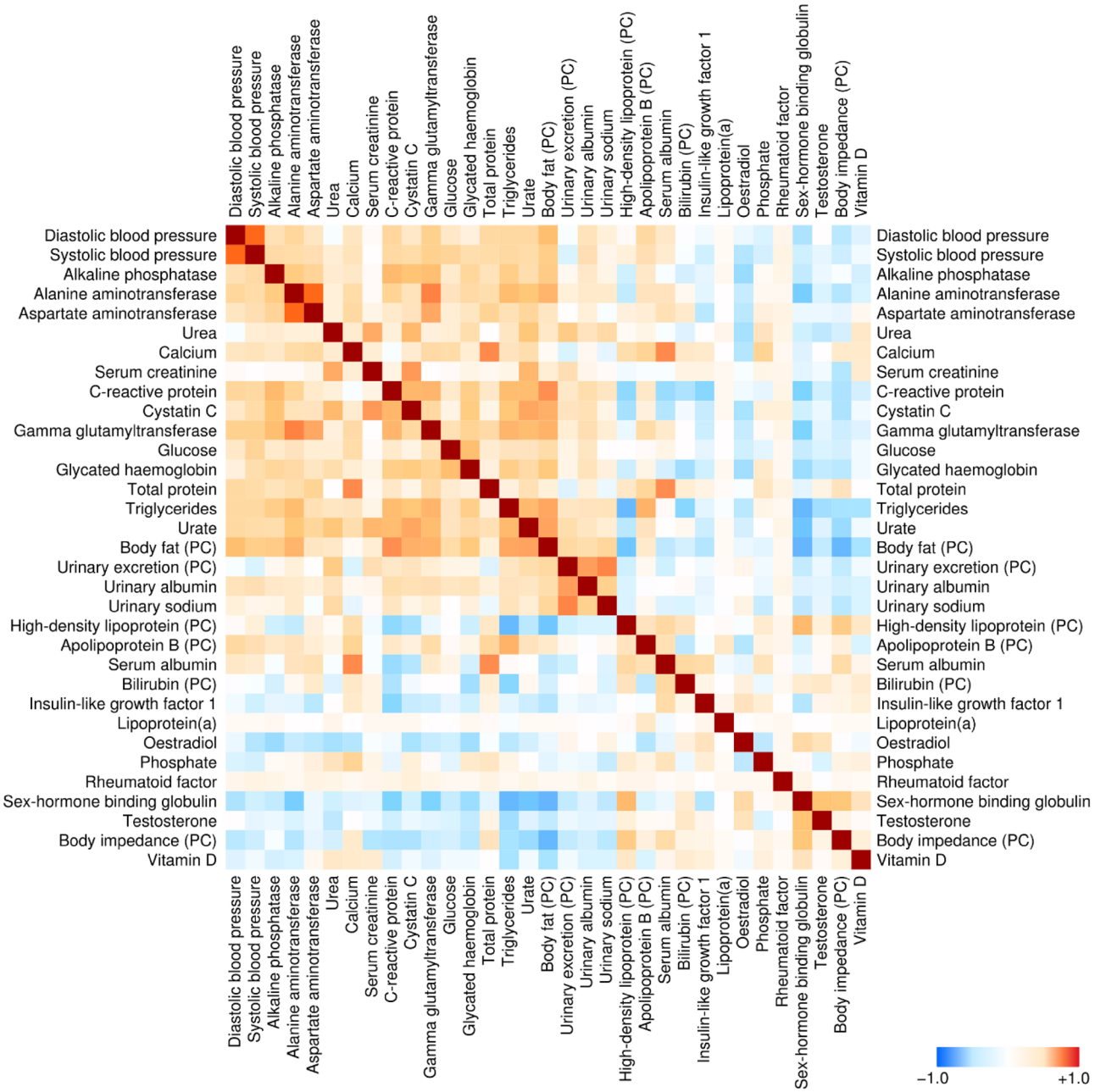

Figure 1. Spearman correlations between anthropometric and biochemical features that comprised the training set for the self-organizing map (adjusted for age and sex). Highly collinear variables were collapsed into the principal component score (PC) prior to correlation analysis.

Succinct descriptive labels based on selected biomarkers were assigned to the subgroups for easier reading (Fig. 4). Unadjusted map colorings in physical units are included in Supplementary Figures S5 and S6. Numerical descriptions of the subgroups are available in Supplementary Table S3.

Disease prevalence and incidence by subgroup. The highest prevalence of IHD was observed in Subgroup III (Fig. 5A). Diabetes prevalence varied the most across the map with small percentages for Subgroups IV and V, but substantially higher in Subgroups II and III (Fig. 5B). The pattern for hypertension was close to that of diabetes (Fig. 5C), but there were also individuals in Subgroup I who had hypertension (see also blood pressure in Fig. 4G). The prevalence of rheumatoid arthritis, dementia and cancer was higher in Subgroup III (Fig. 5D–F). Subgroup IV was associated with the lowest overall burden of disease and was chosen as the control subgroup. The subgroups were similar with respect to age, sex and follow-up time (Fig. 5U–X).

Odds and hazard ratios of diseases between the subgroups are shown in Fig. 5G–T and confidence intervals and *P*-values are available in Supplementary Tables S4 and S5. Subgroup III was associated with the highest prevalence of ischemic heart disease (7.5%, OR = 2.9), hypertension (19.3%, OR = 3.7), rheumatoid arthritis (2.3%, OR = 2.9) and cancer (9.1%, OR = 1.4). High incidence was observed for IHD (9.6 per 1000 person years,

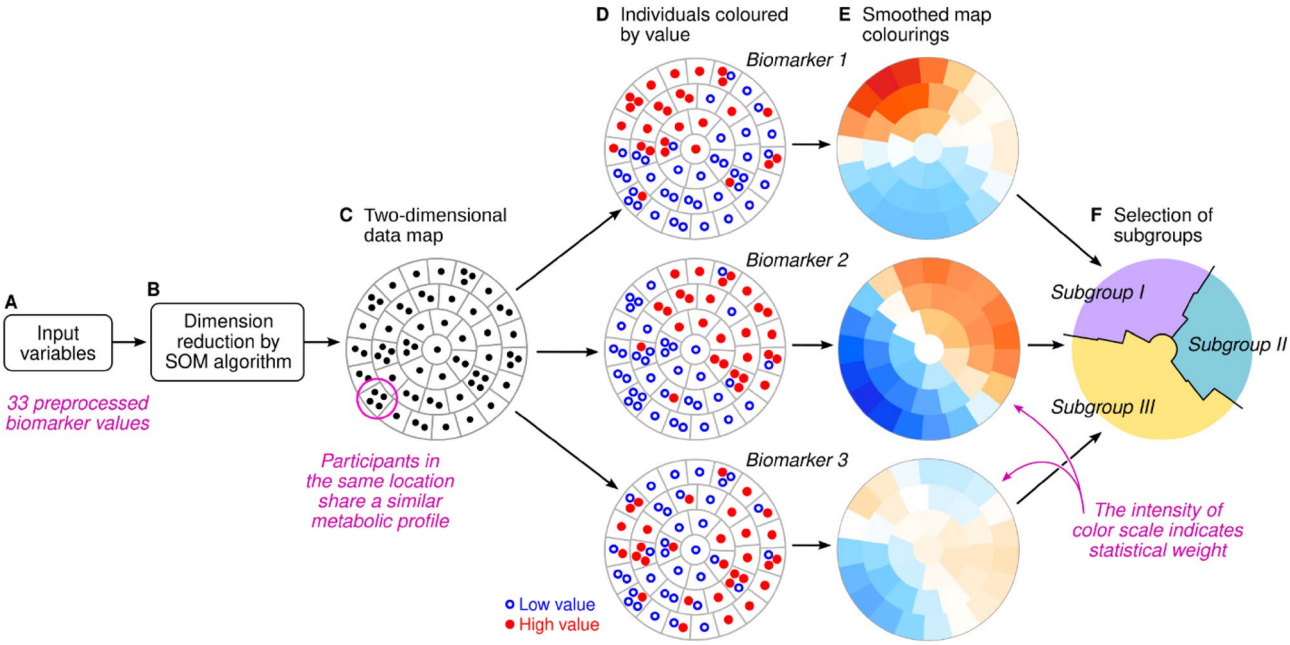

Figure 2. Schematic illustration of the subgrouping procedure. We used the self-organizing map (SOM) algorithm to project high-dimensional data onto a two-dimensional canvas that is divided into districts (**A–C**). The data points can be colored based on the observed values of any variable (**D**). In this study, the statistical weight of regional patterns was encoded in smoothed pseudo-colour representations of the observed values (**E**). The map colorings were used as visual guides to assign map districts and the participants therein into mutually exclusive subgroups (**F**).

HR = 2.1) and the highest incidence for rheumatoid arthritis (1.6, HR = 2.53), cancer (12.8, HR = 1.3), stroke (2.6, HR = 1.9) and mortality (13.4, HR = 2.1).

The prevalence of diabetes was the highest in Subgroup II at 16.7% (OR = 12.6) and the incidence was 14.3 per 1000 person years (HR = 15.8). The incidence of ischemic heart disease in Subgroup II was the same as in Subgroup III (9.6 vs. 9.7, $P > 0.05$). There were no differences in the prevalence of dementia (0.13% vs. 0.14%, $P > 0.05$) or the incidence of dementia (1.4 vs. 1.5, $P > 0.05$) between Subgroups II and III.

Metabolic syndrome and multimorbidity. The metabolic syndrome (MetS) was developed to capture synergistic features associated with high cardiovascular risk[23,24]. The SOM patterns for MetS classification (NCEP ATP III) are shown in Fig. 6A–F and numerical results are available in Supplementary Table S6. High MetS prevalence was observed in Subgroup II (64.2%) and Subgroup III (57.8%) and the lowest in Subgroup IV (5.7%).

The MetS combines risk factors, but we also investigated the combination of established morbidities. The burden of multimorbidity depends on the frequencies of the diseases in the population: if two diseases become more frequent, the random chance of having both increases. For example, younger individuals have fewer diseases compared to older individuals (Fig. 6G, split by the median age of 58 years). This difference in disease frequencies leads to a difference in multimorbidity by mathematics alone (the null model, see Methods). However, the observed excess beyond the null model (i.e. enrichment) was greater in younger individuals (Fig. 6H), which means that having one cardiometabolic disease as a young person increases the probability of having another disease more than it would for an older person.

The highest frequency of multimorbidity was observed in Subgroups II (prevalence 9.8%, incidence 7.7%) and III (prevalence 9.4%, incidence 6.1%) and the lowest in Subgroups IV (2.0%, 1.9%) and V (2.5%, 1.8%). We defined the enrichment ratio (ER) as the ratio between the observed number of individuals with ≥ 2 diseases versus the number predicted by the null model. Multimorbidity was enriched in all subgroups (Fig. 6D, E and Supplementary Tables S7 and S8), with the highest ratios observed in Subgroups IV (prevalent ER = 4.22, incident ER = 4.00), and the lowest in Subgroup II (prevalent ER = 1.74, incident ER = 2.01).

Discussion

Metabolic dysfunction is inextricably linked with ageing demographics and the global obesity pandemic and comes with potentially grave health implications for populations and individuals alike[1–3]. To understand the phenomenon better, we introduced data-driven metabolic subgrouping of the UK Biobank as a model of metabolic diversity (the first aim of the study) and investigated subgroup-specific prevalence and incidence of multiple clinical outcomes (the second aim of the study).

We defined six metabolic subgroups based on the SOM of the UK Biobank. The first three subgroups captured the patterns of classical IHD risk factors and the obesity pandemic (Subgroups I-III). The liver-associated Subgroup II was predictive of diabetes and IHD, which fits with the concept of fatty and insulin resistant liver

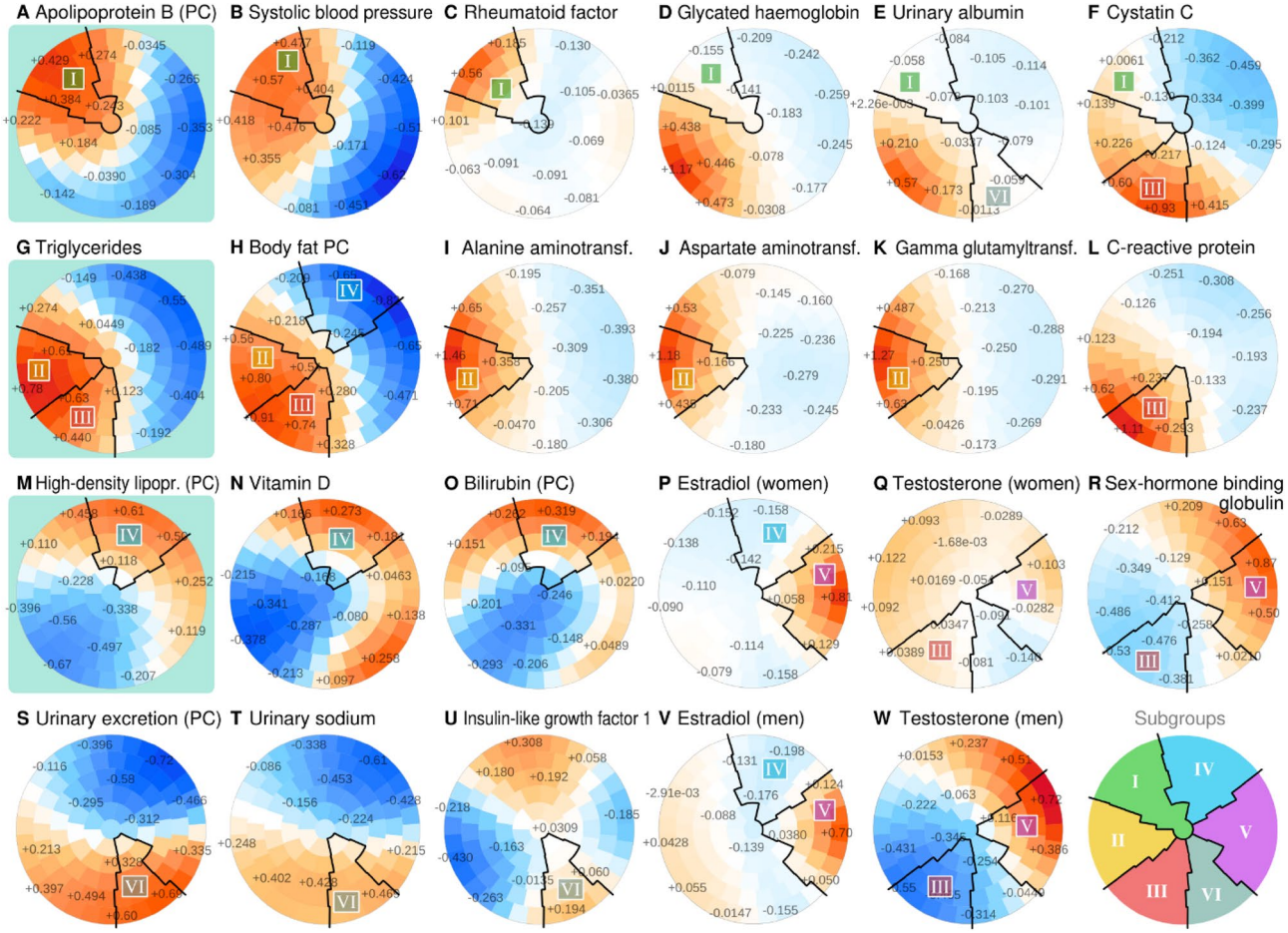

Figure 3. The SOM subgrouping procedure applied to the UK Biobank. In each plot, the same participants reside in the same district. The colors of the districts indicate the regional deviation from the global mean, with color intensity adjusted according to how much the variable contributed to the structure of the map. The numbers on the districts indicate the smoothed mean Z-score of the participants.

as a key player in VLDL-HDL dyslipidemia, insulin resistance and type 2 diabetes[25,26]. The inflammatory and kidney stressed Subgroup III was associated with the highest mortality and overall chronic morbidity (including IHD). This pattern is also compatible with the literature[27,28]. The distinction between the liver and kidney is a notable biological insight from the SOM analysis—for example, the popular definitions of the MetS do not capture the liver-kidney spectrum[24].

We identified a subgroup with elevated sex hormones (Subgroup V). These individuals had a low burden of diabetes and morbidity, which fits the Rotterdam Study[29] and other evidence on insulin resistance[30]. Yet the Rotterdam study also reported that high estradiol in women may indicate increased diabetes risk. Furthermore, we observed multi-fold variation in absolute levels between men, women, young and old that may confound disease associations, as also noted by other studies[31,32]. Longitudinal studies with multiple time points of hormones may be necessary to understand how hormonal levels indicate and predict metabolic dysfunction.

Subgroup VI was characterized by elevated serum urea, elevated serum and urine creatinine and high urinary electrolytes. There was no clear indication of kidney stress nor high morbidity. The biochemical pattern is compatible with the expected effects of habitual high-protein diet[33]. Subgroup VI may also capture a haemodynamic or a fluid balance aspect of metabolic health[34]. Incidental circumstances during sample collection is another possibility: as there is only one biochemical time point, acute illness or other stressors before the baseline visit may have confounded systemic metabolism and resulted in atypical findings for multiple affected and correlated biomarkers.

Obesity and unfavourable lifestyle are risk factors for multimorbidity[1,35]. However, the previous studies did not consider the confounding increase in co-occurrence when the frequency of diseases increases. We observed a synergistic enrichment for cardiometabolic multimorbidity in all subgroups. The most likely explanation is intertwined etiology, partly due to pleiotropic genetic variants and environmental exposures and partly due to secondary effects between the diseases themselves such as the mechanical stress on the vasculature from hypertension[36] or toxicity from excessive glycation in diabetes[37]. Another explanation could be diagnostic procedures: if one disease is detected, it is easier to look for and establish the presence of another.

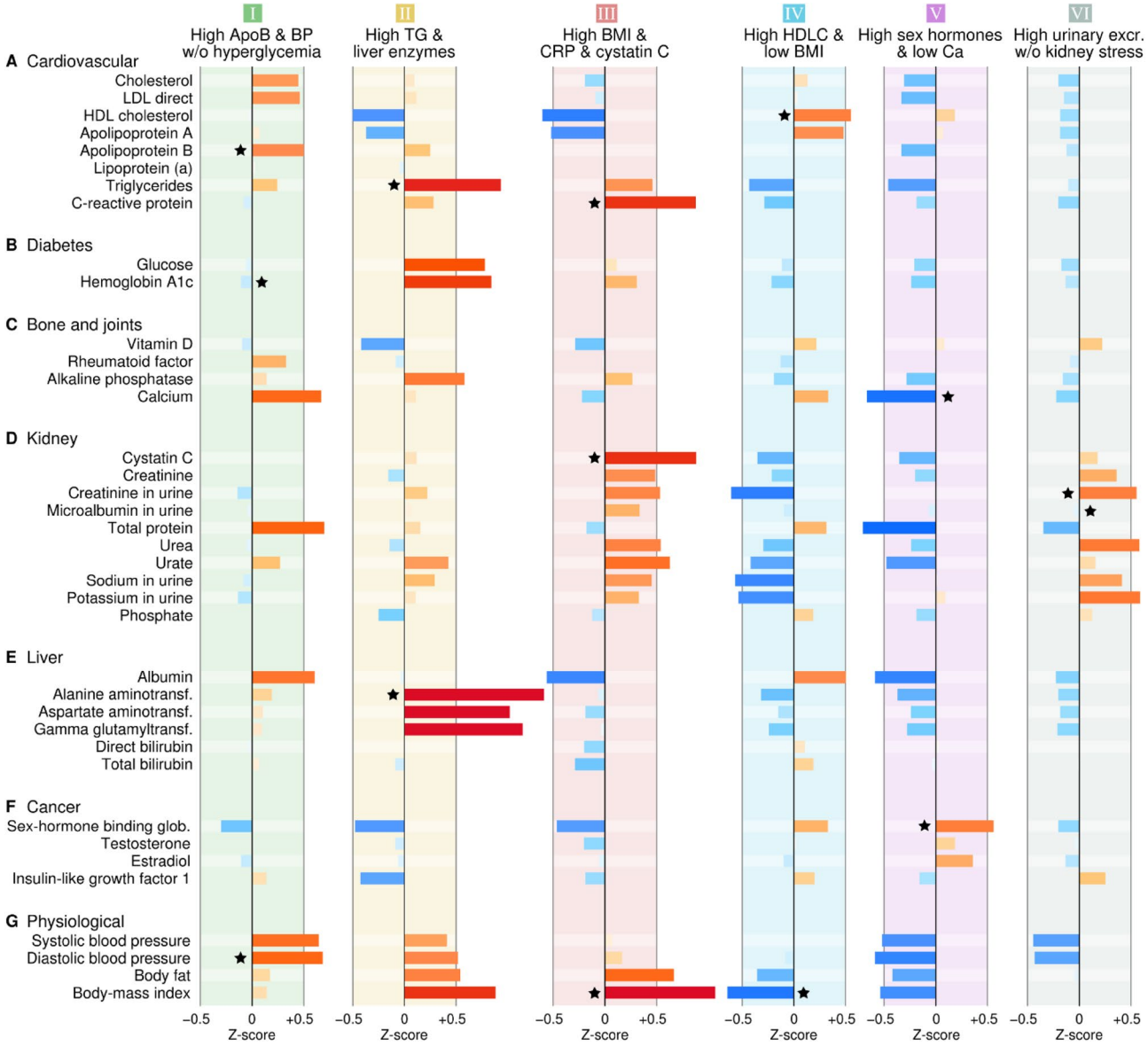

Figure 4. Mean metabolic profiles for SOM subgroups normalized by population SD. The bars are colored according to the direction and magnitude of the deviation from the population mean. The black stars indicate characteristic features that were selected for simplified naming of the subgroups.

Multimorbidity enrichment was pronounced in the metabolically favorable Subgroups IV and V despite them having lower disease burden overall. The paradoxical finding means that the relative risk of co-occurring cardiometabolic disease was higher in the absence of obvious metabolic abnormality. The pattern may reflect genetic and environmental susceptibility that is independent of the typical cardiovascular risk factors but nevertheless pleiotropic to cardiometabolic diseases[38]. The same pattern may also arise from survival bias as people who are simultaneously affected by metabolic dysfunction and multiple morbidities tend to perish younger[39].

The statistical link between metabolic dysfunction and cardiovascular disease is strong on the population level but this does not necessarily translate to accurate prediction of individual events[8], indeed, most cardiovascular risk models show modest predictive ability[9]. For this reason, we envisage the SOM to occupy the inter-mediate space where we can leverage the aggregated statistics over subgroups while interpreting the results as stereotypical individuals that represent meaningful biological phenotypes. Specifically, human observers have visual access to every single variable and its patterns when making the decisions on subgroup boundaries. It is also easy for human observers to verify which subgroup profile matches their own since the profiles are expressible in physical measurement units. Therefore, the SOM model is directly applicable to real-world people and only one SOM is necessary to describe the burden of multiple common disease, as seen in Figs. 5 and 6. Yet a subgroup contains multiple individuals, which enables the calculation of prevalence and incidence rates as subpopulation risk estimates. Indeed, propensity scoring is already used in this manner to identify pools of representative cases within health informatics systems[40]. However, these methods are often presented as black boxes and thus lack the biological context that the SOM colorings can provide. The SOM lets a group of scientists to "see" the data through

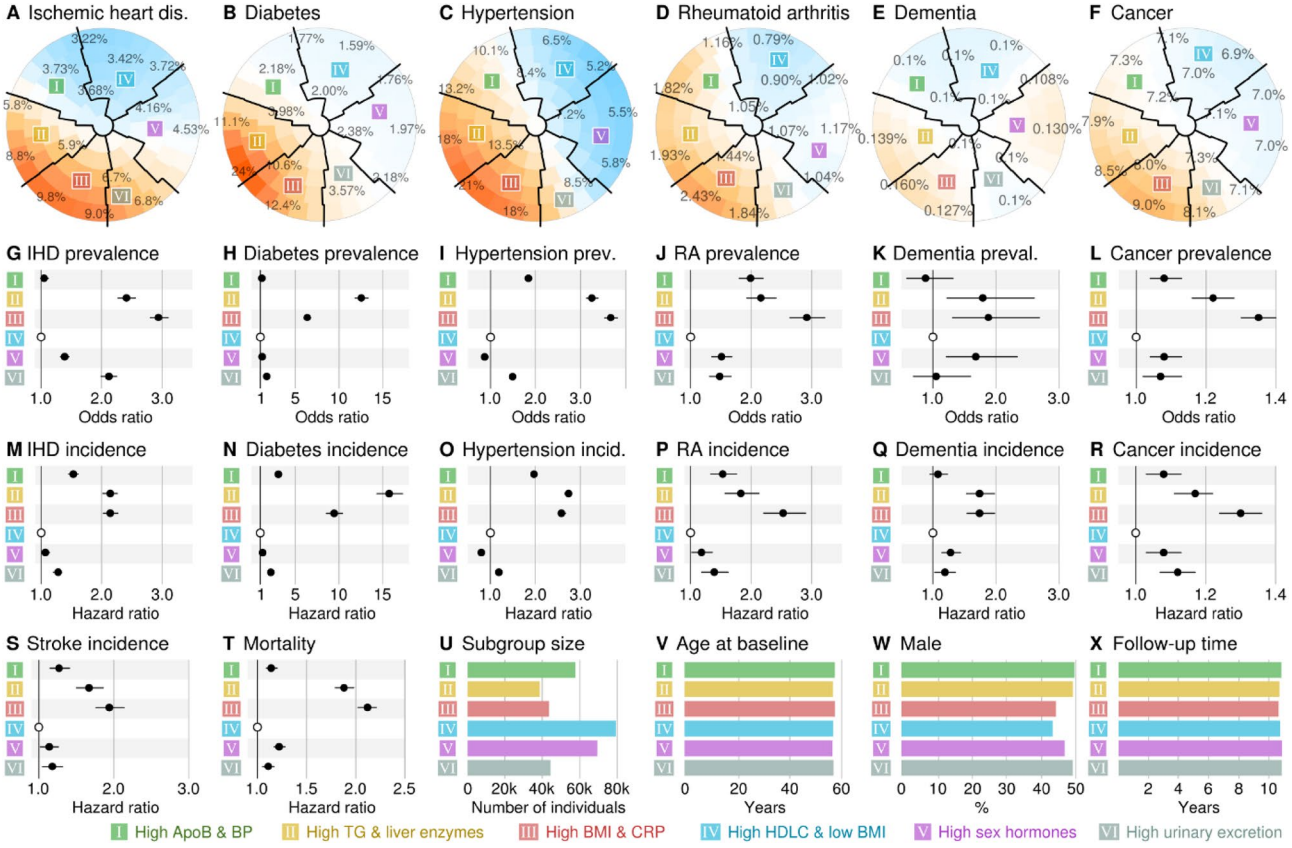

Figure 5. Comparison of morbidity between the SOM subgroups. Percentage of individuals with a disease at baseline across the map districts (**A–F**). Odds ratios for disease prevalence across subgrups based on logistic regression adjusted for age, sex and assessment center (**G–L**). Hazard ratios for incident disease or mortality based on Cox regression adjusted for age, sex and assessment center (**M–T**). Maximum follow-up time available across any clinical end-point (**X**).

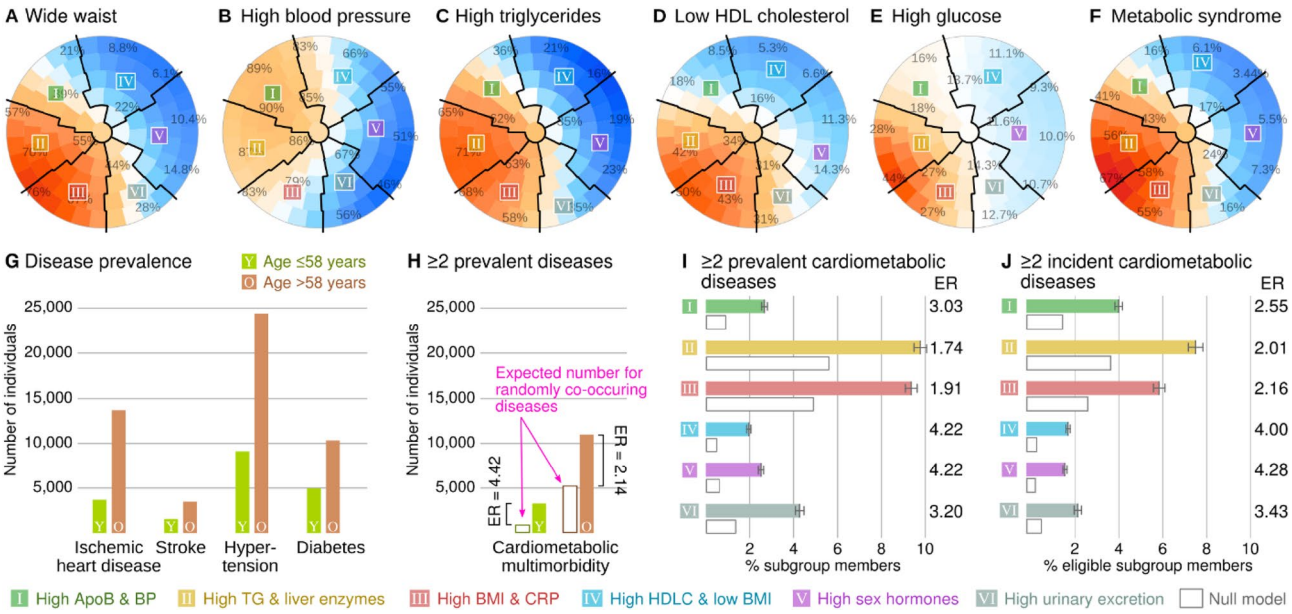

Figure 6. The metabolic syndrome (MetS) and multimorbidity. MetS was defined according to the NCEP ATP III criteria that include five components (**A–E**, the percentages in the plots indicate the proportion of individuals that satisfy a criterion) and subsequent binary classification for those with ≥ 3 points (**F**). The participants were divided into those with age ≤ 58 ($N = 167{,}337$ or 50.7%) and those with age > 58 ($N = 162{,}571$ or 49.3%) to create two equally sized age strata (**G**). The null model represents the number of multimorbid cases if the co-occurrence of diseases was random. Bars for subgroups include 95% confidence intervals (**H–J**).

the statistically standardized colorings in a way no other tool can, and use that information to create a consensus on how to split the population into subgroups that make biological, medical, economic and societal sense.

Limitations. Due to the large sample size, the statistical robustness is high in this study but we urge caution when generalizing the findings of this study to other cohorts, to other ethnicities or to populations of different circumstances. Furthermore, the results are dependent on the selection of available biomarkers and different laboratory panels may produce different subgroups (note also a previously published comprehensive risk factor screening[41]). We also note that the statistical accuracy of population-based data is insufficient to develop a machine learning model for a clinically robust predictive test[8,9]. The UK Biobank recruited volunteers only, thus people with less opportunity to participate due to low socio-economic status or poor health may be underrepresented, however, the disease associations are compatible with other cohorts[42]. Ageing affects metabolism, but the SOM was constructed from cross-sectional data and adjusted for age, thus we are unable to provide information on longitudinal metabolic trajectories and the metabolic subgroups should not be interpreted as part of a temporal sequence.

Conclusions

The SOM subtypes provided a descriptive framework of how combinations of multiple risk factors are associated with diverging cardiometabolic disease outcomes within a population. The new information is useful for the development of targeted interventions for specific subgroups; potential applications include phenotypically guided trials of new treatments where participants are selected based on their full phenotypic profile (e.g. cardiovascular drug trials designed for persons with inflammatory kidney stress vs. persons with diabetogenic liver stress). Such designs will provide more targeted information on the exact type of patient who will benefit the most from the treatment. We also see potential to adopt metabolic profiles as a new approach to assess the health and aggregate disease burden in a population. For example, subtype prevalences can provide phenotypically specific information on how changes in environmental risk factors influence the aggregate disease burden in different segments of the population over time.

Data availability

The UK Biobank data are publicly available (https://www.ukbiobank.ac.uk/). This study was designed and implemented according to project plan #29890.

Received: 13 January 2022; Accepted: 29 April 2022
Published online: 21 May 2022

References

1. Kivimäki, M. *et al.* Overweight, obesity, and risk of cardiometabolic multimorbidity: pooled analysis of individual-level data for 120 813 adults from 16 cohort studies from the USA and Europe. *Lancet Public Health* **2**(6), e277–e285. https://doi.org/10.1016/S2468-2667(17)30074-9 (2017).
2. Bhaskaran, K. *et al.* Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5·24 million UK adults. *Lancet Lond. Engl.* **384**(9945), 755–765. https://doi.org/10.1016/S0140-6736(14)60892-8 (2014).
3. Lee, C. M. *et al.* Association of anthropometry and weight change with risk of dementia and its major subtypes: a meta-analysis consisting 2.8 million adults with 57,294 cases of dementia. *Obes. Rev. Off. J. Int. Assoc. Study Obes.* **21**(4), e12989. https://doi.org/10.1111/obr.12989 (2020).
4. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**(6), 1194–1217. https://doi.org/10.1016/j.cell.2013.05.039 (2013).
5. Ussher, J. R., Elmariah, S., Gerszten, R. E. & Dyck, J. R. B. The emerging role of metabolomics in the diagnosis and prognosis of cardiovascular disease. *J. Am. Coll. Cardiol.* **68**(25), 2850–2870. https://doi.org/10.1016/j.jacc.2016.09.972 (2016).
6. Deelen, J. *et al.* A metabolic profile of all-cause mortality risk identified in an observational study of 44,168 individuals. *Nat. Commun.* **10**(1), 3346. https://doi.org/10.1038/s41467-019-11311-9 (2019).
7. Dennis, J. M., Shields, B. M., Henley, W. E., Jones, A. G. & Hattersley, A. T. Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with models based on simple clinical features: an analysis using clinical trial data. *Lancet Diab. Endocrinol.* **7**(6), 442–451. https://doi.org/10.1016/S2213-8587(19)30087-7 (2019).
8. Sniderman, A. D., Thanassoulis, G., Wilkins, J. T., Furberg, C. D. & Pencina, M. Sick individuals and sick populations by geoffrey rose: cardiovascular prevention updated. *J. Am. Heart Assoc.* **7**(19), e010049. https://doi.org/10.1161/JAHA.118.010049 (2018).
9. Damen, J. A. A. G. *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* **353**, i2416. https://doi.org/10.1136/bmj.i2416 (2016).
10. Ala-Korpela, M. Commentary: data-driven subgrouping in epidemiology and medicine. *Int. J. Epidemiol.* **48**(2), 374–376. https://doi.org/10.1093/ije/dyz040 (2019).
11. Gao, S., Mutter, S., Casey, A. & Mäkinen, V.-P. Numero: a statistical framework to define multivariable subgroups in complex population-based datasets. *Int. J. Epidemiol.* https://doi.org/10.1093/ije/dyy113 (2018).
12. Mäkinen, V.-P. *et al.* Metabolic phenotypes, vascular complications, and premature deaths in a population of 4197 patients with type 1 diabetes. *Diabetes* **57**(9), 2480–2487. https://doi.org/10.2337/db08-0332 (2008).
13. Lithovius, R. *et al.* Data-driven metabolic subtypes predict future adverse events in individuals with type 1 diabetes. *Diabetologia* **60**(7), 1234–1243. https://doi.org/10.1007/s00125-017-4273-8 (2017).
14. Webster, A. J., Gaitskell, K., Turnbull, I., Cairns, B. J. & Clarke, R. Characterisation, identification, clustering, and classification of disease. *Sci. Rep.* **11**(1), 5405. https://doi.org/10.1038/s41598-021-84860-z (2021).
15. Sudlow, C. *et al.* UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779 (2015).
16. Kohonen, T. *Self-organizing maps* (Springer, 2001).
17. Mäkinen, V.-P. *et al.* Metabolic diversity of progressive kidney disease in 325 patients with type 1 diabetes (the FinnDiane Study). *J. Proteome. Res.* **11**(3), 1782–1790. https://doi.org/10.1021/pr201036j (2012).
18. Makinen, V.-P. *et al.* Metabolic phenotypes, vascular complications, and premature deaths in a population of 4197 patients with type 1 diabetes. *Diabetes* **57**(9), 2480–2487. https://doi.org/10.2337/db08-0332 (2008).

19. Mäkinen, V.-P. *et al.* 1H NMR metabonomics approach to the disease continuum of diabetic complications and premature death. *Mol. Syst. Biol.* **4**, 167. https://doi.org/10.1038/msb4100205 (2008).
20. GBD. Diseases and Injuries Collaborators: Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Lond. Engl.* **396**(10258), 1204–1222. https://doi.org/10.1016/S0140-6736(20)30925-9 (2019).
21. Goldstein, J. L. & Brown, M. S. A century of cholesterol and coronaries: from plaques to genes to statins. *Cell* **161**(1), 161–172. https://doi.org/10.1016/j.cell.2015.01.036 (2015).
22. Nichols, H. B. *et al.* From menarche to menopause: trends among US Women born from 1912 to 1969. *Am. J. Epidemiol.* **164**(10), 1003–1011. https://doi.org/10.1093/aje/kwj282 (2006).
23. National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* **106**(25): 3143–3421 (2002)
24. Kassi, E., Pervanidou, P., Kaltsas, G. & Chrousos, G. Metabolic syndrome: definitions and controversies. *BMC Med.* **9**, 48. https://doi.org/10.1186/1741-7015-9-48 (2011).
25. Samuel, V. T. & Shulman, G. I. Nonalcoholic fatty liver disease as a nexus of metabolic and hepatic diseases. *Cell Metab.* **27**(1), 22–41. https://doi.org/10.1016/j.cmet.2017.08.002 (2018).
26. Younossi, Z. M. *et al.* The global epidemiology of NAFLD and NASH in patients with type 2 diabetes: a systematic review and meta-analysis. *J. Hepatol.* **71**(4), 793–801. https://doi.org/10.1016/j.jhep.2019.06.021 (2019).
27. Libby, P. *et al.* Atherosclerosis. *Nat. Rev. Dis. Primer.* **5**(1), 56. https://doi.org/10.1038/s41572-019-0106-z (2019).
28. Sarnak, M. J. *et al.* Chronic kidney disease and coronary artery disease: JACC state-of-the-art review. *J. Am. Coll. Cardiol.* **74**(14), 1823–1838. https://doi.org/10.1016/j.jacc.2019.08.1017 (2019).
29. Muka, T. *et al.* Associations of Steroid sex hormones and sex hormone-binding globulin with the risk of type 2 diabetes in women: a population-based cohort study and meta-analysis. *Diabetes* **66**(3), 577–586. https://doi.org/10.2337/db16-0473 (2017).
30. Wallace, I. R., McKinley, M. C., Bell, P. M. & Hunter, S. J. Sex hormone binding globulin and insulin resistance. *Clin. Endocrinol. (Oxf.)* **78**(3), 321–329. https://doi.org/10.1111/cen.12086 (2013).
31. Bjørnerem, A. *et al.* Endogenous sex hormones in relation to age, sex, lifestyle factors, and chronic diseases in a general population: the Tromsø Study. *J. Clin. Endocrinol. Metab.* **89**(12), 6039–6047. https://doi.org/10.1210/jc.2004-0735 (2004).
32. Honour, J. W. Biochemistry of the menopause. *Ann. Clin. Biochem.* **55**(1), 18–33. https://doi.org/10.1177/0004563217739930 (2018).
33. King, A. J. & Levey, A. S. Dietary protein and renal function. *J. Am. Soc. Nephrol. JASN* **3**(11), 1723–1737 (1993).
34. Armstrong, L. E. & Johnson, E. C. Water intake, water balance, and the elusive daily water requirement. *Nutrients* https://doi.org/10.3390/nu10121928 (2018).
35. Freisling, H. *et al.* Lifestyle factors and risk of multimorbidity of cancer and cardiometabolic diseases: a multinational cohort study. *BMC Med.* **18**(1), 5. https://doi.org/10.1186/s12916-019-1474-7 (2020).
36. Lu, D. & Kassab, G. S. Role of shear stress and stretch in vascular mechanobiology. *J. R. Soc. Interface* **8**(63), 1379–1385. https://doi.org/10.1098/rsif.2011.0177 (2011).
37. de Vos, L. C., Lefrandt, J. D., Dullaart, R. P. F., Zeebregts, C. J. & Smit, A. J. Advanced glycation end products: an emerging biomarker for adverse outcome in patients with peripheral artery disease. *Atherosclerosis* **254**, 291–299. https://doi.org/10.1016/j.atherosclerosis.2016.10.012 (2016).
38. Monte, E. & Vondriska, T. M. Epigenomes: the missing heritability in human cardiovascular disease?. *Proteomics Clin. Appl.* **8**(7–8), 480–487. https://doi.org/10.1002/prca.201400031 (2014).
39. Collaboration, E. R. F. *et al.* Association of cardiometabolic multimorbidity with mortality. *JAMA* **314**(1), 52–60. https://doi.org/10.1001/jama.2015.7008 (2015).
40. Deb, S. *et al.* A review of propensity-score methods and their use in cardiovascular research. *Can. J. Cardiol.* **32**(2), 259–265. https://doi.org/10.1016/j.cjca.2015.05.015 (2016).
41. Madakkatel, I., Zhou, A., McDonnell, M. D. & Hyppönen, E. Combining machine learning and conventional statistical approaches for risk factor discovery in a large cohort study. *Sci. Rep.* **11**(1), 22997. https://doi.org/10.1038/s41598-021-02476-9 (2021).
42. Batty, G. D., Gale, C. R., Kivimäki, M., Deary, I. J. & Bell, S. Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis. *BMJ* **368**, m131. https://doi.org/10.1136/bmj.m131 (2020).

Acknowledgements

We thank the UK Biobank participants and administrators for making this study possible.

Author contributions

A.M. and V.-P.M. analyzed the data, all authors (A.M., E.H., M.A.-K. and V.-P.M.) edited and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-12198-1.

Correspondence and requests for materials should be addressed to V.-P.M.

Reprints and permissions information is available at www.nature.com/reprints.

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