Predicting sensitivity and resilience to modifiable risk factors for cardiometabolic morbidity and mortality

Hugo Pomares-Millan¹, Naemieh Atabaki-Pasdar¹, Ingegerd Johansson², Alaitz Poveda¹, Paul W. Franks^{1,2,3*}

1. Department of Clinical Sciences, Genetic and Molecular Epidemiology Unit, Lund University, Skåne University Hospital Malmö, Malmö, Sweden

2. Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden

3. Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA

Word count: 6917; Abstract word count: 214; Tables: 3; Figures: 2.

Supplementary Figures: 9; Supplementary Tables: 20; References: 34.

Corresponding author:

Paul Franks, Department of Clinical Sciences, Genetic and Molecular Epidemiology Unit, SE-214 28, Malmö, Sweden.

E-mail: paul.franks@med.lu.se

Orcid: orcid.org/0000-0002-0520-7604

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

Abstract

Background: Lifestyle exposures play a major role in the development of disease, yet people vary in their susceptibility. A critical step towards precision medicine is identifying individuals who are resilient or sensitive to the environment, and, assess whether the allocation to these predicted groups are more or less likely to develop cardiometabolic disease.

Methods: We have used repeated data from the VHU study (n=35440) to identify sensitive and resilient individuals using prediction intervals at the 5th and 95th quantile. Three exposure susceptibility groups were derived per cardiometabolic score using quantile regression forests in the training dataset; next, in the validation dataset, we assessed the different risks of the groups using Cox proportional hazard models for CVD and diabetes.

Results: The results of our study suggest that, after ~10 y of follow-up, individuals with sensitivity to the environmental exposures associated with systolic and diastolic blood pressure, blood lipids, and glucose were at higher risk of developing cardiometabolic disease. Moreover, when hazards were pooled with the replication cohort, for those individuals sensitive to the exposures associated with blood pressure traits, the hazards remained significant.

Conclusions: Identifying individuals who are predicted to be sensitive are at higher risk of developing disease, this population may be a clinical target for prevention or early intervention and public health strategies.

Introduction

There is growing recognition that people vary in their susceptibility to environmental risk factors for cardiometabolic diseases, suggesting that one-size-fits-all public health recommendations are unlikely to yield optimal results. Early identification of individuals who are most likely to develop diseases like type 2 diabetes (T2D) and cardiovascular disease (CVD) is desirable, as efficacious therapeutics (both lifestyle and pharmacologic) exist that can help prevent these diseases [1]. Moreover, once manifest T2D and CVD often cause life-threatening health complications can prove difficult and costly to treat [2].

Most statistical models examining susceptibility to lifestyle risk factors, from which public health recommendations are drawn, assume that a given lifestyle exposure conveys a similar effect on disease risk throughout the target population, with variability in these effects viewed solely as a consequence of measurement error and other stochastic factors [3]. However, some of this variability is likely to reflect real differences in exposure-outcome effects, such that for a lifestyle exposure like smoking, physical inactivity or sugar consumption, a gradient of risk throughout the population will exist, with some people being highly sensitive to the adverse effects of these exposures, whilst others will be resilient.

Identifying the sub-populations of those who are environmentally “sensitive” and “resilient” may have utility for the targeted prevention of cardiometabolic disease, with those considered sensitive being the target of early preventive interventions [1, 4], while alternative public health approaches may be warranted in those who appear resilient. Here we used a machine learning approach [5] to differentiate error from true variability in susceptibility to lifestyle risk factors for T2D and CVD. In so doing we sought to identify population subgroups of sensitive and resilient individuals and determine whether allocation to these subgroups affects risk of incident disease and premature mortality. The study was performed in a large, repeated

measures, prospective cohort study (n >35,000) in which detailed assessments of lifestyle exposures, multiple cardiometabolic risk markers, and clinical events were obtained.

Materials and Methods

Study design and participants

The Västerbotten Health Survey (*Västerbottens hälsoundersökning*; VHU) [6, 7] is a prospective, population-based cohort study designed to monitor and improve health of the general population in Västerbotten county, northern Sweden. Adults residing in Västerbotten have been invited to attend their primary care centre to undertake a baseline clinical examination and complete detailed lifestyle questionnaires during the years of their 40th, 50th, and 60th birthdays. We used data derived from VHU (n= 42,887) in our analyses. Owing the small subgroup of participants who were not born in Sweden (n= 7,039), the current analysis focused only on the Swedish-born contingent of VHU. Furthermore, participants in whom diabetes or cardiovascular disease had been diagnosed at baseline (n= 408) were also removed to minimize biases that can occur when people with disease diagnoses are asked to self-report their lifestyle behaviours. Participants with two health examinations between 1985 through 2016 (with ~10 year apart between each visit) were included in the final dataset, which comprised 35,440 participants (see flowchart in Figure 1).

Cardiometabolic risk markers

Clinical measurement protocols in VHU are reported elsewhere [6]. Briefly, height and weight were measured with calibrated tools (stadiometer and scales respectively), with participants wearing light clothing and no shoes. Body mass index (BMI) was calculated as body weight in kilograms divided by height in meters squared. Systolic and diastolic blood pressures (SBP and DBP, respectively) were measured with participants resting supine, using either manual or automated sphygmomanometers. Peripheral blood was drawn after overnight or 4-hour fasting, and a venous blood sample was drawn two hours after the administration of a 75-gram oral

glucose load. Blood glucose [i.e., fasting (FG) and two-hour glucose (2hG)], total cholesterol (TC) and triglycerides (TG) levels were then measured using a Reflotron bench-top analyser (Roche Diagnostics Scandinavia AB); High-density lipoprotein cholesterol (HDL-C) was also measured in a subgroup of participants (n= 5,348) and low-density lipoprotein cholesterol (LDL-C) obtained using the Friedewald formula (n= 5,246) [8]. In September 2009, blood lipids and blood pressure measurements changed, thereafter, blood pressure was measured twice in a sitting position and averaged; TG and TC levels were analysed using standardized chemical analysis in the hospital clinical biochemistry laboratory. Validated conversion equations were used to adjust the blood pressure and lipids measurements taken before and after September 2009 [9]: for participants on lipid lowering and/or blood pressure lowering medications, lipid levels and/or blood pressure levels were corrected by adding published constants (+0.208 mmol/l for TG, +1.347 mmol/l for TC, -0.060 mmol/l for HDL-C, +1.290 mmol/l for LDL-C, +15 mmHg for SBP and +10 mmHg for DBP) suggested in the literature [10, 11]. Cardiometabolic traits' values outside the thresholds suggested by VHU data managers were considered outliers and removed (criteria in Supplementary table 1).

Lifestyle and dietary assessments

All participants were requested to complete a self-administered comprehensively validated lifestyle questionnaire during each visit; the questionnaire queries socio-economic factors, physical/mental health, quality of life, social network and support, working conditions, and alcohol/tobacco use. Physical activity was interrogated using the modified version of the International Physical Activity Questionnaire [12, 13]. A validated semi-quantitative food frequency questionnaire (FFQ) designed to capture habitual diet over the last year was used to obtain information on various dietary factors [14]. In 1996, the FFQ was reduced from 84 to 66 items by merging similar items and removing those redundant. Nutrient and energy contents were calculated based on the Swedish Food Composition Database [15] based on meal

frequency and portion size. Food intake level (FIL) was calculated as total energy intake (TEI) divided by estimated basal metabolic rate, individuals with extreme TEI (below the 5th and above the 97.5th percentile of food intake level in VHU) were excluded from the analyses [16]. Observations with lifestyle values considered biologically implausible were removed (see criteria in Supplementary table 2). Written informed consent was obtained from all living participants at enrolment into VHU. The study was approved by the Region Ethical Review Board in Umeå.

Outcome assessment

Data pertaining to medical diagnoses and mortality were retrieved through record linkage from the National Board of Health and Welfare in Sweden (www.socialstyrelsen.se/register) until December 31st, 2016. Using the participant's civic registration numbers, their records were linked, and the following diagnoses codes were retrieved: ICD-9 code 250 and ICD-10 codes E11.0–E11.9 for T2D; for the composite CVD outcome: ICD-9 code 410 and ICD-10 code I21 were used for myocardial infarction (MI), and ICD-9 codes 430, 431, and 433–436 and ICD-10 codes I60, I61, I63 and I64 for stroke. The first date of a registered event was selected as the outcome for the current analyses.

Replication in UK Biobank

UK Biobank (UKB) is a prospective cohort study since 2007 with more than 500,000 adults aged 40–70 years recruited between 2006 and 2010 in 22 centres across England, Scotland, and Wales [17]. We used the baseline assessment, which involved physical measures, sample collection, and touchscreen questionnaires that included demographics, lifestyle, environmental exposures, and medical history, as described in detail elsewhere [17]. We retrieved the same 9 traits as in VHU, with one exception: HbA1c was available in UKB, but not 2hG (the latter of which was available in VHU). The sample was restricted to individuals who reported *No* to the question: "*Have you made any major changes to your diet in the last 5*

years?" (n= 289,620). We selected those with outcome data in at least one follow-up (i.e., second or third instance) and without secondary withdrawal of consent (n= 43,482), we excluded cases of prevalent T2D, gestational diabetes or CVD, and missing data at baseline; thus, 18,037 participants were included in the analysis.

The data processing was the same as described for VHU. Dietary information was collected via a self-reported dietary questionnaire (Oxford WebQ) [18]. Outcome data, i.e., CVD incidence (*"Has a doctor ever told you that you have had any of the following conditions: angina, heart attack, stroke, high blood pressure?"* at follow-up) and diabetes incidence (*"Has a doctor ever told you that you have diabetes?"* at follow-up) were self-reported. The participants were followed-up for cardiovascular mortality via linkage to information held in national death registers. Date of death was obtained from death certificates held by the National Health Service (NHS) Information Centre (England and Wales) and the NHS Central Register Scotland (Scotland). At the time of analysis, complete data from cause of death registers was available through January 31st, 2018; thus, they were censored at these dates or date of death if this occurred first. The field numbers for the UK Biobank variables can be viewed in the cohort's website: <http://www.ukbiobank.ac.uk/>; and their correspondence with VHU variables are detailed in the supplementary material (Supplementary table 3 and 4). This research has been conducted using the data obtained via UK Biobank Access Application number 18274.

Statistical analysis

We applied a statistical approach termed "environment-wide association study" (EWAS) for the identification of environmental risk markers, as described previously in detail [19]. Briefly, all numeric predictors were inverse normalized to correct skewness and the derived ordinal variables were treated as continuous variables in subsequent analyses. The screened predictors (~300) were those statistically significant at the corrected *p*-value threshold after multiple

testing from the EWAS. We retrieved 167 predictors for BMI, 49 and 37 for SBP and DBP, respectively, 87 for TC, 108 for TG, 50 for HDL-C and 21 with LDL-C. Forty-three predictors were associated with FG and 58 with 2hG [20]. Categorical predictors with more than two levels were converted into dummy variables, using the lowest value as reference. In VHU, reported nutrients were adjusted for TEI with the residual method [21] to obtain an unconfounded effect. We removed correlated (>80%) and zero variance predictors considered to avoid redundancy [22], variables removed are described in Supplementary table 5. For both datasets, we assumed missingness at random (MAR) [23], thus, environmental predictors with <50% missingness were imputed with *missForest* package from R software using a non-parametric approach for mixed data type, to allow a complete cases analysis suitable for the random forest algorithm; continuous predictors were verified by the mean squared error (MSE) and categorical predictors by the proportion falsely classified (PFC) (Supplementary tables 12-20)[24].

We randomly partitioned the VHU dataset into training (50%) and testing (50%) sets. The training set was used to fit quantile regression forest (QRF) models for predictors associated with the cardiometabolic traits and the testing set to predict future intervals. Multicollinearity of the variables within these models was assessed using the variance inflation factor, those >10 were removed [25]. All models were adjusted for age, age², sex, FFQ version, BMI (when not as response variable), follow-up time, and fasting status (for glycaemic and lipid models). UKB models were adjusted for age, sex, BMI (when not as response variable), fasting status, and ethnicity. We utilized QRFs [5], an extension of the supervised machine learning technique random forests, which is an ensemble of simultaneous decision trees derived from bootstrapped samples [26]. Furthermore, we set prediction intervals (PIs) at 90% probability (5th and 95th quantiles) to minimize false positives $[(1 - \alpha) \times 100\%]$, while ensuring a sufficient number of events per category for the time-to-event analysis. The PIs were constructed from the

conditional quantiles of the trait response predicted by QRFs. Briefly, the prediction intervals of a trait response Y given the environmental predictors X was built by $I(x) = [q\alpha/2(Y|X = x), q1 - \alpha/2(Y|X = x)]$. Thus, the 90% prediction interval for the trait value was estimated by

$$I(x) = [q0.05(Y|X = x), q0.95(Y|X = x)]$$

Being that for a given x , the trait response lies within the interval $I(x)$ with high probability. For VHU, based upon the obtained PIs per trait we defined three groups of persistence when baseline visit and follow-up values fall consecutively into one of the categories: persistently ‘resilient’ (below 90% PIs), ‘as expected’ (within 90% PIs) and ‘sensitive’ (above 90% PIs). However, in UKB was not possible to consider two consecutive measures so QRFs were obtained only for the baseline visit (supplementary figures 1-9).

Time-to-event analysis

Cox proportional hazards regression models were used to estimate hazard ratios (HRs) and corresponding 95% confidence intervals (CIs) between persistence category for each cardiometabolic trait derived from the QRF approach and the risk of diabetes and CVD-incidence and mortality. The category ‘as expected’ was used as the reference group. Statistical significance (p -value) was set at 5% level. Three models were constructed per trait, a basic model included age, sex, and BMI; a partially adjusted model included the basic model plus fasting status, FFQ version, TEI and educational level (education has been used as a proxy of socioeconomic status previously in this population) (22); whereas in UKB models we included ethnicity and fasting; and a third model (fully adjusted) consisted of the partially adjusted model adding smoking status, physical activity (except in UKB), and alcohol consumption. The covariates were selected *a priori* owing to their previously established associations with cardiovascular mortality in this population [27], and if a covariate was already in the environmental QRF model, it was not included. The timescale in VHU was the time-on-study

and for the UKB we estimated attained age based on date of birth as the origin point (we defined the 15th day of each month for all participants owing only data on month and year of births were available), date of first attendance as the entry point, and date of exit as date of death, lost to follow-up or censorship (January 31st, 2018). The proportional hazards assumption was tested with Schoenfeld residuals. Incidence rate (IR) ratio (cases per 100,000 person-years) and IR difference (incidence rate of ‘resilient’ or ‘sensitive’ minus incidence rate of ‘as expected’) with its 95% CIs were estimated when too few events had occurred. HRs and 95% CIs from VHU and UKB were pooled for each trait by persistence category to obtain an overall estimate under a random-effects model [28]; heterogeneity was tested with Cochran’s Q statistic [29, 30]. As sensitivity analysis, individuals in the ‘no persistence’ category were assessed in Cox proportional hazards (Supplementary table 6) and two non-modifiable covariates were added to the fully-adjusted model: *‘Parents or siblings have diabetes’* and *‘Parents or siblings had a cerebral haemorrhage/thrombosis or cardiac infarction before the age of 60’*. All statistical analyses were performed using R software version 3.6.1 [31] and packages used are listed in Supplementary table 10.

Results

Baseline characteristics for each trait of VHU and UKB cohorts are shown in Table 1. Median follow-up [interquartile range (IQR)] for VHU was 9.7 (5.8). Fully-adjusted Cox model results per cardiometabolic trait for CVD and T2D are in table 2 and 3, respectively; basic and partially-adjusted Cox models are shown in Supplementary tables 7-9. In VHU, DBP ‘sensitive’ individuals had IRR of 2.4 (95%CI 1.5, 3.7) and the HRs were 2.25 (95%CI 1.49, 3.4; $p = 1.05E-04$) for CVD, conversely, the ‘resilient’ group had lower IRR of 0.4 (95%CI 0.2, 0.9) and hazards 0.39 (95%CI 0.17, 0.87; $p = 2.1E-02$) when compared with the reference group. For LDL-C, the IRR was 1.5 (95%CI 1.1, 3.8) in the ‘sensitive’ group and exhibited

higher HRs for CVD when compared with the reference, 2.17 (95%CI 1.17, 4 ; $p=0.01$); for TC, those ‘sensitive’ had IRR 2.5 (95%CI 1.8, 3.5) and HR 2.2 (95%CI 1.6, 3.1; $p=2.00E-06$) for CVD and ~2 times the rate (95%CI 1, 3.5) of T2D, which was also observed in UKB. In UKB cohort, the FG ‘sensitive’ group had higher rates IRR 3.5 (95%CI 2, 5.6) and risk of developing T2D than the reference, HR: 4.4 (95%CI 2.58, 7.51; $p=5.8E-08$); the risk of T2D in the ‘resilient’ group was lower, yet not statistically significant. For the SBP ‘sensitive’ group, the HRs for CVD were 2.4 (95%CI 1.12, 5.14; $p=0.02$) compared with the reference group, moreover, when HRs from both cohorts were pooled the overall SBP ‘sensitive’ HR for CVD was 1.9 (95%CI 1.1; 3.0), Cochran’s Q test= 8.42, $P_Q=0.0037$ (Fig. 2A); Similarly, in DBP ‘sensitive’ individuals had IRR 2.5 (95%CI 1.2, 5) and HR for CVD was 2.8 (95%CI 1.38, 5.79; $p=0.437E-03$) and, when aggregated with VHU, the pooled DBP ‘sensitive’ HR for CVD was 2.3 (95%CI 1.5; 3.4), Cochran’s Q test= 13.3, $P_Q=0.0003$ (Fig. 2B). In UKB, the HDL-C ‘resilient’ group were at higher hazards for T2D, 2.4 (95%CI 1.17, 4.99; $p=0.02$) and those ‘sensitive’ had IRR 0.03 (95%CI 0.001, 0.23) and IRD -0.44 (95%CI -0.5, -0.4) when compared with the reference. For BMI, no fully-adjusted Cox model was significant yet the IRR of T2D in those ‘resilient’ was 0.2 (95%CI 0.06, 0.96) and, on the contrary, 4 times more (95%CI 2.2, 7.2) the rate for the ‘sensitive’ population. In UKB, for TG the hazards for CVD in the ‘sensitive’ group were twice (HR: 2; 95%CI 1.0, 4.2; $p=0.04$), this association was also observed for T2D outcome, but the statistical significance disappeared when the model was maximally adjusted. For CVD-mortality outcome there were no sufficient number of events to assess and compare among groups, thus, IRR and IRD were reported (see Supplementary table 11).

Discussion

Public health guidelines for disease risk stratification rely on population-averaged estimates of risk factor susceptibility, often involving categorization of intermediate clinical markers such as blood pressure or serum cholesterol levels. Public health guidelines for cardiometabolic disease prevention focus on reducing levels of harmful exposures such as smoking and consumption of processed foods, as well as promoting health-enhancing behaviours such as exercise and fresh fruit and vegetable consumption. However, public health guidelines do not provide recommendations tailored to the individual's level of risk factor susceptibility. To explore whether doing so might be of clinical value, we used machine learning methods to identify, in an agnostic manner, population subgroups that are especially sensitive or resilient to modifiable risk exposures for cardiometabolic disease. We showed that those who are especially sensitive to these risk exposures tend to develop T2D and CVD more rapidly. For example, people who are sensitive to the modifiable risk factors associated with elevated blood pressure, CVD risk was twice that seen in people who responded as anticipated (i.e., had the population average response) to these risk factors; similarly, people who were sensitive to the risk factors for elevated triglycerides, LDL, and total cholesterol concentrations were twice as likely to develop CVD or diabetes. This type of risk classification is important, as it highlights individuals with "normal" or "low" levels of intermediate clinical markers who are at relatively high risk of clinical events and who would be overlooked by conventional screening and risk-classification approaches.

Of the nine intermediate clinical markers studied here, not all provided information that led to clinically relevant stratification of risk. For example, risk factor susceptibility for obesity was not associated with CVD or diabetes in either of the UK or Swedish cohorts. Similarly, sensitivity to the risk factors for elevated 2hr glucose (only available in VHU) were not associated with risk of clinical events. The scarcity of events in subgroups for traits like, LDL and HDL cholesterol in VHU and FG, HbA1c, BMI and TC in UKB, gave relatively wide

confidence intervals (Tables 2 and 3). It is likely that some of the analyses performed here are underpowered and the negative findings may, by consequence, be false positive. It is also likely that some of the risk factor exposure patterns are specific to the UK and Swedish populations studied. Furthermore, although the Swedish study is characterised by its broad array of exposure assessments, the UK cohort is less well characterised, which is a key limitation for the analysis performed here. Moreover, because total energy intake was not assessed in the UK cohort, dietary exposures cannot be fully standardized, which may lead to confounding related to body composition and other features of energy balance such as exercise. One should also consider that the exposure assessments in both cohorts were obtained through questionnaires, which are prone to reporting biases, and the observational nature of the study makes causal inference challenging. Thus, while analyses of this nature are likely to yield valid estimates of clinical risk, intervening on the modifiable exposures used to define susceptibility may not necessarily reduce risk. As with most findings generated from epidemiological studies, intervention trials are required to address this limitation.

Most current clinical guidelines for T2D and CVD discuss the importance of personalized care, but include generic healthy lifestyle recommendations [32, 33]. These recommendations do not accommodate an individual's variability in susceptibility to environmental risk factors. There has been extensive debate about the role of precision medicine in disease prevention, which typically focuses on population subgroups with distinct risk factor and treatment response profiles, such that efficacy is maximized and costs and risk are minimized [1]. The approach described here is aligned with the objectives of precision prevention [1], by aiding the identification of people at high risk of cardiometabolic disease and helping determine which modifiable exposures to intervene on.

In conclusion, the approach to cardiometabolic risk stratification presented here may help improve the precision with which at-risk subgroups of the population are identified. In practice,

the implementation of this approach would require combined assessments of modifiable risk exposures and intermediate markers of cardiometabolic risk. Calculating an individual's level of risk using the current approach is more complicated than convention risk algorithms, because it leverages conditional probabilities; however, this could be managed through app-based assessment and decision support systems, which have proven successful elsewhere [34].

Data availability:

The individual-level data from NSHDS is not publicly available due to privacy and confidentiality constraints of Swedish regulation, but data is available from the Department of Biobank Research, Umeå University, upon reasonable request.

Acknowledgements:

We thank the participants, health professionals, and data managers involved in the Västerbotten Health Survey and UK Biobank. This research was conducted using the UK Biobank resource (application ID: 18274).

Funding:

The work described in this paper was supported by the European Research Council (CoG-2015_681742_NASCENT), Swedish Research Council, Strategic Research Area Exodiab, (Dnr 2009-1039), the Swedish Foundation for Strategic Research (IRC15-0067), and the Swedish Research Council, Linnaeus Grant (Dnr 349-2006-237).

Conflicts of interest:

PWF has received research grants from numerous diabetes drug companies and fees as consultant from Novo Nordisk, Lilly, and Zoe Global Ltd. He is currently the Scientific Director in Patient Care at the Novo Nordisk Foundation. Other authors declare non competing interests.

Reference:

- [1] Chung WK, Erion K, Florez JC, et al. (2020) Precision medicine in diabetes: a Consensus Report from the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetologia*: 1-23
- [2] Zhou X, Siegel KR, Ng BP, et al. (2020) Cost-effectiveness of diabetes prevention interventions targeting high-risk individuals and whole populations: a systematic review. *Diabetes Care* 43(7): 1593-1616
- [3] Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media
- [4] Franks PW, Poveda A (2017) Lifestyle and precision diabetes medicine: will genomics help optimise the prediction, prevention and treatment of type 2 diabetes through lifestyle therapy? *Diabetologia* 60(5): 784-792. 10.1007/s00125-017-4207-5
- [5] Meinshausen N (2006) Quantile regression forests. *Journal of Machine Learning Research* 7(Jun): 983-999
- [6] Hallmans G, Agren A, Johansson G, et al. (2003) Cardiovascular disease and diabetes in the Northern Sweden Health and Disease Study Cohort - evaluation of risk factors and their interactions. *Scandinavian journal of public health Supplement* 61: 18-24. 10.1080/14034950310001432
- [7] Norberg M, Wall S, Boman K, Weinshall L (2010) The Vasterbotten Intervention Programme: background, design and implications. *Glob Health Action* 3. 10.3402/gha.v3i0.4643
- [8] Friedewald WT, Levy RI, Fredrickson DS (1972) Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem* 18(6): 499-502
- [9] Ng N, Carlberg B, Weinshall L, Norberg M (2012) Trends of blood pressure levels and management in Vasterbotten County, Sweden, during 1990-2010. *Global Health Action* 5: 1-12. 10.3402/gha.v5i0.18195
- [10] Wu J, Province MA, Coon H, et al. (2007) An investigation of the effects of lipid-lowering medications: genome-wide linkage analysis of lipids in the HyperGEN study. *BMC Genet* 8: 60. 10.1186/1471-2156-8-60
- [11] Tobin MD, Sheehan NA, Scurrah KJ, Burton PR (2005) Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Stat Med* 24(19): 2911-2935. 10.1002/sim.2165
- [12] Hallal PC, Victora CG (2004) Reliability and validity of the International Physical Activity Questionnaire (IPAQ). *Med Sci Sports Exerc* 36(3): 556. 10.1249/01.mss.0000117161.66394.07
- [13] Craig CL, Marshall AL, Sjostrom M, et al. (2003) International physical activity questionnaire: 12-country reliability and validity. *Med Sci Sports Exerc* 35(8): 1381-1395. 10.1249/01.MSS.0000078924.61453.FB
- [14] Johansson I, Hallmans G, Wikman A, Biessy C, Riboli E, Kaaks R (2002) Validation and calibration of food-frequency questionnaire measurements in the Northern Sweden Health and Disease cohort. *Public Health Nutr* 5(3): 487-496. 10.1079/PHNPHN2001315
- [15] Livsmedelsverket The food database. Available from <https://www.livsmedelsverket.se/en/food-and-content/naringsamnen/livsmedelsdatabasen>. Accessed 29th June 2018
- [16] Winkvist A, Hörnell A, Hallmans G, Lindahl B, Weinshall L, Johansson I (2009) More distinct food intake patterns among women than men in northern Sweden: a population-based survey. *Nutrition journal* 8(1): 12

- [17] Sudlow C, Gallacher J, Allen N, et al. (2015) 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
- [18] Bradbury KE, Young HJ, Guo W, Key TJ (2018) Dietary assessment in UK Biobank: an evaluation of the performance of the touchscreen dietary questionnaire. *Journal of nutritional science* 7
- [19] Patel CJ, Bhattacharya J, Butte AJ (2010) An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PloS one* 5(5)
- [20] Poveda A (2021) Environment-wide association study (EWAS) on cardiometabolic traits: A systematic assessment of the association of lifestyle variables on a longitudinal setting. In. *medRxiv*
- [21] Willett WC, Howe GR, Kushi LH (1997) Adjustment for total energy intake in epidemiologic studies. *The American journal of clinical nutrition* 65(4): 1220S-1228S
- [22] Kuhn M, Johnson K (2013) *Applied predictive modeling*. Springer
- [23] Zhang Z (2015) Missing data exploration: highlighting graphical presentation of missing pattern. *Annals of translational medicine* 3(22)
- [24] Stekhoven DJ, Bühlmann P (2012) MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28(1): 112-118
- [25] Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. *Methods in ecology and evolution* 1(1): 3-14
- [26] Breiman L (2001) Random forests. *Machine learning* 45(1): 5-32
- [27] Padyab M, Blomstedt Y, Norberg M (2014) No association found between cardiovascular mortality, and job demands and decision latitude: Experience from the Västerbotten Intervention Programme in Sweden. *Social science & medicine* 117: 58-66
- [28] DerSimonian R, Kacker R (2007) Random-effects model for meta-analysis of clinical trials: an update. *Contemporary clinical trials* 28(2): 105-114
- [29] Borenstein M, Hedges LV, Higgins JP, Rothstein HR (2011) *Introduction to meta-analysis*. John Wiley & Sons
- [30] Borenstein M, Higgins JP (2013) Meta-analysis and subgroups. *Prevention science* 14(2): 134-143
- [31] Team RC (2017) *R: A Language and Environment for Statistical Computing*. 887 R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/> 888
- [32] Arnett DK, Blumenthal RS, Albert MA, et al. (2019) 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Journal of the American College of Cardiology* 74(10): 1376-1414
- [33] American Diabetes Association (2021) 15. Diabetes Care in the Hospital: Standards of Medical Care in Diabetes—2021. *Diabetes Care* 44(Supplement 1): S211-S220
- [34] Berry SE, Valdes AM, Drew DA, et al. (2020) Human postprandial responses to food and potential for precision nutrition. *Nature medicine* 26(6): 964-973

Figure 1. Overview of VHU and UKB studies, data processing and model training.

Figure 2. Forest plots of VHU and UKB studies by persistence category and CVD event.

Panel A: Systolic blood pressure [SBP (mm/Hg)]; Panel B: Diastolic blood pressure [DBP

(mm/Hg)]. Random-effects meta-analysis of the association between trait-persistence category

and CVD. The square and diamond shapes represent summary estimates while the horizontal

bars represent the 95% confidence intervals. HR: Hazard ratio.ES: effect estimate. SE: standard

error.

448

Table 1. Baseline characteristics of VHU and UKB cohorts

	VHU	UKB
n	35440	18037
Male (%)	15599 (46.8)	7881 (43.7)
Age (y)	42.96 (7.02)	55.32 (7.58)
BMI (kg/m²)	25.10 (3.71)	26.02 (3.98)
TC (mmol/L)	5.47 (1.14)	5.88 (1.00)
HDL-C (mmol/L)	1.32 (0.57)	1.53 (0.37)
LDL-C (mmol/L)	3.92 (1.16)	3.69 (0.76)
TG (mmol/L)	1.32 (0.76)	1.60 (0.86)
FG (mmol/L)	5.31 (0.63)	4.91 (0.58)
2hG (mmol/L)	6.39 (1.30)	-
HbA1c (mmol/mol)	-	34.77 (3.67)
SBP (mmHg)	123.27 (15.77)	133.07 (15.00)
DBP (mmHg)	77.25 (10.86)	78.65 (8.20)

All values are mean (SD) otherwise stated. VHU: Västerbotten Health Survey; UKB: UK biobank; BMI: Body Mass index; HDL-C: High-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; HbA1c: Glycated haemoglobin; 2hG: 2h glucose tolerance; FG: Fasting glucose; TG: Triglycerides; TC: Total cholesterol.

449

450

451

452 Table 2. Fully-adjusted Hazard ratios and 95%CI of prediction interval categories and CVD

Trait/study	Categories	n	Events	person-years	IRR*	95% (CIs)		IRD*	95% (CIs)		HR	95% (CIs)		p
<i>Fasting glucose</i> VHU	Ref (Expected)	13596	446	10854.5							1.00			
	Resilient	261	11	193.1	1.39	0.76	2.52	0.02	-0.02	0.05	1.21	0.82	1.80	0.34
	Sensitive	576	26	486.5	1.30	0.88	1.93	0.01	-0.01	0.03	1.54	0.85	2.81	0.16
	Ref (Expected)	11578	148	264.4										
	Resilient	493	4	11.1	0.64	0.24	1.74	-0.20	-0.56	0.17	0.71	0.26	1.92	0.50
	Sensitive	567	9	13.4	1.20	0.61	2.35	0.11	-0.34	0.56	1.20	0.61	2.36	0.59
<i>2-h Glucose/ HbA1c</i>														
VHU	Ref (Expected)	12233	392	10063.7							1.00			
	Resilient	535	16	457.9	0.90	0.54	1.48	-0.004	-0.02	0.01	0.77	0.47	1.27	0.31
	Sensitive	192	6	160.3	0.96	0.43	2.15	-0.002	-0.03	0.03	1.33	0.59	2.98	0.49
	Ref (Expected)	11541	136	263.4							1.00			
	Resilient	569	7	13.1	1.03	0.48	2.21	0.02	-0.39	0.42	1.17	0.55	2.51	0.68
	Sensitive	519	4	11.8	0.66	0.24	1.79	-0.17	-0.52	0.17	0.63	0.23	1.70	0.36
<i>DBP</i>														
VHU	Ref (Expected)	13127	429	10873.5							1.00			
	Resilient	456	6	374.1	0.41	0.18	0.91	-0.02	-0.04	-0.01	0.39	0.17	0.87	0.02
	Sensitive	252	19	202.6	2.38	1.50	3.76	0.05	0.01	0.10	2.25	1.49	3.40	1.05E-04
	Ref (Expected)	11976	135	273.5							1.00			
	Resilient	382	4	8.7	0.93	0.34	2.52	-0.03	-0.49	0.42	0.87	0.32	2.35	0.78
	Sensitive	282	8	6.5	2.49	1.22	5.09	0.74	-0.12	1.59	2.83	1.38	5.79	4.37E-03
<i>HDL</i>														
VHU	Ref (Expected)	1855	90	1416.9							1.00			
	Resilient	95	3	70.5	0.67	0.21	2.12	-0.02	-0.07	0.03	0.64	0.20	2.03	0.45

Trait/study	Categories	n	Events	person-years	IRR*	95% (CIs)		IRD*	95% (CIs)		HR	95% (CIs)		p
UKB	Sensitive	91	5	66.9	1.18	0.48	2.89	0.01	-0.06	0.08	1.22	0.49	3.01	0.67
	Ref (Expected)	12015	153	274.5							1.00			
	Resilient	288	2	6.7	0.54	0.13	2.16	-0.26	-0.68	0.16	0.61	0.15	2.45	0.48
	Sensitive	319	3	7.2	0.75	0.24	2.34	-0.14	-0.62	0.34	0.73	0.23	2.29	0.59
BMI														
VHU	Ref (Expected)	12951	423	10640.4							1.00			
	Resilient	636	19	557.1	0.86	0.54	1.36	-0.01	-0.02	0.01	1.18	0.72	1.94	0.50
	Sensitive	257	5	198.2	0.63	0.26	1.53	-0.01	-0.04	0.01	0.53	0.20	1.36	0.19
UKB	Ref (Expected)	12009	143	274.5							1.00			
	Resilient	345	4	7.9	0.97	0.36	2.63	-0.01	-0.52	0.49	1.31	0.47	3.62	0.60
	Sensitive	272	2	6.3	0.62	0.15	2.50	-0.20	-0.65	0.26	0.45	0.11	1.84	0.26
LDL														
VHU	Ref (Expected)	1921	91	1447.2							1.00			
	Resilient	68	4	49.8	1.28	0.47	3.48	0.02	-0.06	0.10	1.52	0.56	4.18	0.41
	Sensitive	114	12	90.8	2.10	1.15	3.84	0.07	-0.01	0.15	2.17	1.17	4.02	0.01
UKB	Ref (Expected)	12475	157	284.9							1.00			
	Resilient	97	0	2.2	-	-	-	-	-	-	-	-	-	-
	Sensitive	57	0	1.3	-	-	-	-	-	-	-	-	-	-
Total Cholesterol														
VHU	Ref (Expected)	12637	403	10425.3							1.00			
	Resilient	651	14	523.3	0.69	0.41	1.18	-0.01	-0.03	0.00	0.78	0.46	1.33	0.37
	Sensitive	466	39	405.2	2.49	1.79	3.46	0.06	0.03	0.09	2.23	1.60	3.10	2.0E-06
UKB	Ref (Expected)	12537	158	286.2							1.00			
	Resilient	34	0	0.8	-	-	-	-	-	-	-	-	-	-
	Sensitive	58	0	1.3	-	-	-	-	-	-	-	-	-	-

Trait/study	Categories	n	Events	person-years	IRR*	95% (CIs)		IRD*	95% (CIs)		HR	95% (CIs)		p
<i>Triglycerides</i>														
VHU	Ref (Expected)	10498	330	8592.6							1.00			
	Resilient	1	0	1.1	-	-	-	-	-	-	-	-	-	-
	Sensitive	378	16	311.0	1.34	0.81	2.21	0.01	-0.01	0.04	1.11	0.67	1.83	0.70
UKB	Ref (Expected)	11775	137	269.2							1.00			
	Resilient	468	6	10.7	1.10	0.49	2.50	0.05	-0.40	0.51	1.37	0.60	3.12	0.45
	Sensitive	376	8	8.5	1.85	0.91	3.77	0.43	-0.23	1.09	2.05	1.00	4.20	0.05
<i>SBP</i>														
VHU	Ref (Expected)	13021	408	10773.1							1.00			
	Resilient	474.00	6.00	377.83	0.71	0.10	5.05	-0.01	-0.06	0.04	0.46	0.21	1.04	0.06
	Sensitive	201.00	10.00	162.55	1.66	0.53	5.18	0.02	-0.05	0.10	1.63	0.87	3.05	0.13
UKB	Ref (Expected)	11961	136	273.1							1.00			
	Resilient	388	4	8.8	0.91	0.34	2.47	-0.04	-0.50	0.41	0.85	0.31	2.30	0.75
	Sensitive	291	7	6.7	2.10	0.98	4.48	0.55	-0.23	1.33	2.40	1.12	5.14	0.02

" - " it was not possible to estimate the number. IRR: incidence rate ratio. IRD: incidence rate differences. *Per 100,000 person-years; SBP: systolic blood pressure; DBP: diastolic blood pressure; HDL-C: High-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; CVD: Cardiovascular disease. T2D: Type 2 diabetes.

Table 3. Fully-adjusted Hazard ratios and 95%CI of prediction interval categories and T2D Fully-adjusted Hazard ratios and 95%CI of prediction interval categories and T2D

Trait/study	Categories	n	Events	person-years	IRR*	95% (CIs)		IRD*	95% (CIs)		HR	95% (CIs)		p
<i>Fasting glucose</i>	Ref (Expected)	13566	186	10838.4							1.00			
	VHU Resilient	574	8	485.1	0.96	0.47	1.95	-0.001	-0.01	0.01	0.73	0.18	2.97	0.67
	Sensitive	261	2	193.1	0.60	0.15	2.43	-0.01	-0.02	0.01	0.83	0.41	1.68	0.60
	Ref (Expected)	11567	110	264.1							1.00			
	UKB Resilient	567	4	13.4	0.72	0.26	1.94	-0.12	-0.42	0.18	0.74	0.27	2.01	0.55
	Sensitive	492	16	11.1	3.46	2.05	5.85	1.02	0.31	1.74	4.40	2.58	7.51	5.81E-08
<i>2-h Glucose/HbA1c</i>														
	Ref (Expected)	12210	110	10049.4							1.00			
	VHU Resilient	535	8	457.9	1.60	0.78	3.27	0.01	-0.01	0.02	1.42	0.69	2.91	0.34
	Sensitive	192	2	160.3	1.14	0.28	4.61	0.002	-0.02	0.02	1.46	0.36	5.94	0.59
	Ref (Expected)	11531	95	263.2							1.00			
	UKB Resilient	568	0	13.1	-	-	-	-	-	-	-	-	-	-
	Sensitive	516	46	11.7	10.89	7.66	15.49	3.57	2.43	4.71	8.48	5.78	12.44	6.96E-28
<i>DBP</i>														
	Ref (Expected)	13094	175	10851.3							1.00			

Trait/study	Categories	n	Events	person-years	IRR*	95% (CIs)		IRD*	95% (CIs)		HR	95% (CIs)		p
VHU	Resilient	456	4	374.1	0.66	0.25	1.79	-0.01	-0.02	0.01	0.94	0.39	2.28	0.89
	Sensitive	251	3	201.7	0.92	0.29	2.89	0.00	-0.02	0.02	0.78	0.25	2.44	0.67
	Ref (Expected)	11963	122	273.2							1.00			
UKB	Resilient	382	3	8.7	0.77	0.25	2.43	-0.10	-0.50	0.30	0.82	0.26	2.57	0.73
	Sensitive	282	1	6.5	0.34	0.05	2.47	-0.29	-0.60	0.02	0.43	0.06	3.10	0.40
HDL														
VHU	Ref (Expected)	1849	31	1413.3							1.00			
	Resilient	95	0	70.5	-	-	-	-	-	-	-	-	-	-
	Sensitive	90	2	66.1	1.38	0.33	0.54	0.13	2.16	-0.26	1.71	0.40	7.28	0.47
UKB	Ref (Expected)	12003	125	274.2							1.00			
	Resilient	287	10	6.7	0.31	0.16	0.59	-0.31	-0.43	-0.20	2.41	1.17	4.99	0.02
	Sensitive	319	1	7.2	0.03	0.00	0.23	-0.44	-0.53	-0.36	0.26	0.04	1.87	0.18
BMI														
VHU	Ref (Expected)	12921	161	10621.1							1.00			
	Resilient	636	2	557.1	0.24	0.06	0.96	-0.01	-0.02	-0.01	0.58	0.14	2.36	0.45
	Sensitive	255	12	196.7	4.02	2.24	7.23	0.05	0.01	0.08	0.64	0.27	1.53	0.32

Trait/study	Categories	n	Events	person-years	IRR*	95% (CIs)		IRD*	95% (CIs)		HR	95% (CIs)		p
UKB	Ref (Expected)	11993	120	274.2							1.00			
	Resilient	345	4	7.9	1.16	0.43	3.13	0.07	-0.43	0.57	1.63	0.40	6.73	0.50
	Sensitive	270	4	6.2	1.47	0.54	3.99	0.21	-0.43	0.84	0.40	0.14	1.18	0.10
LDL														
VHU	Ref (Expected)	1914	38	1442.3							1.00			
	Resilient	67	1	49.0	0.78	0.11	5.65	-0.01	-0.05	0.03	0.82	0.11	5.99	0.84
	Sensitive	112	1	89.9	0.42	0.06	3.08	-0.02	-0.04	0.01	0.39	0.05	2.85	0.35
UKB	Ref (Expected)	12462	146	284.6							1.00			
	Resilient	97	3	2.2	2.66	0.85	8.34	0.85	-0.69	2.40	2.93	0.93	9.22	0.07
	Sensitive	56	1	1.3	1.50	0.21	10.72	0.26	-1.25	1.77	1.31	0.18	9.36	0.79
Total Cholesterol														
VHU	Ref (Expected)	12610	150	10408.0							1.00			
	Resilient	649	9	522.4	1.20	0.61	2.34	0.003	-0.01	0.01	1.18	0.59	2.36	0.63
	Sensitive	463	11	404.0	1.89	1.02	3.49	0.01	-	0.003	1.70	0.92	3.14	0.09
UKB	Ref (Expected)	12524	138	285.9							1.00			
	Resilient	649	9	522.4	1.20	0.61	2.34	0.003	-0.01	0.01	1.18	0.59	2.36	0.63
	Sensitive	463	11	404.0	1.89	1.02	3.49	0.01	-	0.003	1.70	0.92	3.14	0.09

Trait/study	Categories	n	Events	person-years	IRR*	95% (CIs)		IRD*	95% (CIs)		HR	95% (CIs)		p
UKB	Resilient	34	0	0.8	-	-	-	-	-	-	-	-	-	-
	Sensitive	58	3	1.3	4.78	1.52	15.01	1.83	-0.79	4.44	5.05	1.24	20.59	0.02
Triglycerides														
VHU	Ref (Expected)	10469	148	8574.8							1.00			
	Resilient	1	0	1.1	-	-	-	-	-	-	-	-	-	-
	Sensitive	376	4	309.6	0.75	0.28	2.02	0.004	-0.02	0.01	0.69	0.25	1.86	0.46
UKB	Ref (Expected)	11763	120	268.9							1.00			
	Resilient	468	2	10.7	0.42	0.10	1.69	-0.26	-0.53	0.01	0.47	0.11	1.92	0.29
	Sensitive	376	7	8.5	1.85	0.86	3.95	0.38	-0.24	0.99	1.80	0.73	4.43	0.20
SBP														
VHU	Ref (Expected)	12996	170	10757.8							1.00			
	Resilient	474.00	5.00	377.83	0.84	0.34	2.04	0.003	-0.01	0.01	0.65	0.26	1.64	0.37
	Sensitive	200.00	5.00	161.41	1.96	0.81	4.77	0.02	-0.01	0.04	1.74	0.71	4.24	0.22
UKB	Ref (Expected)	11948	121	272.8							1.00			
	Resilient	388	5	8.8	1.28	0.52	3.13	0.12	-0.38	0.63	1.35	0.55	3.31	0.52
	Sensitive	291	0	6.7	-	-	-	-	-	-	-	-	-	-

Trait/study	Categories	n	Events	person-years	IRR*	95% (CIs)	IRD*	95% (CIs)	HR	95% (CIs)	p
" - " it was not possible to estimate the number. IRR: incidence rate ratio. IRD: incidence rate differences. *Per 100,000 person-years; SBP: systolic blood pressure; DBP: diastolic blood pressure; HDL-C: High-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; CVD: Cardiovascular disease. T2D: Type 2 diabetes.											

456

457 SUPPLEMENTARY INFORMATION

458 Table 1. VHU Criteria for exclusions on cardiometabolic traits.

459 Table 2. VHU Criteria for implausible values for lifestyle variables.

460 Table 3. VHU variable meaning.

461 Table 4. UKBB and VHU variable correspondence.

462 Table 5. Variables removed during data processing.

463 Table 6. 'No persistence' category hazard ratios (HR) and 95% CIs in VHU cohort.

464 Table 7. Hazard ratios (HR) and 95%CI of prediction interval categories and CVD.

465 Table 8. Hazard ratios (HR) and 95%CI of prediction interval categories and CVD mortality.

466 Table 9. Hazard ratios (HR) and 95%CI of prediction interval categories and diabetes.

467 Table 10. R packages used for the analyses in the current study.

468 Table 11. Fully-adjusted Hazard ratios and 95%CI of prediction interval categories and CVD-mortality.

469 Table 12 (.xlsx file). Mean squared error (MSE) and proportion falsely classified (PFC) for TC.

470 Table 13 (.xlsx file). Mean squared error (MSE) and proportion falsely classified (PFC) for SBP.

471 Table 14 (.xlsx file). Mean squared error (MSE) and proportion falsely classified (PFC) for LDL-C.

472 Table 15 (.xlsx file). Mean squared error (MSE) and proportion falsely classified (PFC) for BMI.

473 Table 16 (.xlsx file). Mean squared error (MSE) and proportion falsely classified (PFC) for HDL-C.

474 Table 17 (.xlsx file). Mean squared error (MSE) and proportion falsely classified (PFC) for DBP.

475 Table 18 (.xlsx file). Mean squared error (MSE) and proportion falsely classified (PFC) for FG.

476 Table 19 (.xlsx file). Mean squared error (MSE) and proportion falsely classified (PFC) for 2hG.

477 Table 20 (.xlsx file). Mean squared error (MSE) and proportion falsely classified (PFC) for TG.

478 Figure 1. Scatter plots of predicted and observed values of Fasting glucose (FG) in mmol/L.

479 Figure 2. Scatter plots of predicted and observed values of 2-hour glucose in mmol/L and HbA1c (UKB) mmol/mol.

480 Figure 3. Scatter plots of predicted and observed values of Systolic blood pressure (SBP) in mm/Hg.

481 Figure 4. Scatter plots of predicted and observed values of Diastolic blood pressure (DBP) in mm/Hg.

482 Figure 5. Scatter plots of predicted and observed values of high-density lipoprotein cholesterol (HDL-C) in mmol/L.

483 Figure 6. Scatter plots of predicted and observed values of low-density lipoprotein cholesterol (LDL-C) in mmol/L.

484 Figure 7. Scatter plots of predicted and observed values of total cholesterol (TC) in mmol/L.

485 Figure 8. Scatter plots of predicted and observed values of triglycerides (TG) in mmol/L.

486 Figure 9. Scatter plots of predicted and observed values of body mass index (BMI) in kg/m².