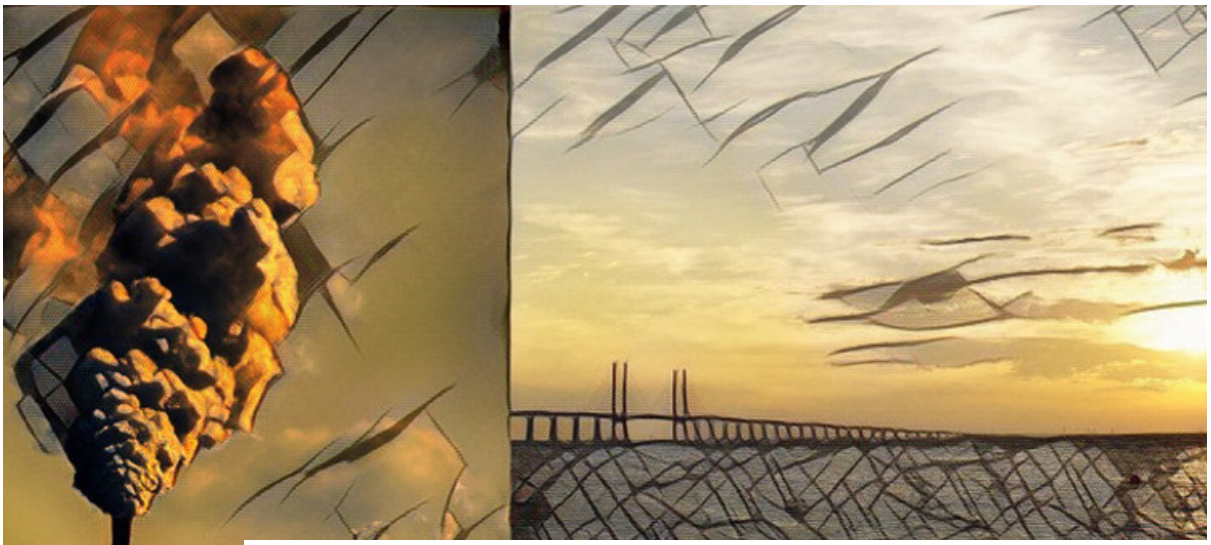
Optimizing Exposome-wide Assessments in Cardiometabolic Risk

Pomares-Millan, Hugo

2022

Document Version:
Publisher's PDF, also known as Version of record



Optimizing Exposome-wide Assessments in Cardiometabolic Risk

HUGO POMARES-MILLAN
DEPARTMENT OF CLINICAL RESEARCH | FACULTY OF MEDICINE | LUND UNIVERSITY

Printed by Media-Tryck, Lund 2022 NORDIC SWAN ECOLABEL 3041 0903

HUGO POMARES-MILLAN received his medical degree (MD) from the National University of Mexico (UNAM). In 2018, he graduated with a MSc in Public Health (MPH) from Lund University and completed his Ph.D. at the Genetic and Molecular Epidemiology unit at the Lund University Diabetes Center. His main interests lie within the fields of genetic epidemiology, cardiometabolic diseases and cancer. Hugo's thesis focuses on the causal effects of the exposome on cardiometabolic disease.

LUND UNIVERSITY

FACULTY OF MEDICINE

Department of Clinical Research

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2022:130
ISBN 978-91-8021-292-2
ISSN 1652-8220

Optimizing Exposome-wide Assessments in Cardiometabolic Risk

Optimizing Exposome-wide Assessments in Cardiometabolic Risk

Hugo Pomares-Millan

LUND
UNIVERSITY

DOCTORAL DISSERTATION

Doctoral dissertation for the degree of Doctor of Philosophy (Ph.D.) at the Faculty of Medicine at Lund University to be publicly defended on the 6th of October, 2022 at 13.00 in Agardhsalen, Clinical Research Centre, Jan Waldenströms gata 35, Malmö.

Faculty opponent

Professor Rikard Landberg

Chalmers University of Technology

Thesis advisors

Paul W. Franks, Giuseppe N. Giordano

Organization: LUND UNIVERSITY	Document name: DOCTORAL DISSERTATION
	Date of disputation: 2022-10-06
Author(s): Hugo Pomares-Millan	Sponsoring organization: RHAPSODY, NASCENT

Title:
Optimizing exposome-wide assessments in cardiometabolic risk

Abstract

This thesis is focused on cardiovascular disease (CVD) and type 2 diabetes mellitus (T2D), two concomitant conditions that appear with growing concern. In our work, we aim to improve the identification of individuals at-risk of cardiometabolic disease through the characterization of complex environmental exposures (i.e. diet, physical activity), that temporally vary, and the health effects on cardiometabolic traits and disease. Our projects were based upon the Västerbotten Health Survey (VHU) and the Malmö Diet and Cancer (MDCS) studies, which included extensive data on lifestyle, biological intermediates, and clinical outcomes.

In **Paper I**, we utilized the so-called environmental-wide association approach (EWAS), using longitudinal data from > 31,000 adults in VHU study. Under generalized linear models, from ~ 300 candidate exposures, 11 modifiable variables were associated with most of the cardiometabolic traits; the prioritised variables belonged to smoking, coffee intake, physical activity, alcohol intake, and context-specific lifestyle domains.

In **Paper II**, we implemented a machine learning-based model to identify individuals with variable susceptibility to lifestyle risk factors for T2D and CVD. Individuals with sensitivity to blood lipids, and blood pressure associated predictors were at higher risk to develop cardiometabolic disease. Furthermore, when pooling across sensitive groups from the two cohorts, the findings suggest a particular vulnerable subpopulation with different risk profile.

In **Paper III**, a series of causal-inference experiments from VHU and publicly available genome-wide association study (GWAS) summary statistics were used to triangulate evidence of the direct and mediated effects by adiposity and physical activity, of macronutrient intake (fat, carbohydrates, protein and sugar) and cardiometabolic disease. Using structural equation modelling, the mediation analyses enhanced with Mendelian randomization analysis, showed a likely causal putative association between carbohydrate intake and T2D. In addition, the integrative genomic analyses suggested a candidate causal variant localized to the established T2D gene *TCF7L2*.

In **Paper IV**, we conducted a systematic review and metanalysis of observational studies, complemented by Mendelian randomization analysis using GWAS summary statistics, investigating causal associations of individuals with high, yet normal, glycaemia associated with cardiovascular complications. Prediabetes was likely causally associated with coronary heart disease; suggesting higher, but not diabetic levels of blood glucose confer a risk, thus, effective preventive strategies may prove successful in prediabetes.

Key words: cardiometabolic risk, cardiovascular disease, type 2 diabetes, prediabetes, machine learning, causal inference, mediation, Mendelian randomization, nutritional epidemiology

Classification system and/or index terms (if any)

Supplementary bibliographical information		Language: English
ISSN and key title:1652-8220 Lund University, Faculty of Medicine Doctoral Dissertation Series 2022:130		ISBN 978-91-8021-292-2
Recipient's notes	Number of pages: 75	Price
	Security classification	

Signature  Date 2022-08-30

Optimizing Exposome-wide Assessments in Cardiometabolic risk

Hugo Pomares-Millan

LUND
UNIVERSITY

Cover photo by Hugo Pomares-Millan. Images courtesy of www.pixabay.com

Faculty of Medicine
Department of Clinical Research

ISBN 978-91-8021-292-2
ISSN 1652-8220

Lund University, Faculty of Medicine Doctoral Dissertation Series
2022:130

Printed in Sweden by Media-Tryck, Lund University
Lund 2022

To my parents, brothers and Ivy.

Table of Contents

List of publications

Poveda, A., **Pomares-Millan, H.**, Chen, Y., Kurbasic, A., Patel, C.J., Renström, F., Hallmans, G., Johansson, I. and Franks, P.W., 2022. Exposome-wide ranking of modifiable risk factors for cardiometabolic disease traits. Scientific reports, 12(1), pp.1-10.
PubMed PMID: 35260745.
https://doi.org/10.1038/s41598-022-08050-1

Pomares-Millan, H., Poveda, A., Atabaki-Pasdar, N., Johansson, I., Björk, J., Ohlsson, M., Giordano, G.N. and Franks, P.W., 2022. Predicting sensitivity to adverse lifestyle risk factors for cardiometabolic morbidity and mortality. Nutrients, 14(15), p. 3171.
PubMed PMID: 35956347.
https://doi.org/10.3390/nu14153171

Pomares-Millan, H., Atabaki-Pasdar, N., Coral, D., Johansson, I., Giordano, G.N. and Franks, P.W., 2022. Estimating the direct effect between dietary macronutrients and cardiometabolic disease, accounting for mediation by adiposity and physical activity. Nutrients, 14(6), p.1218.
PubMed PMID: 35334875.
https://doi.org/10.3390/nu14061218.

Mutie, P.M.*, **Pomares-Millan, H.***, Atabaki-Pasdar, N., Jordan, N., Adams, R., Daly, N.L., Tajes, J.F., Giordano, G.N. and Franks, P.W., 2020. An investigation of causal relationships between prediabetes and vascular complications. Nature communications, 11(1), pp.1-11.
PubMed PMID: 32929089.
https://doi.org/10.1038/s41467-020-18386-9.
** contributed equally*

Publications not included in this thesis

Atabaki-Pasdar, N., Ohlsson, M., Viñuela, A., Frau, F., **Pomares-Millan, H.**, Haid, M., Jones, A.G., Thomas, E.L., Koivula, R.W., Kurbasic, A. and Mutie, P.M., 2020. Predicting and elucidating the etiology of fatty liver disease: A machine learning modeling and validation study in the IMI DIRECT cohorts. PLoS medicine, 17(6), p.e1003149.

Franks, P.W. and **Pomares-Millan, H.**, 2020. Next-generation epidemiology: the role of high-resolution molecular phenotyping in diabetes research. Diabetologia, 63(12), pp.2521-2532.

Wilman, H.R., Parisinos, C.A., Atabaki-Pasdar, N., Kelly, M., Thomas, E.L., Neubauer, S., Jennison, C., Ehrhardt, B., Baum, P., ….**Pomares-Millan, H.,** Schoelsch, C., and Freijer, J., 2019. Genetic studies of abdominal MRI data identify genes regulating hepcidin as major determinants of liver iron concentration. Journal of hepatology, 71(3), pp.594-602.

Acknowledgements

First, **Paul W. Franks** my mentor and supervisor, thank you. After a few years of this amazing journey, I am grateful that during the last months of my master's you replied to my email and we met. After some time, you were already helping me with the master's thesis, guiding me into this wonderful research journey. **Paul**, from the beginning you were more than just a mentor, you inspired me as a leader and scientist to address research problems creatively and aim for greater goals. You were always thoughtful, supportive, and attentive to my personal and professional growth, I will never forget that every time we were holding our weekly meetings, the opening question was about how I was doing outside the workplace. Thanks for your mentoring and friendship.

I am also thankful to **Giuseppe (Nick) Giordano** who took the leadership of the GAME unit after changes came into place. **Nick**, you were always interested in the well-being of the members of the unit, and I admire how you managed more responsibilities during the transition and kept the group moving forward, therefore, thank you. To **Anders Rosengren**, my other co-supervisor, I thank your trust deposited in me, although we briefly met and chat, it was heartening to know you have my back during the half-time review.

I would like to express my gratitude to the fabulous members and former members of the Genetic and Molecular Epidemiology (GAME) unit. **Alaitz Poveda**, thanks for showing me how to keep order in the analysis pipeline, neat scripts, and how to use many bioinformatic tools. **Tibor V. Varga**, more than a colleague you are a good friend always with a great advice. **Naemieh Atabaki-Pasdar**, it was a pleasure to work together with you, I am thankful for your help and support on my projects; "**Naemieh-nem"**, I admire your intelligence, tenacity, and kindness; thank you for your friendship and the good laughs. **Pascal Mutie**, we collaborated on my first Ph.D. project, Amigo! thanks for your interesting conversations and good humour; **Hugo Fitipaldi**, it was an honour to share this journey by your side since the MPH, I admire your curiosity, coding (and musical) skills, and determination to get things done. **Daniel Coral**, I appreciate your kindness, dedication, and willingness to help always with a big smile and great attitude. **Sebastian Kalamajski**, thanks for the weekly lunch company (either Indian or falafel) and the jokes while tasting a beer; **Mi Huang**, I admire your courage and capacity to surprise, thanks for showing me how to keep humble. **Pernilla Siming**, even after leaving LU, you will always be part of the unit and my training, thank you for helping me on how to navigate the

Swedish bureaucracy; **Marketa Sjögren**, thanks for being very thoughtful and being always willing to help, you are a brilliant scientist and a caring person. To **Juan Fernandez Tajes** and **Neli Tsereteli**, thanks for helping me to either debug my scripts or visualize my findings, your input was crucial. **Angela Estampador** and **Robert Koivula** thanks for your openness to discussing research questions. Thanks for the good memories to all my friends who made my life easier in Sweden (**Marie, Maria, Soley, John, Jeanette**) and thanks to those who our paths crossed during my studies and contributed to my growth (**Siham, Masoud, Simon Timpka, Peter M. Nilsson, Nicole Prinz, Ewan Pearson, Ian Forgie, Ana Viñuela,** among many, many others).

Thanks to all my co-authors, especially to **Ingegerd Johansson**, I hope one day to meet you in person, but thanks for all your support; every time I received reviewers' feedback you were always the first willing to help me. Thanks to the LUDC faculty members and staff, I greatly appreciate the support from **Mattias Borell** and **Johan Hultman**, especially when using the server. To my friend and colleague from another research unit: **Esther Gonález-Padilla**, your passion in nutritional epidemiology is inspirational. More recently I would like to thank **Brock Christensen, Michael Pasarrelli**, and the Epi department at Dartmouth for welcoming and opening the door to cancer research, **Brock** thank you for your trust and let me join your amazing lab during my visit.

Finally, to **Ivy Lorena**, the love of my life, this process would not have been possible without your love, example, and support despite the distance. Thanks to my brothers, **César** and **Oscar**, for always being supportive and encouraging to me during this stage of my life. There are not enough words to thank my parents, you always gave me unconditional love and support; my mother **María Elizabeth** thank you for your patience and understanding, you always guided me in rough times; to my father **Ramiro**, you were always proud and loving with me no matter what. I owe this accomplishment to you both.

Hugo Pomares-Millan
New Hampshire, August, 2022

Popular summary

Our lifestyles influence, to a large extent, our health status. However, many other factors also affect our health, including genetics and social determinants, such as education and economy. Currently, energy-dense processed foods are often cheaper than healthier alternatives. Moreover, with wide-spread automation of tasks that only a generation ago involved physical labour, sedentary behaviours are becoming more frequent. As a result, the individual is exposed to adverse environments that have, in recent decades, increased the prevalence of overweight/obesity. Although the connection of environmental exposures and cardiometabolic disease has been widely explored, many unanswered questions remain. For instance, (i) Which exposures cause specific diseases? (ii) Which exposures are the most important to intervene upon? (iii) To what extent do specific exposures affect disease risk? (iv) Which are the optimal approaches to disease prevention? each of these questions require evidence-based responses, if the burden of chronic disease that blights so many societies worldwide is to be adequately addressed.

Environmental exposures are often studied when there is prior evidence of association between specific agents and disease (e.g. saturated fat in the diet and atherosclerosis). However, to identify and elucidate unrecognized risk factors, hypothesis-free data-driven approaches may be required. One such example of novel associations related to cardiometabolic diseases are exposures related to heavy metals and sleep patterns. In the study described here, we conducted an analysis of hundreds of exposures in relation to cardiovascular biomarkers. Thereafter, we obtained a shortlist of candidate variables for further investigation. We found that the strongest signals were for well-established risk factors for cardiometabolic disease (cardiovascular disease, (CVD) and type 2 diabetes mellitus (T2D)). Physical activity, smoking, and overall health status explained most of the variation in cardiovascular biomarkers. Thus, it is plausible that intervening upon these modifiable variables will have a larger impact when preventing disease.

However, correlation does not always reflect causality. The presence of an exposure-outcome association does not necessary mean that the exposure directly impacts the outcome. Therefore, observational associations help inform the design of interventions but may fail if the factors upon which the interventions focus are not causally related with the outcomes of interest.

Causal relationships are generally determined at a population level, yet there is often considerable heterogeneity between individuals in exposure and response relationships, with some people incurring large effects and others little or no effect. Clinical algorithms sometimes use cut-off values based on biomarker levels to diagnose disease; thus, conventional diagnostic approaches may overlook people with underlying pathology that does not meet diagnostic criteria. To help address some of the heterogeneity in disease presentation, methods to reclassify some diseases (e.g. T2D and obesity) into narrower diagnostic categories may be needed.

In this project, we used a quantile approach to identify individuals that were at higher risk of disease based on their predicted cardiometabolic biomarker values. We evaluated those subgroups of the population with elevated predicted values of blood glucose, lipids and blood pressure. We found that these individuals were at higher risk of heart disease and premature death. We also found that individuals living with prediabetes have a higher risk to develop disease. We found that elevated, but non-diabetic, blood glucose levels, often considered relatively harmless, are causally related with clinical complications. However, prediabetes was not causally associated with other diabetes complications such as stroke or kidney disease.

In analyses focusing on the causal effects of modifiable exposures, we found evidence of a causal effect of carbohydrate intake and T2D, suggesting a higher risk if the diet composition is predominantly carbohydrate-based over other macronutrients (e.g. fat, protein). However, we could not disentangle the independent effects of sugar intake (a type of carbohydrate often associated with higher risk of diabetes) or fibre (associated with a lower risk of diabetes). These findings imply that when recommending nutritional strategies to prevent disease, the direct impact of carbohydrates should be considered if individuals are especially sensitive to the environment or have prediabetes, emphasizing the importance of maintaining a healthy and balanced diet to avoid health complications.

Overall, the exposures we are subjected to in our daily lives have a varying influence in our health status. We demonstrated causal associations between prediabetes and CVD and carbohydrate intake and T2D. Moreover, we ranked the most important variables for further investigation and used these to elucidate subgroups of the population that are at higher risk of disease without been clinically recognized by conventional screening methods. Our findings emphasize the notion that adopting and maintaining a healthy lifestyle is beneficial to prevent cardiometabolic disease and mortality.

List of abbreviations

2-hour glucose - 2-hr glucose

2SLS - two-stage least squares regression

AHA - American Heart Association

ADA - American Diabetes Association

BMI - body mass index

CVD - cardiovascular diseases

CIs - confidence intervals

CHD - coronary heart disease

CKD - chronic kidney disease

DBP - diastolic blood pressure

DIAGRAM - DIAbetes Genetics Replication and Meta-analysis Consortium

DIAMANTE - DIAbetes Genetics Replication and Meta-ANalysis-TransEthnic

E% - percentage of energy

EWAS - environment-wide association study

FDR - false discovery rate

FG - fasting glucose

FFQ - food frequency questionnaire

GDM - gestational diabetes mellitus

GRS - genetic risk score

GWAS - genome-wide association study

HbA1c - glycated haemoglobin

HR - hazard ratio

HDL-C - high-density lipoprotein cholesterol

IFG - impaired fasting glucose

IGR - impaired glucose regulation

IGT - impaired glucose tolerance

IV - instrumental variable

ICD - International Classification of Diseases

IDF- International Diabetes Federation

IVW - inverse variance-weighted

LD - linkage disequilibrium

LDL-C - low-density lipoprotein cholesterol

ML - machine learning

MDCS - Malmö Diet and Cancer Study

MDCS-CC - Malmö Diet and Cancer Study cardiovascular cohort

MODY - Maturity-Onset Diabetes of the Young

MR - Mendelian randomization

MAGIC - Meta-Analyses of Glucose-and Insulin-related traits Consortium

MetS - metabolic syndrome

MI - myocardial infarction

PRS - polygenic/partitioned risk score

PIs - prediction intervals

PCA - principal component analysis

QRF - quantile regression forest

ROC - receiver operating characteristic

SNP - single nucleotide polymorphism

SEM - structural equation modelling

SBP - systolic blood pressure

TEI - total energy intake

T1D - type 1 diabetes mellitus

T2D - type 2 diabetes mellitus

TCF7L2 - transcription factor 7-like 2

VHU - Västerbottens Health survey

Chapter 1

Introduction

Associations between environmental exposures and disease have been at the core of epidemiology since its inception. In recent decades, the 'exposome-wide' paradigm has been advanced by the rapid development of cheaper and accessible phenotyping technologies (e.g. fitness wearables, sleep trackers) that can generate large amounts of data for researchers to examine[1].

Though existing therapies are fairly safe and efficacious leading to an improvement in healthcare management and life-expectancy, chronic diseases such as type 2 diabetes mellitus (T2D) and cardiovascular disease (CVD) remain among the most frequent causes of death in all countries, irrespective of income[2]. Although CVD death rates have been steadily declining in parallel with improved healthcare[3], the clinical presentation, complications and response to interventions are highly heterogeneous within and between populations, indicating that 'one-size-fits-all' interventions are suboptimal. In recognition of this, a field termed 'precision medicine' has been born[4].

The primary goal of precision medicine is to ensure *"the right therapy, for the right individual, at the right time"*. Though physicians have been conducting individualised care for decades, the plethora of data derived from large, well-characterised studies has allowed to operationalize the partition of individuals into subclasses (clusters or subgroups) according to their risk and the variation in biomarkers (i.e. omics). This has been done for prediabetes[5], T2D[6, 7, 8] and CVD[9], with these approaches showing promise for enhanced risk prediction and treatment stratification.

Traditionally, traits such as body mass index (BMI), high- and low-density lipoprotein cholesterol (HDL-C and LDL-C, respectively), triglycerides, total cholesterol, fasting glucose (FG), 2-hour glucose (2-hr glucose), and systolic and diastolic blood pressures (SBP and DBP, respectively) are used as biomarkers for cardiovascular risk assessment. However, cardiometabolic conditions such as CVD and T2D are diseases driven by the complex interplay between genetic predisposition and a myriad of environmental exposures.

Population variance in disease explained by genetic factors is limited by the discovery of genome-wide variants or single nucleotide polymorphism (SNPs); for

complex traits, SNPs typically have small effect sizes. Moreover, it is unlikely that genetic contributions outweigh the effects of environmental factors in disease onset. The predominant role of lifestyle in cardiometabolic disease development is mostly drawn from results of observational studies and clinical trials, which suggest that CVD and T2D could, to a large extent, be prevented by adopting and maintaining a healthy lifestyle. Such changes convey more sustained and profound beneficial effects on health compared with pharmacotherapy[10, 11].

Intricacies in exposome-wide assessments include diversity in exposure sources (e.g. physical, chemical agents, etc.), the dynamic nature of risk factors (rarely remain the same through lifetime) in contrast with fixed genetic factors (see Figure 1), and statistical challenges to harmonise and analyse exposures across populations.

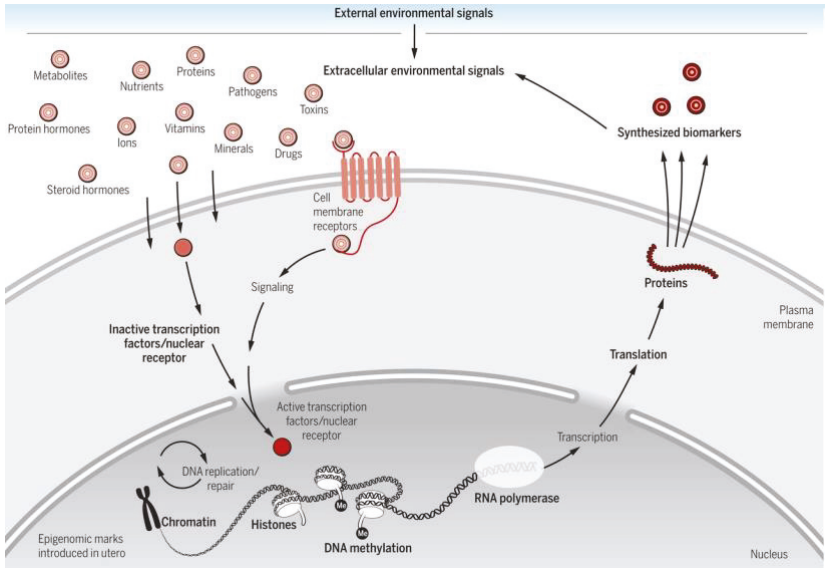


Figure 1. Mechanisms through which the environment interacts with the genome to affect health. Source reference[12].

The first section of this thesis describes the hallmarks of T2D and CVD. Section two, through a brief review of the literature, provides an overview of the mechanisms underlying both conditions; the third section revisits common approaches for studying the aetiology of these diseases. The last section summarise the findings of the papers included in the thesis and elaborate on prospective plans. Overall, this work explores the independent and likely causal effects of risk factors and their roles in disease onset, with focus on data-driven methods to identify subgroups of the population at higher risk of disease.

Cardiometabolic risk

Cardiometabolic disorders represent a group of interrelated risk factors (e.g. LDL-C, BMI, etc.) that may lead to the clinical outcomes of T2D and CVD, both of which are modifiable and preventable diseases. Early attempts to group risk factors into singular scores include the metabolic syndrome (MetS) (also known as 'syndrome X'), which helps screen people at risk of T2D and CVD. However, there has been extensive criticism surrounding the true clinical value of MetS[13], and it is rarely used today in research or practice.

Many initiatives across the globe have called for action to prevent CVD[14]. However, many challenges are posed by multifaceted conditions like MetS, highlighting the need for comprehensive strategies to prevent it. One example is the 'Life's Simple 7' from the American Heart Association (AHA), which summarizes the cardiometabolic risk factors (see Figure 2) and recommends interventions to lower a person's likelihood of developing CVD (i.e. cease smoking, improve diet, increase physical activity, maintain a healthy weight; lower blood pressure, cholesterol, and blood glucose levels)[15, 16].

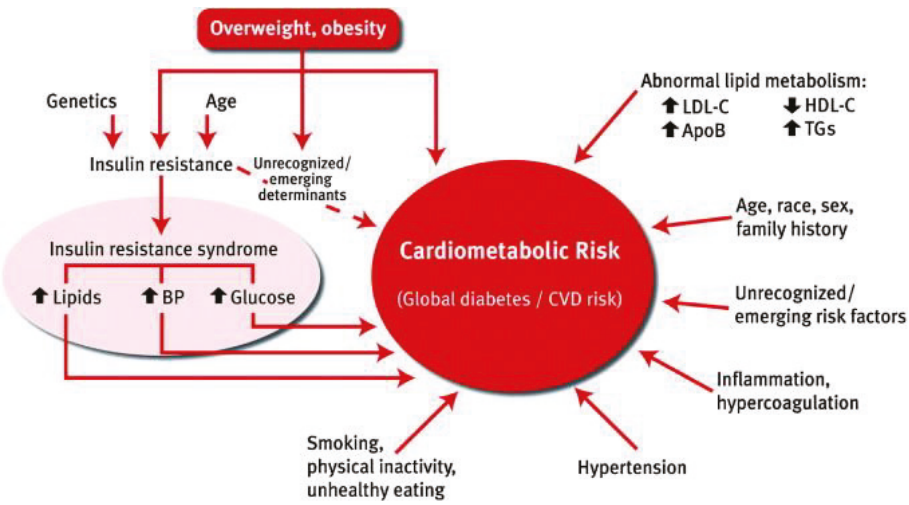


Figure 2. Factors contributing to cardiovascular disease and T2D risk. Source reference[17].

Diabetes and prediabetes

A common unifying feature of diabetes is chronically elevated blood glucose (hyperglycaemia) whilst fasting or following consumption of food or beverages containing carbohydrates. Insulin is the hormone produced in the pancreas that facilitates glucose transport from the blood into cells in the body, where glucose can either be burned or converted to stored energy. The two driving features of hyperglycaemia are insufficient insulin secretion and the inability of the tissues to respond to insulin (i.e. insulin resistance). Providing enough insulin is secreted, blood glucose concentrations may remain within the normal range even if cells are resistant to insulin, and vice versa. However, a reduction in insulin production coupled with insulin resistance typically results in hyperglycaemia. The criteria to diagnose the most common form of diabetes (i.e. T2D) which accounts for > 90% of all diabetes cases worldwide[18] are based on glycaemic measures: (i) FG ≥ 7.0 mmol/L (126 mg/ dL), or (ii) 2-hr glucose levels ≥ 11.1 mmol/L (200 mg/ dL), or (iii) glycated haemoglobin (HbA1c) ≥ 6.5% (48 mmol/mol)[19, 20]. Moreover, elevated yet non-diabetic glucose levels have been grouped under the term 'prediabetes', which comprises impaired glucose tolerance (IGT) defined as 2-hr glucose of 7.8 to 11.0 mmol (140 to 199 mg/dL) on the 75-g oral glucose tolerance test, and impaired fasting glucose (IFG), defined as 5.6 to 6.9 mmol/L (100 to 125 mg/dL) when fasting[21], and impaired glucose regulation (IGR) when IFG and IGT co-occur.

Diabetes and prediabetes cut-off values for detection remain controversial (summarized in Table 1). The term 'prediabetes' has been criticized because it implies that a person with this condition will inevitably progress to full-blown diabetes, and 'medicalizing' the prediabetic state may adversely affect attempts to minimize its detrimental impact[22]. Alternatively, because prediabetes is not typically considered a disease, but a risk factor for disease, this can impact the extent to which clinical interventions are initiated and/or adopted. Perhaps more importantly, though, there are multiple tissue- or organ-specific defects that can cause blood glucose to rise, with glucose levels merely one indicator of these underlying pathologies. Thus, the 'glucocentric' perspective that is embodied in the concept of prediabetes may inhibit recognition and use of other biomarkers that aid the prediction and prevention of T2D.

Table 1. Type 2 diabetes and prediabetes diagnostic criteria values.

	WHO/IDF	ADA
Type 2 diabetes mellitus (T2D)		
Fasting plasma glucose	≥ 7.00 mmol/L	≥ 7.00 mmol/L
	126 mg/dL	126 mg/dL
2-hr glucose	OR, ≥ 11.1 mmol/L	OR, ≥ 11.1 mmol/L
	200 mg/dL	200 mg/dL

	WHO/IDF	ADA
Impaired glucose tolerance (IGT)		
Fasting plasma glucose	< 7.00 mmol/L	Not required
	126 mg/dL	
2-hr glucose	AND, 7.8 to 11.0 mmol/L	7.8 to 11.0 mmol/L
	140 to 199 mg/dL	140 to 199 mg/dL
Impaired fasting glucose (IFG)		
Fasting plasma glucose	6.1 to 6.9 mmol/L	5.6 to 6.9 mmol/L
	110 to 125 mg/dL	100 to 125 mg/dL
2-hr glucose	If measured, < 7.8 mmol/L	If measured, < 7.8 mmol/L
	140 mg/dL	140 mg/dL

WHO: World Health Organization; IDF: International Diabetes Federation; ADA: American Diabetes Association; 2-hr glucose concentration after ingestion of 75-g of glucose load.

Other less common forms of diabetes include type 1 diabetes mellitus (T1D), which is driven by an autoimmune reaction that leads to destruction of the pancreatic β-cells[23], and gestational diabetes (GDM), where it is believed that pregnancy hormones raise glucose to levels that exceed the exocrine function of the pancreas[24, 25]. These classical diagnostic categories provide the framework for clinical guidelines. However, because T2D is essential a diagnosis of exclusion (of the known causes of hyperglycaemia), and because it is by far the most common form of diabetes, there is an unmet need to refine the diabetes phenotype. Several recent studies have attempted to do this, using combinations of data, much of which is not used in the classical diagnosis such as genetics, omics, and non-glucose biomarkers[6, 7, 26, 27]. Moreover, some of these newly identified diabetes subtypes share clinical features with those autoimmune and rare forms of monogenic diabetes (e.g. Maturity-Onset Diabetes of the Young (MODY)). Other infrequent forms of diabetes include neonatal and secondary diabetes (i.e. drug-induced)[28].

Glucose metabolism

The preferred source of energy for human cells is glucose. Glucose metabolism after meal ingestion (postprandial) follows two core pathways: storage (glycogenesis) or breakdown (glycolysis). Both processes work in a sophisticated feedback system that maintains the stable supply of glucose to the brain, despite significant transitory

shifts in the supply of glucose to other organs and tissues that occur, for example, when eating or exercising.

Blood glucose levels are the product of several processes: (i) the intestinal absorption from macronutrient intake, (ii) the liberation of glucose from glycogen (glycogenolysis), and (iii) synthesis of glucose from non-carbohydrate substrates, mainly in the liver (gluconeogenesis). Glycogenesis is the formation of glycogen in liver or muscle tissue, similar to lipogenesis where energy in the form of lipids are stored in adipocytes. Conversely, in glycolysis, glucose is converted to adenosine triphosphate (ATP) by cellular respiration, the basic constituent of energy in the body to function[29].

Glucose homeostasis is finely regulated by the complementary action of two peptide hormones secreted from pancreatic islets: glucagon and insulin; the latter being secreted by β-cells in response to increased blood glucose and amino acid levels after substrate intake. The action of insulin, when bound with its receptor, promotes glucose uptake in peripheral tissue and increases gluconeogenesis in adipose tissue and liver, whilst glucagon (secreted from pancreatic α-cells) inhibits glucose output from glycogenolysis and gluconeogenesis[30]. Insulin has other important effects; it inhibits intracellular lipase, driving release of fatty acids (lipolysis) and uptake of triglycerides by adipose tissue. Other pancreatic hormones such as somatostatin, ghrelin and amylin, also have less marked but nevertheless essential roles in glucose homeostasis[31].

In prediabetes and T2D, an underlying condition that disrupts glucose homeostasis is insulin resistance where response of glucose-sensitive tissue is impaired. The mechanistic causes remain poorly understood, but being overweight/obese is a common antecedent (see Figure 3)[32]. A likely mechanism is the peripheral tissue resistance which leads to continuous insulin secretion. Moreover, as a compensatory response to hepatic and muscle insulin resistance, the pancreas will produce larger quantities of insulin, a state known as hyperinsulinemia[33]. Often, physical activity, irrespective of the exercise regime improves peripheral sensitivity which reduces insulin requirements during fasting status, moreover, other mechanisms such as glucose uptake independent of insulin (GLUT 4 transportation) are enhanced[34]. Finally, as time progresses, the β-cells fail to overcome insulin resistance, which leads to chronic hyperglycaemia and often T2D.

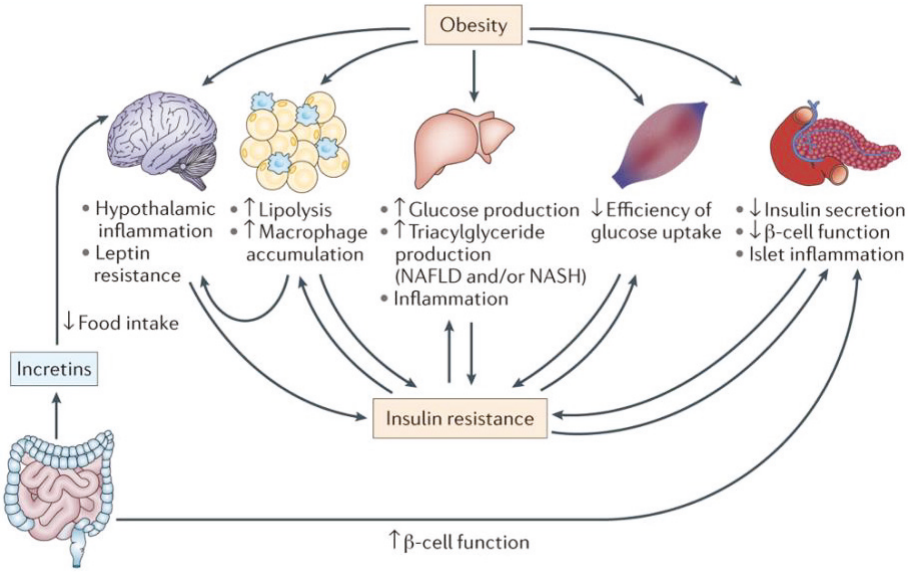


Figure 3. Obesity's contribution to systemic insulin resistance and β-cell dysfunction. Source reference[32].

Years before T2D onset (and during 'prediabetes'), the body relies on the over-production of insulin to counterbalance insulin resistance (mainly in peripheral tissue) to maintain glucose homeostasis. From studies assessing insulin resistance through the euglycemic hyperinsulinemic clamp (gold-standard for assessing insulin sensitivity)[35], in individuals with isolated IFG, hepatic insulin sensitivity is reduced whilst muscle sensitivity is almost normal. This suggests inadequate glucose output suppression. In contrast, in IGT individuals, hepatic insulin resistance is lower than in individuals with IFG, but with markedly reduced peripheral insulin sensitivity[36]. Overall, insulin resistance contributes to CVD through promoting the release of free fatty acids, apolipoprotein B (apoB) production in the liver, and further synthesis of low- and very-low density lipoproteins (VLDL)[37].

Insulin influences blood pressure in multiple ways. However, persistent elevated levels of insulin that are often observed in obesity, have been consistently associated with hypertension. Despite no clear mechanistic pathway, the activation of the sympathetic nervous system and reduction in urine excretion coincide with an inflammatory response mediated through cytokines such as tumour necrosis factor α (TNF-α), leptin, and interleukin-6 (IL-6). This eventually contributes to endothelial dysfunction and increased overall CVD risk[38]. Other reported downstream biomarkers in inflammation such as fibrinogen, factor VII, and

plasminogen activator inhibitor 1 (PAI-1), are also involved in thrombotic processes[39].

Cardiovascular disease (CVD)

CVD encompasses: (i) coronary heart disease (CHD), including myocardial infarction (MI), angina, and coronary death; (ii) cerebrovascular disease, defined by stroke and transient ischemic attack; (iii) peripheral arterial disease (e.g. claudication); and, (iv) atherosclerosis. Worldwide, CVD remains the leading cause of morbidity and mortality, despite improvements in life-expectancy. CHD accounts for almost half of total CVD cases, being the top cause of death in adults, globally[40]. Since the 1980s, CHD-mortality has declined, owing major progress in healthcare. However, its prevalence is on the rise due to increased aging populations, coupled with higher rates of T2D[2, 41]. CVD prevention strategies often include promoting healthy habits, and eradicating unhealthy lifestyles (e.g. poor-quality diet, physical inactivity, smoking) before or in parallel with pharmacotherapy[42]. With such measures, up to 80% of premature CVD events can be prevented[43, 44].

CVD risk factors

The majority of individuals in the general population have one or more risk factors for CVD. The main five leading modifiable risk factors (hypercholesterolemia, diabetes, hypertension, obesity, and smoking) contribute to more than half of cardiovascular death[45]. Total cholesterol ≥ 6.22 mmol/L (≥ 240 mg/dL), SBP ≥ 140 mmHg, DBP ≥ 90 mmHg, smoking, and T2D are considered major risk factors[46]. Other independent factors include social determinants of health and genetic burden[47]. Lowering risk factors is fundamental to disease prevention (i.e. lowering total cholesterol and LDL-C levels reduce CHD events and mortality). From large international cohorts (e.g. MONICA project[48], INTERHEART study[49]), which focused on changes in risk factors with lipid-lowering medication for primary and secondary prevention, results showed the importance of maintaining optimal levels of risk markers to prevent premature death[49, 50]. In the INTERHEART study, after adjustment for conventional CVD risk factors, every 1% increase in HbA1c was associated with 19% higher chances for MI[51, 52]. Therefore, screening people at risk of T2D (i.e. individuals living with overweight/obesity, and those considered prediabetic), and monitoring cardiometabolic risk biomarkers may help prevent complications.

Another major risk factor is the lack of physical activity. Exercising has beneficial effects in reducing blood pressure and promoting weight loss[53]. Obesity, which is defined as BMI > 30 kg/m^2, is highly prevalent worldwide. This condition is associated with a number of risk factors for atherosclerosis, CVD and hypertension, including insulin resistance, prediabetes, and dyslipidaemia.

Link between T2D and CVD

Insulin resistance, hyperinsulinemia and elevated blood glucose are all associated with CVD. Only having a diagnosis of diabetes has been considered equivalent to having a prior MI event in terms of relative risk[54]. The mechanism linking T2D and CVD is multifactorial, though the main pathway is through atherosclerosis, which is responsible for almost all cases of CHD (see Figure 4)[39]. This pathological process begins with vascular fatty deposits in major blood vessels during adolescence (or even earlier), which progress into plaques and eventual thrombotic occlusion. The formation of such lipid and atheromatous plaques promote smooth muscle cell proliferation and the recruitment of macrophages and inflammatory proteins in the intima, such as TNF-α and IL-6. In turn, these alter the release of vasoactive molecules (i.e. nitric oxide (NO)), thus consolidating the atherosclerotic plaque and subsequent overall vasoconstriction[55]. Local oxidative stress (through reactive oxygen species (ROS)) on lipoproteins increase the susceptibility of the plaque to rupture, with further ischaemia and necrosis[56, 57]. The rupture of the plaque activates platelets and thrombin formation that eventually leads to thromboembolism events. All these processes are accelerated in T2D, in the so-called 'prothrombotic' state where sustained hyperglycaemia promotes vascular dysfunction through oxidative stress[58]. Moreover, either prolonged blood glucose levels or excursions (spikes) in glucose levels, impair fibrinolytic proteins[59].

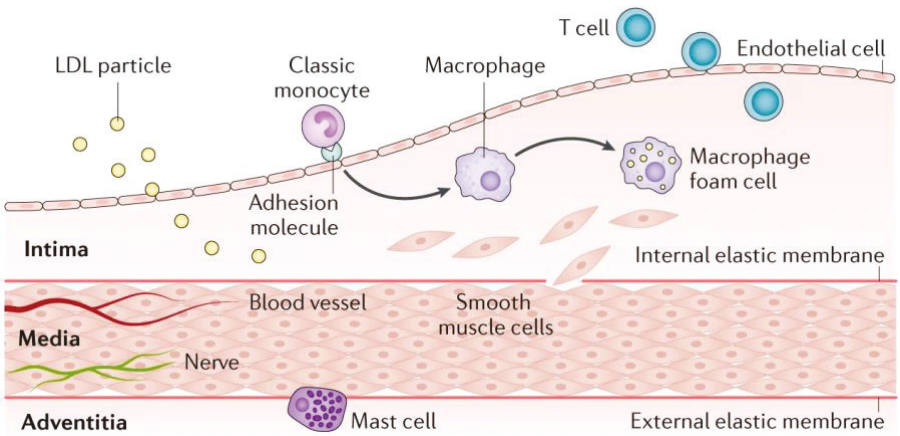


Figure 4. Atherosclerotic plaque initial process. Cell mediators that contribute to vessel occlusion and thrombosis. Source reference[39].

Of relevance here is that if no appropriate correction is established, sustained high blood glucose levels and circulating free fatty acids stimulate higher intracellular

concentrations of ROS leading to mitochondrial DNA damage, membrane permeability alterations, and apoptosis[60]. These processes confer systemic and organ-specific damage to peripheral nerves, blood vessels, heart, eyes, and kidneys. These pathogenic processes render T2D among the leading causes of non-traumatic lower limb amputation, chronic kidney disease (CKD), and visual impairment in relatively young populations[61].

Genetic and environmental factors

Cardiometabolic diseases have genetic components. Genome-wide association studies (GWAS) have revealed thousands of relatively independent DNA variants (or SNPs) associated with cardiometabolic traits[62]. For T2D, large-scale GWAS, mainly in populations of European descent, have identified reproducible genetic associations linked with the cardiometabolic disease. Efforts like DIAGRAM (DIAbetes Genetics Replication and Meta-analysis) and DIAMANTE (DIAbetes Genetics Replication and Meta-ANalysis-TransEthnic) consortia collected data from 74,124 T2D cases and 824,006 controls[63]. For continuous glycaemic traits (i.e. FG[64], 2-hr glucose[65], and HbA1c[66]), genetic association data from the Meta-Analyses of Glucose-and Insulin-related traits (MAGIC) consortium have been made publicly available to the research community. For CVD, several consortia exist; to date, the largest is the Coronary Artery Disease (C4D) Genetics (CARDIoGRAMplusC4D)[67], which includes 60,801 cases and 123,504 controls; GWAS of stroke have been conducted by the MEGASTROKE consortium[68].

In the context of T2D, the latest metanalysis of GWAS[63] has discovered up to 403 relatively independent loci associated with the disease (most identified in European-ancestry individuals), yet fewer than 20 loci localize in coding regions, many of the remaining variants are implicated in transcriptional regulation [69]. In CVD, the latest GWAS discovered 163 loci at genome-wide level significance associated specifically with CHD[70, 71, 72], and 35 loci for stroke[73]. As in T2D, most of the variants are located in non-coding regions, being related to CHD development by affecting different pathways such as acute phase response signals during inflammation[74]. Many genetic variants that confer risk of CHD or stroke also influence other traits, such as BMI, SBP and DBP. Moreover, some variants overlap with monogenic forms of the disease[68, 75], which have aided in our understanding of putative mechanisms.

In the context of precision medicine, GWAS discoveries have led to hypothesise about mechanisms of action. For example, the genes *LDLR*, *PCSK9*, *ANGPTL4*, and *ANGPTL3* have been explored as targets for drug development to prevent CHD[76]. In contrast, for T2D, the main use of genetics has been to develop genetic, polygenic, and partitioned risk scores (GRS and PRS, respectively), which use hundreds of variants to predict disease. However, when genetic markers are added to clinical risk scores, the predictive value increases modestly[77, 78].

Despite the many association signals discovered using GWAS, these explain only a small fraction of disease susceptibility. The estimated heritability from linkage analyses (parent to offspring) vary greatly for T2D (20 to 80%)[79], stroke (30 to 40%)[80], and CHD (40 to 60%)[81]. The so-called 'missing heritability' is likely due to the common presence of undetected variants (with minor allele frequencies (MAF) ≥ 0.01) and/or rare variants not captured or imputed by standard chips[82], as well as measurement error. CVD and T2D, like other complex diseases, develop as a consequence of exposure to environmental risk factors in combination with genetic susceptibility. Thus, to study these metabolic diseases, where environmental factors converge with genomic factors to confer risk, requires a systematic and comprehensive assessment to identify causal exposures and particular subgroups at elevated risk[83, 84].

Aims

Current evidence supports the notion of varying degrees of cardiometabolic risk within populations, emphasised by distinctive subphenotypes of disease. The contribution of lifestyle to disease susceptibility suggests several mechanistic hypotheses, yet it is likely that disease onset is mainly driven by the complex interplay with genetic factors and the environment. In light of the rising numbers of cardiometabolic cases and its high economic and clinical burden to society, understanding the role of environmental exposures may inform more impactful preventive strategies.

The overall aim of this thesis is to optimize exposome-wide assessment to predict disease and evaluate the causal links of environmental exposures and disease. This work focuses on data-driven approaches, paired with causal inference methods using primarily two independent Swedish cohort studies: Västerbottens Health Survey (VHU) and Malmö Diet and Cancer Study (MDCS).

The specific aims of the papers included in this thesis are the following:

Paper I

In this work, the aim was to investigate free of any hypothesis, the environmental associations with cardiometabolic traits in a population-based cohort. We undertook an environmental-wide association study (EWAS) using longitudinal data from > 31,000 adults in VHU study. Generalized linear models were used to assess the relationships of nearly 300 candidate exposures, where eleven modifiable prioritised variables were associated with most of cardiometabolic traits; most of these related to lifestyle i.e. smoking, coffee intake, physical activity, and alcohol intake. The

prioritised variables may be used for further research or to inform clinical trial designs.

Paper II

This paper was focused on evaluating the variable susceptibility to lifestyle risk factors for T2D and CVD, by applying machine learning methods. We aimed to identify individuals by estimating prediction intervals (PIs) around the association of the prioritised variables from **Paper I** with cardiometabolic traits. Moreover, we quantified the risk of being allocated inside and outside the PIs. Those individuals with 'sensitivity' (above 95th quantile) of blood glucose, lipids and blood pressure were at higher risk to developing cardiometabolic disease and premature death. In this investigation, we identified an environmentally sensitive subpopulation for risk stratification in VHU and MDCS, whether this population may or may not be captured by conventional screen strategies, these individuals can be prioritised for further evaluation or establish early preventive therapies.

Paper III

In this work, we combined structural equation modelling (SEM) and Mendelian randomization (MR) approaches to estimate the direct and mediated effect in a range of putative causal associations between macronutrient intake and cardiometabolic traits and disease. VHU was the primary dataset and freely available GWAS summary statistics were interrogated in an integrative genomic approach. In this study, we characterised the role of macronutrient intake, its association with cardiometabolic traits and disease to inform nutritional recommendations.

Paper IV

In this paper, we aggregated (in a metanalysis) secondary-data coupled with an MR approach to investigate whether prediabetes is causally linked with T2D-complications or if the association is confounded by the progression of T2D. We evaluated the risk of prediabetes associated with cardiovascular (i.e. CHD, stroke) and kidney disease outcome, yet only CHD was causally associated.

Chapter 2

Cohort studies

Västerbotten Health Survey (VHU)

The VHU (Västerbottens hälsoundersökning) is a prospective, population-based cohort study designed to improve health outcomes among the general population (~ 225,000 inhabitants) in Västerbotten county, northern Sweden[85, 86] that started in 1985. This study was motivated by the highest rates of CVD and mortality throughout Sweden at that time[87]. All adults residing in the region of Västerbotten were invited via mail to attend their primary care centre to undertake a baseline clinical examination and to complete detailed lifestyle questionnaires during the years of their 40th , 50th , and 60th birthdays. In some locations, until 1996, 30-year-olds were also included. Since the early 1990s, participation rates ranged between 58 to 66%[88]. For this thesis, I focused on VHU participants born in Sweden, as well as residing in the region to minimize confounding by population stratification; relatively few participants were excluded on the basis of non-Swedish ancestry (~ 6%)[89]. Additionally, in our analyses, we excluded people with prevalent diabetes and/or CVD to minimise the risk of respondent bias that can occur when people with diagnosed disease are asked health-related questions. Lastly, given that VHU participants had the opportunity to undertake several study visits, this enabled us to perform analyses longitudinally (**Paper I**) and cross-sectionally (**Paper III**). Moreover, in **Paper II**, we exploited the repeated measures (~ 10-year interval) data for a more stringent definition of subgroup populations and to reduce regression dilution bias.

Lifestyle and dietary assessments

All participants were requested to complete a self-administered lifestyle questionnaire during each visit. All questionnaires were optically read, and the domains included socio-economic factors, physical/mental health, quality of life, social network and support, working conditions, and alcohol/tobacco consumption. Physical activity was assessed using the modified version of the International Physical Activity Questionnaire[90, 91]. A validated semi-quantitative food frequency questionnaire (FFQ), designed to capture habitual diet over the last year, was used to retrieve information on dietary factors[92]. In 1996, the FFQ was reduced from 84

to 66 items by merging similar items and removing those considered redundant. Nutrient and energy contents were calculated based on the Swedish Food Composition Database, 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) were excluded from the analyses, as per recommended by data managers[88]. The VHU data are organized, curated, and data/samples stored under the administrative authority of Umeå University. All participants provided written informed consent before they engaged in the study.

Cardiometabolic risk markers

Clinical measurements in VHU were performed under standardized practices. To calculate BMI, calibrated tools (scale and stadiometer) provided body weight in kilograms divided by height in meters squared from participants wearing light clothing and no shoes. Systolic and diastolic blood pressures were measured in resting participants in a supine position using either manual or automated sphygmomanometers, taken by trained nurses. Peripheral blood was drawn after overnight fasting and a venous blood sample was drawn two hours after the administration of a 75-g oral glucose load. Blood glucose, total cholesterol and triglycerides levels were then measured using a Reflotron® bench-top analyser (Roche Diagnostics Scandinavia AB). HDL-C was also measured in a subgroup of participants and LDL-C was estimated using the Friedewald formula[93]. In September 2009, blood lipids and blood pressure measurements changed: the blood pressure was measured twice in a sitting position and averaged, and triglycerides and total cholesterol levels were analysed using clinical chemical analysis in the Umeå University Hospital laboratory. Thus, validated conversion equations were used to adjust the blood pressure measurements taken before and after September 2009[94]. For participants on lipid lowering and/or blood pressure lowering medications, lipid levels and/or blood pressure levels were also corrected by adding published constants (+ 0.208 mmol/L for triglycerides, + 1.347 mmol/L for total cholesterol, − 0.060 mmol/L for HDL-C, + 1.290 mmol/L for LDL-C, + 15 mmHg for SBP and + 10 mmHg for DBP)[95, 96]. Cardiometabolic traits' values considered outside the thresholds suggested by VHU data managers were removed.

Outcome assessments

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

Malmö Diet and Cancer Study (MDCS)

The MDCS is a prospective, population-based cohort study conducted between 1991 and 1996. All men and women residing in the city of Malmö (south of Sweden) born between 1923 to 1945 and 1923 to 1950, respectively, were invited to participate through a personal letter or media advertisements. Participants who did not speak or read Swedish and those with mental incapacity were not eligible to participate. Participation rate was ~ 70% (30,446 participants with ~ 40% men) at baseline[97, 98, 99]. For the analyses described we focused on the same serological traits interrogated in VHU, thus, I utilized data from the subgroup of MDCS in whom these were available; these individuals were randomly selected for deeper cardiometabolic risk marker assessment within the MDCS cardiovascular cohort (MDCS-CC) (n = 6,103) carried out between 1991 and 1994[100], in which fasting blood samples were collected to measure cardiometabolic risk markers. Just like for VHU, in our analysis using MDCS data, we excluded non-Swedish participants. The Ethical Committee at Lund University approved the MDCS (LU 51-90) and all participants provided written informed consent.

Lifestyle and dietary assessments

All participants were requested to complete a self-administered validated comprehensive lifestyle questionnaire at baseline and during follow-up. The questionnaire was designed to assess participants' medical history, medication and diet supplementation, as well as socioeconomic, demographic and lifestyle factors, such as leisure-time physical activity, smoking/tobacco habits, alcohol consumption and quality of life. Disease history was assessed through Swedish national medical registers. The 'MDCS modified diet history survey' is a validated method for dietary data collection in this population[101, 102]. The method consisted of three parts: (i) a 7-day food diary, collecting information regarding prepared meals (lunch and dinner), cold drinks and supplement intake; (ii) a FFQ, covering 168 items consumed regularly (breakfast, snacks and others not covered by the food diary) and hot drinks; moreover, portion sizes were estimated with the help of a picture booklet with 4 portion sizes as reference for up to 48 food items; and (iii) a 45 to 60 minutes interview with a trained interviewer, covering information about cooking methods and portion sizes of the items recorded in the food diary. Thus, the interviewer could check that there was no overlap in the information collected through methods (i) and (ii). The combined dietary data obtained were then introduced into a software with the Malmö Food and Nutrient Database (based on Swedish Food Database PC KOST-93) to calculate nutrient and energy intake[103, 104]. As in VHU, participants on lipid lowering and/or blood pressure lowering medications, lipid levels and/or blood pressure levels were also corrected by adding published constants (+ 0.208 mmol/L for triglycerides, + 1.347 mmol/L for total cholesterol, − 0.060 mmol/L for HDL-C, + 1.290 mmol/L for LDL-C, + 15 mmHg for SBP and + 10 mmHg for DBP)[95, 96]. For MDCS and VHU, we estimated the macronutrient percentage of energy intake

(E%) by multiplying intake by the metabolizable energy conversion factors and dividing this by TEI[105], these variables were used as exposures in **Paper III**.

Cardiometabolic risk markers

Clinical measurement protocols in MDCS followed standardized practices, where BMI, SBP, DBP, FG, HDL-C, and LDL-C were measured as described for VHU. Yet, there was no change in how blood pressure was registered and peripheral blood was collected after fasting, moreover, HbA1c was only measured in MDCS-CC using standard procedures at the Department of Clinical Chemistry, University Hospital Malmö[98, 100].

Outcome assessments

As in VHU, data pertaining to medical diagnoses and mortality were retrieved through record linkage from the National Board of Health and Welfare in Sweden, the Swedish National Tax Agency, and Statistics in Sweden until December 31st, 2014. Using the participant's civic registration numbers, records were linked, and the same diagnoses codes used in VHU study for T2D and composite CVD (MI and stroke) with the first date of a registered event was selected as the outcome for the analysis.

Chapter 3

Methods

Background

Epidemiology focuses on assessing the distribution, rates, and patterns of disease across and within populations, as well as predicting and elucidating disease aetiology. Recent developments in affordable high-throughput genetic and molecular phenotyping technologies have driven the emergence of a new type of epidemiological analyses, where large, multi-dimensional datasets can be analysed without *a priori* hypothesis. The analysis of such data sometimes exceeds the capacity of traditional statistical methods used in epidemiology. This problem has been addressed through the use of machine learning (ML) methods, which are better suited to handling very complex datasets[106] and allow both 'supervised' or 'unsupervised' analyses to be performed. While supervised ML methods are often used to reinforce or 'teach' algorithms and to test specific hypotheses, unsupervised ML methods are typically used to uncover hidden structures in the data that might be undetectable with supervised approaches[107]. Although different in many ways, both ML approaches can be highly complementary.

Environmental-wide association study (EWAS)

EWAS is an approach somewhat analogous to GWAS, in which multiple environmental factors can be systematically screened for their associations with disease traits. This approach is agnostic to prior knowledge about disease associations, thus, bias from predetermined hypothesis is minimised. EWAS was first proposed in 2010, where it was used to study associations between environmental exposures and T2D[108]. We used the same approach on a Swedish population in a longitudinal study, to investigate the relationship between modifiable lifestyle exposures and cardiometabolic disease[109]. This approach was used in **Paper I** and the prioritised variables served as input variables in the predictive modelling in **Paper II**.

Principal component analysis (PCA)

A large, multi-dimensional dataset is generally composed of observations, often found as rows (n) and features (i.e. variables of interest) in the columns (p). High dimensional data, is when $p > n$ and this pose several important hurdles. Thus, techniques to reduce the dimensions are useful to retrieve the important variables and eliminate redundant and uninformative features. One popular approach is PCA, where the features are transformed into new variables (i.e. principal components) whilst retaining the maximum variation[110, 111]. In **Paper III**, we used PCA to represent real-world dietary patterns. New variables were created as linear combinations of the original dietary variables and were used to test linear associations with cardiometabolic outcomes as a complimentary analysis.

Machine learning (ML): random forest, quantiles, and prediction intervals

Amongst ML algorithms, a popular supervised technique is random forests. Decision trees have been implemented for decision making for several decades; however, the ensemble of different trees using bootstrapped datasets and randomly selecting a subset of variables for each tree decision was first introduced in 2001[112]. After averaging all trees, the model with the highest average probability is selected. Next, variables are ranked based on the 'importance' in the model. Caveats to this technique are that it is only suitable for complete-case analysis (i.e. does not permit missing values) and the relative contribution of variables to the model (e.g. variance explained or effect sizes) can be hard to determine. Random forests can be applied when the response (dependent variable) is continuous (regression) or categorical (classification). For the former, a quantile extension was introduced in 2006 called Quantile Regression Forest (QRF)[113]; this method does not assume any prior distribution and allows different quantiles of the response to be defined, such that responses at the tails of a distribution can be quantified. This contrasts linear models, which assume the response is normally distributed and focus on group average responses. Thus, by using quantiles, we can identify different response values compared with standard random forests, which like linear regression models focuses on group means.

A good practice in prediction modelling is the random split of the sample into training, testing, and validation sets (when no external dataset is available). The training share of the sample (usually the largest) allows to 'teach' the ML algorithm when applied to the data; in the remaining set(s), the fitted model from the training step is applied to 'unseen' data in the testing set. This approach allows to quantify the performance of the built model and benchmark against other techniques and models. Appropriate interpretation of performance metrics from ML algorithms depends upon the type of response (i.e. regression or classification), the underlying assumptions, and the purpose of the model (e.g. predictive, diagnostic).

Performance metrics can be derived from building a confusion matrix, and allocating the number of observations classified as true or false, positives or negatives, respectively. Briefly, a confusion matrix is a contingency table that summarizes the performance of any 'classifier' or model (e.g. logistic regression, random forest classification), based on the correct or incorrect allocation of the predicted values against the true values (see Figure 5). The most common metrics to assess the performance is accuracy (defined as proportion of correctly predicted cases, i.e. true positives plus true negatives), sensitivity or recall (true positives divided by the sum of true positives and false negatives), specificity (true negatives divided by the sum of true negatives and false positives); and precision (true positives divided by the sum of true positives and true negatives).

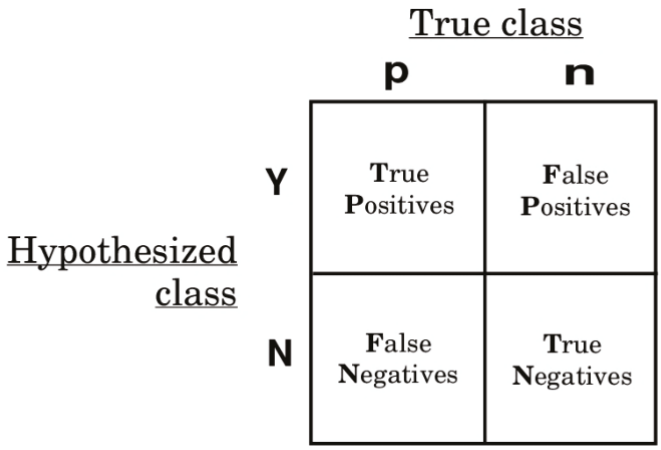


Figure 5. Confusion matrix. p: positive; n: negative; Y: yes; N: no.

Another popular metric derived from a graphical model using sensitivity and specificity is the receiver operating characteristic (ROC) curve. When a ROC is plotted, sensitivity is usually displayed on the y-axis and 1-specificity on the x-axis; the ROC area under the curve (AUC) provides an indication of overall performance of the prediction model[114, 115].

As part of prediction modelling, the forecast of a new observation is likely to have error. Uncertainty arise from the methods used to measure responses and when fitting a model to approximate to the true response. Thus, to account for uncertainty, instead of having a single estimate we can calculate PIs. Those are defined as a range of future values where new observations are likely to fall with a given probability. Similar to confidence intervals (CIs), the probability range is set to include the true observation. However, the main difference between PIs and CIs is that the latter only includes the sampling uncertainty, whether PIs (which are wider) include the

uncertainty of the population mean plus the random variation of the individual observation[116]. I used QRF and prediction intervals in **Paper II** to identify individuals and use the allocation label to estimate the hazard ratios (HR) for cardiometabolic disease. Moreover, I used the AUCs to assess the performance of the created label when applying standardized risk scores.

Mediation analysis

In graphical modelling, SEM is a data analysis technique often used to explore and test causal relations among variables in a structure[117]. When written out, SEMs provide a graphical representation of a causal hypothesis, enabling simultaneous quantification of the structural relationships among the variables, defined as indirect effects (mediated), direct effects (causal, or the exposure effect without mediators) and total effects (the product of the sum of direct and indirect effects); by doing so, one can estimate if the effect of the exposure is likely to be causal after accounting for the mediators. In SEM, a pathway of relationships between variables (i.e. exposure, mediator, and outcomes) can be estimated as generalized linear models, following any graphical configuration hypothesized (see Figure 6)[118].

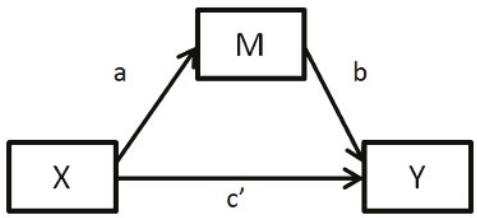


Figure 6. Graphical representation of the mediation analysis. X: independent variable; M: mediator; Y: Outcome. SEM Pathways: a is the coefficient of the effect of X on M; b is the effect of M on Y adjusting for the explanatory variable, c' is the coefficient of the effect of X on Y adjusting for M.

Mendelian randomization (MR)

Amongst the methods available with which causal relationships can be assessed, the instrumentalization of variables (IVs) and the use of proxy variables (i.e. a variable correlated with the variable of interest) were introduced to deal with confounding and control for measurement error when graphical modelling was first described[119, 120]. A plethora of GWAS have been performed to date and summary data (and sometimes individual-level data) are often available through managed-access

repositories[121]. These data can be used to identify genetic variants as IVs to investigate the causal relationships between exposures (E) and outcomes (O)[122, 123, 124] using MR approach. MR is a versatile causal inference method that can be implemented using person-level or summary statistics data; the latter called two-sample MR.

The advantage of MR is the leverage of the random and independent assortment of homologous chromosomes during meiosis, which makes it less prone to confounding or reverse causality. Two-stage least squares (2SLS) regression, is a pragmatic MR method where first the exposure is predicted from the genetic instrument, and secondly, the outcome is regressed on the predicted exposure, this is used in 'one-sample MR'. On the contrary, two-sample MR relies mostly on summary statistics, which can be rapidly interrogated in statistical packages or in web-based platforms[125], and the most common method used is the inverse-variance weighted (IVW), which combines ratios of estimates (Y_{SNPs}/X_{SNPs}) weighted by the inverse of the variance.

For MR to be valid, an IV should meet some criteria to obtain an unbiased estimation of the causal association between E and O, thus, (i) the IV should directly be associated with E, (ii) the IV should not be associated with any confounder (U) of the E−O association, (iii) the IVs associated with O should only be through the E and there should not be any other causal pathway from IV to O[122] (see Figure 7). If the latter is violated, this serves as evidence of pleiotropy.

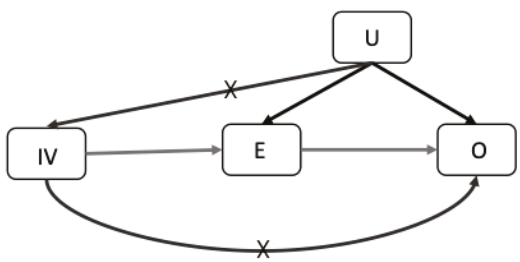


Figure 7. Graphical representation of MR. IVs: Instrumental variables. Causal association between exposure and outcome (E−O) using IV as a proxy of E. Crossed lines highlight the violation of the IV criteria, where IV should not be associated with confounders (U) of E−O association and IV should be associated with O only via E and not through another pathway.

The first criterion for IV to be valid in MR can be verified by the strength of the genetic variant with the exposure of interest. In contrast, for the second assumption

the variants must influence the outcome only through the exposure and, thirdly, the instruments must not associate with measured or unmeasured confounders or with the outcome via other biological pathways (i.e. horizontal pleiotropy). However, when assumptions are violated (when there is pleiotropy, outlier variants, and weak instrument bias) statistical solutions and robust MR methods (i.e. Mendelian randomization pleiotropy residual sum and outlier (MR-PRESSO), MR-Egger, Mendelian randomization Robust Adjusted Profile Score (MR-RAPS))[126, 127] can be implemented. The MR approach was used in **Paper III** using two-sample MR and robust MR methods to account for weak instrument bias. Moreover in **Paper IV**, we designed an instrument of SNPs associated with FG and HbA1c, but not T2D, to test the causal association of genetically-proxied prediabetes and CVD.

Colocalization

Genetic colocalization is a probabilistic technique to identify where genetic factors at particular loci are shared between two or more traits[128]. Overlapping genetic variants between traits may be driven by chance. However, colocalization aims to identify if (i) there is a causal variant for trait 'A' that is distinct from the causal variant for trait 'B' or if, (ii) the causal variant for trait 'A' and trait 'B' are shared whilst being at the same locus. Multiple algorithms for distinguishing between these two scenarios have been published[129, 130]. However, the common assumption in all these algorithms is that there is one causal variant (or a variant in very strong linkage disequilibrium (LD) with the true causal variant) in the region[128]. In **Paper III**, we utilized the Hypothesis Prioritisation for multi-trait Colocalization (HyPrColoc) algorithm[129], which is a method to identify a putative causal variant shared between two traits. Colocalization is regularly used to infer putative causal relationships between 'omics' and complex traits, in our case we found evidence of colocalization of a likely causal variant near the transcription factor 7-like 2 (*TCF7L2)*, yet having a locus colocalized to two traits is necessary but not sufficient for causality.

Metanalytic research

The 'meta' prefix denotes 'comprehensiveness'; thus, the statistical methodology to aggregate quantitative evidence from studies (i.e. effect sizes, measures of association, *p* values, etc.) is called metanalysis. To pool individual- or aggregate-level data from similar studies for further analytical procedures, involves a systematic and comprehensive approach to harmonize studies and collate published evidence with similar study designs, populations, exposures/treatments and outcomes.

One example are GWAS, which typically include the statistical aggregation of separate GWAS to achieve larger sample sizes (e.g. ~ 700,000 individuals for BMI GWAS[131]). Another example is in MR where individual SNP effect sizes are pooled

for the IV and regressed on the exposure and outcome. In **Paper IV**, we pooled published studies using retrospective and prospective cohort studies. Metanalysis, when appropriately conducted, is often regarded as yielding unbiased estimates. One example are Cochrane reviews[132], where findings often inform clinical guidelines. However, this approach is not without limitations, a major concern when aggregating studies is heterogeneity, arising from biological and/or statistical sources. To address this, solutions exist to control for differences among studies (i.e. subgroup analysis) and quantify heterogeneity (i.e. Cochran's Q and I^2). Moreover, specific biases pertaining to metanalysis of secondary data are difficult to anticipate before endeavouring in a systematic review. These include publication bias, which is a consequence of preferential reporting of positive findings. A second common error in metanalysis exists when an incorrect decision to use one of the two main statistical models is used: (i) fixed effects (if the researcher assumes the population are from the same source and there is a 'true' effect) and (ii) random-effects (where there is no single true effect, but rather a normal distribution of effects)[133]. The random-effects model often involves assigning a weight to each interrogated study according to the precision of the estimates taking into account both the within and between variance[134]. This particular approach left a profound teaching in my training given it was the last analysis conducted as part of my Masters' thesis and the first project of my doctoral studies.

Finally, the use of these methods and tools allowed me to optimise and assess causal relationships in epidemiological studies. Whether determining if an exposure-outcome association is causal, it entails the triangulation of evidence using various approaches rather than a process guided by established principles (i.e. Hill's criteria). However, there is no absolute criteria to establish causation, moreover, processes like multi-causation, dynamic competing exposures, and interactions among causal exposures, represent more complex challenges that still remain to be addressed. Traditionally, causation can only be drawn from randomised clinical trials, yet, novel and robust models to assess causal associations using observational data are now possible owing to the rapid development of techniques to analyse and integrate large amounts of data as discussed in this thesis.

Chapter 4

Results and discussion

The present section summarizes the papers included in this thesis. In **Paper I**, we implemented the EWAS approach to assess, in a relatively agnostic fashion, the relationships between modifiable lifestyle exposures and established cardiometabolic risk markers (i.e. BMI, LDL-C, HDL-C, triglycerides, total cholesterol, FG and 2-hr glucose, SBP, and DBP). This analysis enabled us to examine a wide range of exposures representing different domains (diet, working conditions, well-being). The associations were modelled using generalized linear regressions adjusted for standard covariates (e.g. age, sex, socioeconomic status, etc.). We used linear mixed models to account for longitudinal data and individual-level variation (random effects) to screen for variables that are likely to explain meaningful degrees of variance in cardiometabolic traits. In addition, we fitted linear regression models where the outcomes were the 10-year change in the level of the nine cardiometabolic traits mentioned previously[89]. In **Paper II**, using pre-selected variables (those prioritised by the EWAS in **Paper I**), I built non-parametric prediction models using an ML algorithm for each cardiometabolic trait and estimated prediction intervals to identify individuals at the tails of the distribution and quantify risk attributable to this type of classification in relation to incident cardiovascular events and T2D. In **Paper III**, two causal inference methods (i.e. mediation analysis and MR) were employed to assess the direct and indirect effects of macronutrient intake and cardiometabolic traits and disease[135]. In **Paper IV**, we conducted a large metanalysis of published cohort studies, and enhanced this analysis using MR to established the directionality and magnitude of the association of prediabetes and CVD[136]. In the following paragraphs, I discuss the four papers, highlighting key aspects during the analysis, and some of the main findings therein.

Paper I

In **Paper I**, we conducted an EWAS in > 31,000 free-living Swedish individuals within VHU. Roughly 300 lifestyle exposures grouped in 10 domains: (i) alcohol consumption; (ii) non-alcoholic beverage consumption; (iii) food; (iv) nutrients; (v) general health; (vi) physical activity and fitness; (vii) psychosocial; (viii) sleep; (ix) social conditions; and (x) tobacco use, were interrogated for their associations with

BMI, LDL-C, HDL-C, triglycerides, total cholesterol, FG, 2-hr glucose, SBP and DBP. We assessed these associations using linear mixed models, to account for repeated measures, with a random intercept for each participant (equation 1).

$$Trait = \alpha + \beta_{age} + \beta_{age^2} + \beta_{sex} + \beta_{FFQ\ version} + \beta_{other\ covariates} + (1|participant) + \varepsilon \quad (1)$$

where 1|participant represents different random intercepts for each participant and ε is error.

Moreover, we utilized standard linear regressions for the change in each cardiometabolic trait during 10 years of follow-up (equation 2).

$$\Delta Trait = \alpha + \beta_{age.B} + \beta_{age.B^2} + \beta_{age.F} + \beta_{sex} + \beta_{FFQ\ version} + \beta_{other\ covariates} + \varepsilon \quad (2)$$

where *age.B* is the age at baseline and *age.F* is the age at follow-up, and ε is error.

Initially, we transformed ordinal variables into continuous variables, and categorical variables were dichotomized. Next, we removed outliers, biologically implausible values, and missing observations. After adjusting for covariates such as age, sex, food frequency, BMI (when appropriate) and education level, the exposures 'physical activity' and 'general health' yielded at least tentative signals in analyses focused on average associations with lifestyle variables. 'Tobacco use' was amongst the top-ranking exposure when the 10-year changes were the outcome. After ranking the top 5 variables corresponding to each domain for each cardiometabolic trait, we found 11 variables with a consistent effect across the majority of the cardiometabolic traits, including 'Exercise during the last three months', 'Informed of having high blood pressure', 'Overall state of health during the last year', 'Years smoking', and 'Fitness status'. When we applied hierarchical clustering, we found the variables were grouped mostly into two groups corresponding to the smoking and physical activity domains.

As the second author of **Paper I**, my contribution was mainly through data analysis (i.e. cluster analysis) to complement the standard analyses performed by the first author. I further contributed by critically revising the manuscript and interpreting the results. Then, the first author left our research unit before the review process had concluded, thus, I addressed most of the reviewers comments and made the modifications accordingly. Throughout the data analysis process, I provided feedback and discussion to the first author on the need to correct for overfitting by partitioning our dataset and pooling effect estimates. In **Paper I**, we utilized longitudinal data from the VHU (two instances), which included ~ 300 variables. These variables were inverse normalized to correct skewedness, scaled for comparability and dietary data was residualized for total energy intake (see Figure 8).

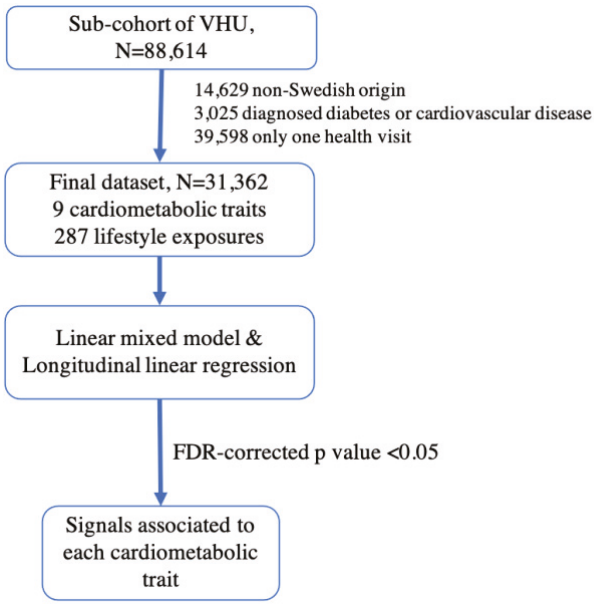


Figure 8. Flowchart of the study of **Paper I**.

There is growing recognition that lifestyle, physical activity, diet and cardiometabolic traits are intimately linked. Prior studies have identified lifestyle and environmental variables (i.e. sleep patterns, heavy metals exposures) associated with cardiometabolic traits and diseases such as T2D[108, 109], that would have been missed with hypothesis-driven approaches[137]. Moreover, most of the exposures interrogated were categorized as either 'non-modifiable' (e.g. age and sex) or 'modifiable' (e.g. diet and smoking). Doing so placed emphasis on variables (i.e. modifiable exposures) that might be components of interventions intended for cardiometabolic disease prevention as well as those that represent background risk factors (i.e. non-modifiable exposures).

Through the conducted analyses, we adjusted for multiple testing to minimise false positive signals; those that passed the false discovery rate (FDR; $p < 0.05$) were considered 'tentative' signals. For instance, these included exposures associated to the different lipid fractions investigated in the study (see Figure 9).

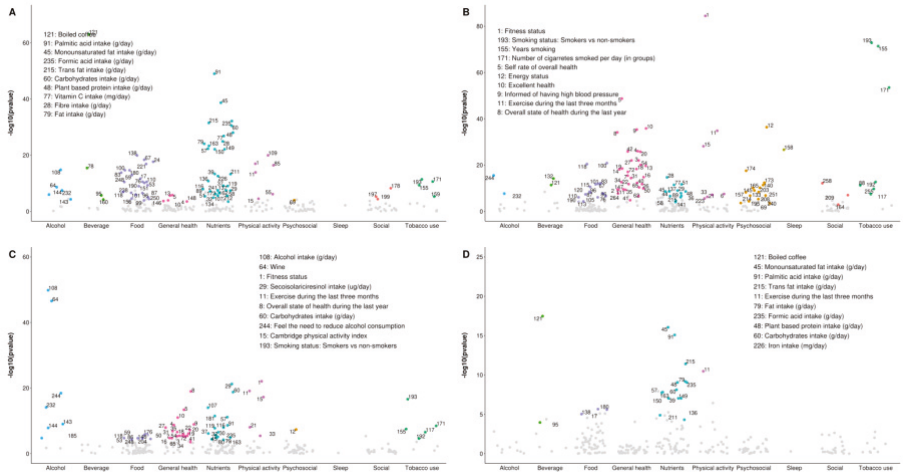


Figure 9. Manhattan plots of lipid fraction representing the distribution of *p* values in the (y-axis) of the association of lifestyle variables (x-axis) and lipid traits by domain. Panel A) total cholesterol; panel B) triglycerides; panel C) HDL-C, and panel D) LDL-C. Tentative signals are coloured and number labelled in the figure. The top 10 variables are listed in the plot and the rest detailed in the original paper[89].

Next, we rank-ordered the screened variables within each domain by percentage of the variance explained for each of the outcome traits. These variables were then clustered to identify domain-specific targets for subsequent analyses. We found that established lifestyle risk factors (i.e. 'tobacco use'), variables within the physical activity domain (e.g. 'Exercise during the last three months'), and 'alcohol intake (g/day)', explained the largest variances across all traits, whereas in the long-term lifestyle association analysis, most of the 'tentative signals' were within the 'tobacco use' and 'General health' domains. Using the prioritised variables we fitted predictive models in **Paper II**.

Paper II

During the third and fourth year of my doctoral studies, **Paper II** was completed, it was a collaborative project conceptualized by another team member, my supervisor, and I. My role was to progress the project to the point of completion. This involved comparing and selecting parametric and non-parametric models with the goal to identify individuals susceptible to environmental exposures and assess the risk of cardiometabolic disease. In these analyses, we utilized data from cohorts located in the north (VHU) and south (MDC) of Sweden. An outline of the studies is shown in Figure 10. One of the main challenges was the direct comparison between a ML

approach and generalized linear regressions. At this time, there is no clear consensus about how ML and conventional statistical models should be compared. Therefore, I use the percentage covered by the prediction intervals as a metric of performance, where those models close to the 95% coverage (equivalent to the set prediction interval probability) were selected.

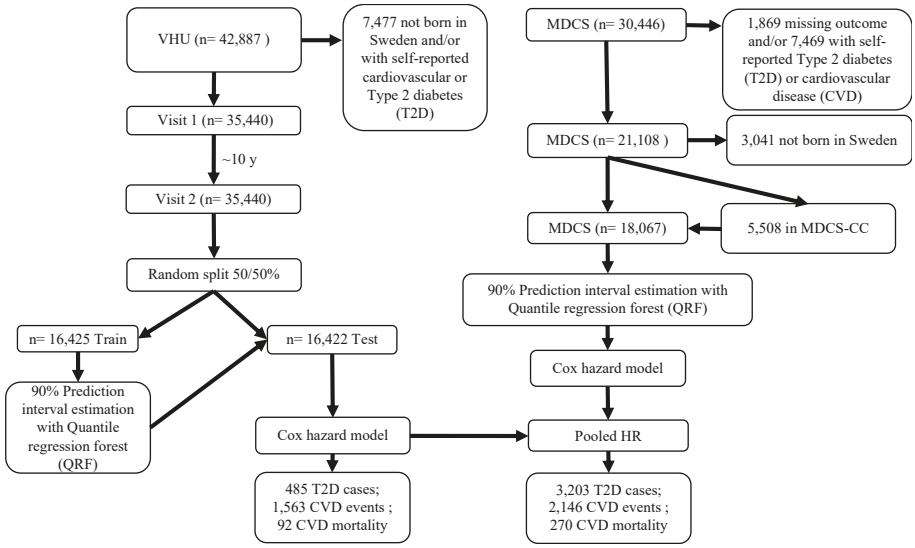


Figure 10. Flowchart of the studies in **Paper III**.

The input data for the models included the prioritised environmental exposures obtained through a comprehensive lifestyle questionnaire that queried socio-economic factors, physical/mental health, quality of life, social network and support, working conditions, alcohol/tobacco use, physical activity, and also included an FFQ. Cardiometabolic markers and cardiovascular outcomes were retrieved and measured as described in the Methods section.

To build the models, I used the prioritised exposures from **Paper I**. Initially, variables were inverse normalized and ordinal variables transformed into continuous variables. Next, I removed zero and near-zero variance predictors (i.e. mostly unique values) to avoid redundancy using the 'caret' R package. Variables were then assessed for collinearity and those with a variance inflation factor > 10 were removed. For diet data, prior to undertaking the QRF analysis, I residualized macro- and micro-nutrient intake for total energy intake to control for confounding.

Before conducting the analyses, I partitioned the data to 50% of the complete data (n=16,425) for validation and the remaining 50% (n=16,422) for model training. A common ratio to split the dataset for ML training is 70 to 80%[107]; however, because

the absolute number of disease events is relatively low in the partitioned datasets, we pragmatically determined a 50/50 data split. An example of input variables is depicted in the feature selection step from the random forest algorithm (see Figure 11).

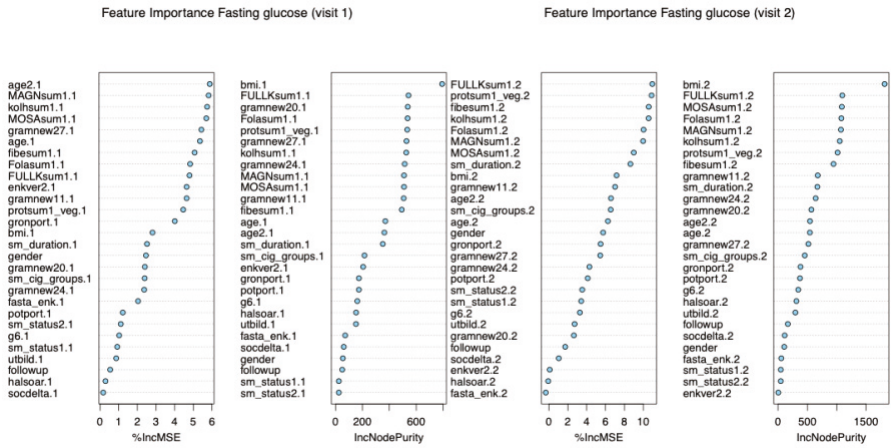


Figure 11. Variable importance plot of FG model in VHU per visit. The x-axis shows each model variable. %IncMSE: percentage in mean square error is estimated upon mean decrease of accuracy in predictions with out-of-bag samples; IncNodePurity: increase in node purity is the total decrease of squared errors for each decision tree.

The most informative features are included as input variables for the model testing step. After fitting the QRF models we determined the prediction intervals at 90% probability (5^{th} and 95^{th} quantiles) using bootstrapping to obtain a wide range of values where a future value is likely to fall, as in equation 3:

$$I(x) = [q\ 0.05\ (Y|X = x), q\ 0.95\ (Y|X = x)] \quad (3)$$

where for a given x, the response value lies within the interval $I(x)$.

Next, for observations with values above 95^{th} quantile were considered as 'sensitive', and those below 5^{th} quantile as 'resilient', the remaining were considered as the reference group. This categorization was hypothesised to describe environments with more or less likely influence on intermediate cardiometabolic traits, and thereby identify individuals with varying degrees of susceptibility to adverse health consequences of the environment. To minimize misclassification owing to regression dilution, we used repeated-measured data where an individual was classified consistently on consecutive occasions (only possible in VHU).

In subsequent time-to-event analyses (to determine risk of fatal and non-fatal disease events) the reference group included all those not classified as 'sensitive' or those below their 5th quantile. Moreover, we evaluated the predictive performance of group membership by obtaining two known cardiovascular risk-scores (i.e. Framingham risk score and 2013 American College of Cardiology/American Heart Association Task Force score) and assessed the AUCs of two logistic regression models with and without a term for sensitivity classification (0/1); overall, AUCs were higher in the models where the sensitivity term was included (see Table 2).

Table 2. AUCs of risk scores for each trait in VHU.

Traits	FRS non-laboratory-based risk score	FRS non-laboratory-based + Sensitivity status	FRS laboratory-based risk score	FRS laboratory-based risk score + Sensitivity status	ACC/AHA risk score	ACC/AHA risk score + Sensitivity status
Total cholesterol	0.72	0.73	-	-	-	-
SBP	0.73	0.74	-	-	-	-
DBP	0.73	0.73	-	-	-	-
LDL-C	0.71	0.76	0.72	0.75	0.74	0.76
HDL-C	0.69	0.62	0.67	0.63	0.70	0.68
BMI	0.73	0.74	0.71	0.73	0.73	0.74
2-hr glucose	0.73	0.74	0.70	0.71	0.71	0.71
FG	0.73	0.73	0.71	0.73	0.72	0.72
Triglycerides	0.70	0.69	-	-	-	-

" - " it was not possible to estimate the number; HDL-C: High-density lipoprotein cholesterol; LDL-C: Low-density lipoprotein cholesterol; FG: Fasting glucose; SBP: Systolic blood pressure; DBP: Diastolic blood pressure; FRS: Framingham risk score; ACC/AHA: American College of Cardiology/American Heart Association.

Although the VHU and MDCS are both prospective cohort studies of Swedish adults, the data collection methods differ in numerous ways, as noted in the Methods section. However, as MDCS represented the best available replication cohort, we proceeded with these analyses, resulting in similar results across studies and pooling them for a single measure of association to appraise risk. The findings suggest that identifying population subgroups that are especially sensitive to the adverse consequences of environmental risk factors for cardiometabolic disease may aid in the prevention of these diseases.

Paper III

I finished **Paper III** during the third year of my doctoral studies, with the analytical methods and data analysis primarily my responsibility. It is within this project where I worked most independently of all those included in this thesis. The data analyses were conducted by me, under the guidance of my supervisor. Moreover, the analysis plan, which included mediation and MR analyses, was largely designed by me.

The main goal of this analysis was to assess the causal relationships between genetically-predicted macronutrient intake and cardiometabolic conditions, which might aid to better understand the role of macronutrients in the diet and the risk of CVD and T2D. To assess a causal link, we utilized mediation analyses by accounting the mediated effects of physical activity and BMI, and to triangulate our findings we undertook a series of two-sample MR analyses. We utilized data from VHU with key input variables including macro- and micro-nutrient data, physical activity, BMI, education, portion size (as covariates) and summary statistics from large consortia studies for the genomic integrative analysis.

Before performing the MR analysis, we undertook pairwise mediation analyses of the variables in VHU dataset, as it was anticipated, adiposity and physical activity (both variables correlated) had a large proportion of the effect mediated in cardiometabolic disease (see Figure 12 and 13).

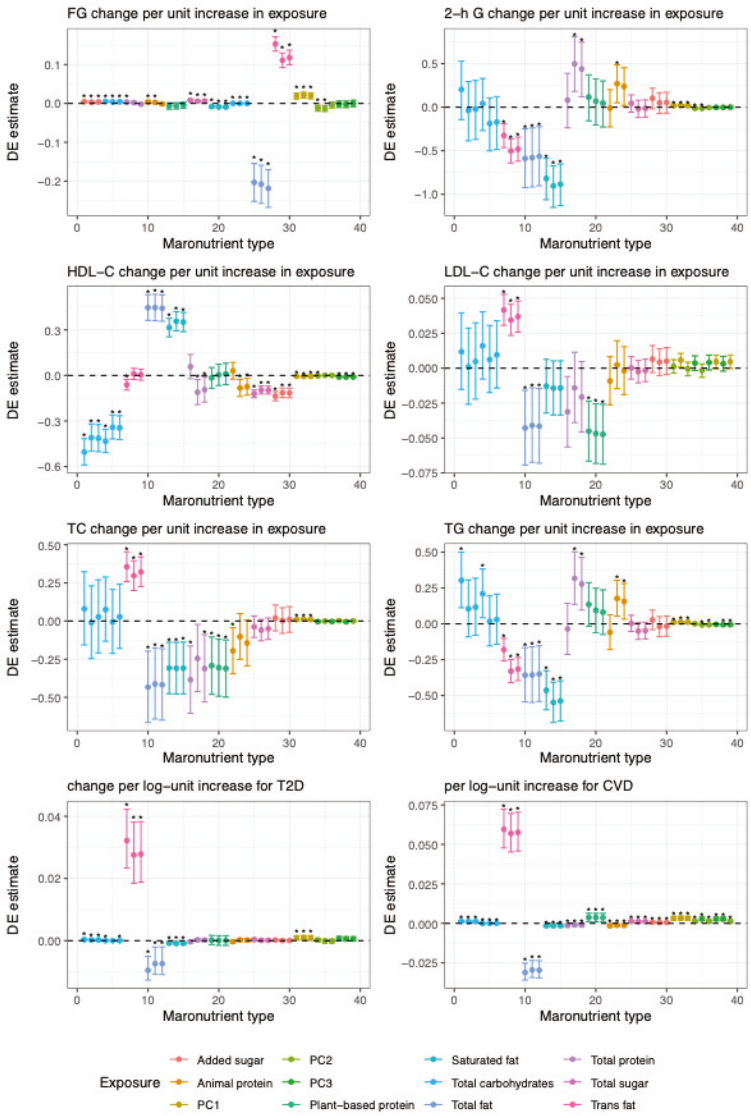


Figure 12. Direct estimates between macronutrients and outcome in pairwise mediation analysis. Macronutrients are arrayed on the x-axis in colour codes. Data are presented as direct estimates (DE) and 95% CIs (coefficient * ±1.96 (standard error(coefficient)); (*) significant after FDR correction at < 0.05; Units: fasting glucose (FG) mmol/L; 2-hr glucose (2-h G) mmol/L; total cholesterol (TC) mmol/L; LDL-C mmol/L; HDL-C mmol/L; triglycerides (TG) mmol/L; For T2D and CVD, the unit increase corresponds to the probability.

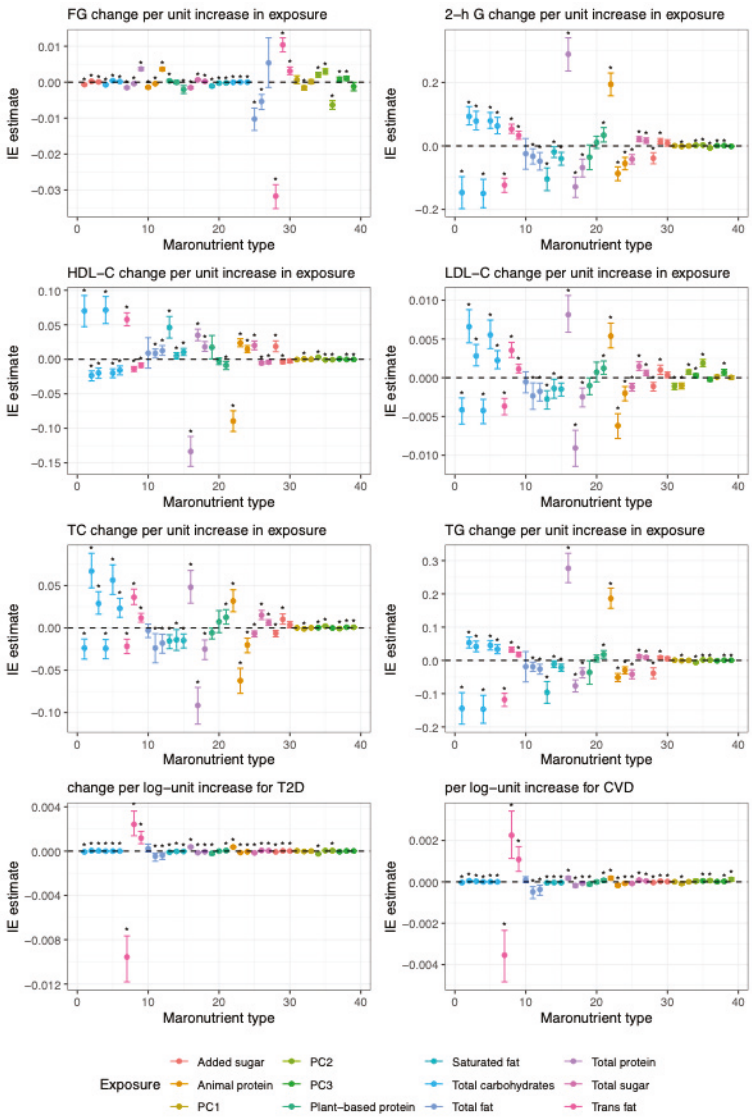


Figure 13. Indirect estimates between macronutrients and outcome in pairwise mediation analysis. Macronutrients are arrayed on the x-axis in colour codes. Data are presented as indirect estimates (IDE) and 95% CIs (coefficient * ±1.96 (standard error(coefficient)); IDE is the estimated average increase in the dependent variable as a result of the mediators; (*) significant after FDR correction at < 0.05; Units:

fasting glucose (FG) mmol/L; 2-hr glucose (2-h G) mmol/L; total cholesterol (TC) mmol/L; LDL-C mmol/L; HDL-C mmol/L; triglycerides (TG) mmol/L; For T2D and CVD, the unit increase corresponds to the probability.

Other mediation analyses were undertaken using two more configurations (i.e. serial and parallel) for building structures of more realistic scenarios. To conduct mediation analyses, I fitted several SEM pathways composed by linear or logistic regression between mediators, i.e. a pathway including BMI and physical activity and the outcomes (cardiometabolic trait or disease) adjusted for changes in macronutrient intake; another pathway including linear or logistic regression between macronutrient intake and outcomes, having adjusted for mediators (direct pathway). Given we were mainly interested in the direct effect of our exposures, we compared partially and fully-mediated models using the chi-squared difference test, an example is shown in Table 3. In addition, the statistically significant associations in the MR approach between carbohydrate intake and outcomes are shown in Table 4.

Table 3. Serial and parallel mediation analysis estimates between dietary carbohydrate intake and T2D.

		Partially mediated				Fully-mediated				Partially mediated				Fully-mediated			
	Parallel	Estimate	SE	*p*		Estimate	SE	*p*	*Difference Test *p*	*Serial*	Estimate	SE	*p*		Estimate	SE	p	*Difference Test *p*
Total carbohydrates	IE.M1	0	0	0.339		0	0	0.339		IE	0	0	0.968		0	0	0.021	
	IE.M2	0	0	< 0.001		0	0	< 0.001										
	DE	0.01	0	0.001						DE	0.01	0	0.001					
	TE	0.01	0	0.014	TIE	0	0	< 0.001		TE	0.01	0	0.004					
	chi^2	0				0			8.87E-04	chi^2	0				0.02			8.87E-04

* under ANOVA; SE: standard error; IE: indirect effect, IE.M1 indirect effect for moderator 1; IE.M2: indirect effect for moderator 2; DE: direct effect; TE: total effect.

Table 4. Two-sample MR exposure-outcome associations for dietary carbohydrate intake.

			IVW						MR-Egger					MR–PRESSO		MR–RAPS		
Exposure	Outcome	# SNPs	F	β	95% CI		*p*	Q statistic	*p*	β	95% CI		*p*	Q statistic	*p*	Global test *p* value	Distortion test *p* value	β	βSE	*p*
Carbohydrates	Glycaemic traits																			
	FG	28	5	-0.07	-0.17	0.03	0.16	44.25	0.02	-0.12	-0.59	0.35	0.61	44.16	0.01	0.02	-	-0.13	0.06	0.02
	2-hr glucose	31	5	-0.08	-0.6	0.44	0.76	36.6	0.16	-0.16	-2.83	2.5	0.91	36.58	0.13	0.16	-	-0.09	0.27	0.74
	Lipid traits																			
	HDL-C	44	4	-0.12	-0.32	0.09	0.27	1272.13	1.59E-238	-0.35	-0.98	0.28	0.28	1254.88	1.22E-235	< 1E-04	0.7563	-0.13	0.05	0.02
	LDL-C	44	4	0.44	0.05	0.82	0.03	3784.41	-	1	-0.17	2.18	0.1	3698.32	-	< 1E-04	< 1E-04	0.08	0.07	0.25

				IVW						MR-Egger						MR–PRESSO		MR–RAPS		
Exposure	**Outcome**	**# SNPs**	**F**	**β**	**95% CI**		***p***	**Q statistic**	***p***	**β**	**95% CI**		***p***	**Q statistic**	***p***	**Global test *p* value**	**Distortion test *p* value**	**β**	**βSE**	***p***
	TC	45	4	0.33	0.03	0.63	0.03	652	3.30E-109	0.76	-0.19	1.7	0.12	638.86	3.95E-107	< 1E-04	< 1E-04	0.11	0.08	0.16
	TG	44	4	0.19	0.03	0.34	0.02	663.34	4.11E-112	0.37	-0.1	0.84	0.13	653.47	1.06E-110	< 1E-04	0.1338	0.15	0.02	0
	Clinical endpoints																			
				Odds Ratio	**95% CI**		***p***			**Odds Ratio**	**95% CI**		***p***							
	T2D	6	5	0.1	0.01	0.71	0.02	80.3	-	0.19	6.56E-05	560.9	0.69	79.55	-	2.00E-04	< 1E-04	-1.68	0.63	0.01
	Stroke	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	*T2D	5	5	0.47	0.3	0.75	0.001	3.51	0.48	0.18	0.04	0.72	0.02	1.42	0.7	0.4343	-	-0.82	0.29	0.004
	CHD	44	4	1.23	0.92	1.64	0.2	95.94	6.50E-06	1.12	0.44	2.87	0.76	95.85	4.30E-06	< 1 E-04	0.2603	0.17	0.12	0.16

* adjusted for BMI; ** Wald ratio method for single SNP; (-) Not possible to estimate; We considered significant if the directions of the estimates by IVW, weighted median and MR-Egger were directionally consistent with $p < 0.05$, and no significant evidence of pleiotropy tested by MR-PRESSO ($p > 0.05$). F statistics (median) for the strength of correlation between instrument and exposure. IVW: inverse variance weighted; MR-RAPS: Robust adjusted profile score ;MR-PRESSO: Pleiotropy residual sum and outlier.T2D: Type 2 diabetes; CHD: Coronary heart disease; HDL-C: high-density lipoprotein; LDL-C: low-density lipoprotein; TG: triglycerides; TC: total cholesterol. F-statistic correspond to the median.

Next, to identify shared causal pathways among traits (i.e. genetically-predicted dietary carbohydrate intake and T2D), we employed the HyPrColoc algorithm for multi-trait colocalization[138], to explore a putative shared causal pathway. I conducted pairwise and multi-trait colocalization analyses of the shared causal variants between T2D, adjusted (T2DadjBMI) and unadjusted for BMI, carbohydrate, sugar, protein, and fat intake; where for T2DadjBMI and carbohydrate intake colocalized with posterior probability > 0.9 in the region 10 near the known T2D, *TCF7L2* locus[139, 140] (see Figure 14).

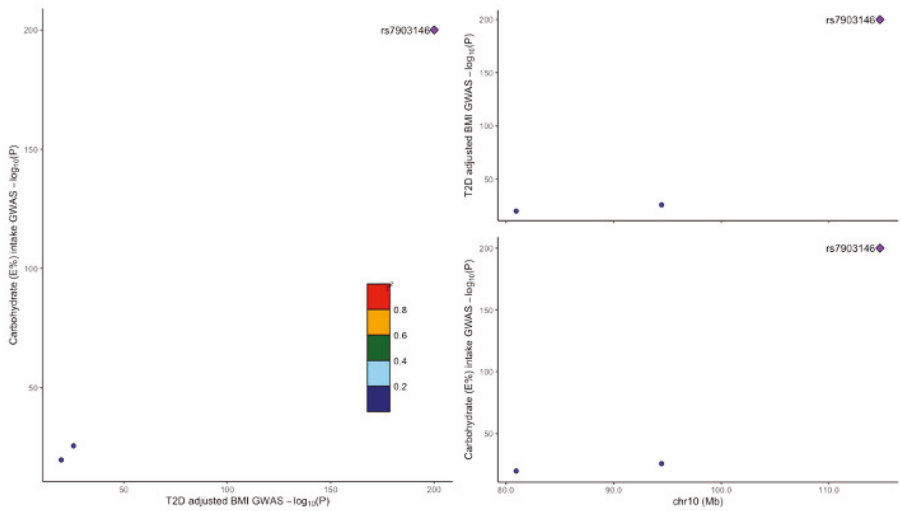


Figure 14. From left to right. Trait-trait scatterplot of 3 SNPs where *rs7903146* (diamond) is the likely causal SNP; Upper right panel: Plot of locus region (near *TCF7L2* gene) and T2D adjusted for BMI; Lower right panel: Plot of locus region (near *TCF7L2* gene) and carbohydrate (E%) trait.

In summary, our analyses suggested that genetically-predicted carbohydrate intake and T2D have a likely causal relationship, with a putative mechanism involving *TCF7L2* gene.

Paper IV

Complex diseases, like T2D and its complications, can be prevented. Although observational studies have been equivocal about the risk associated with high blood glucose levels without diabetes, clinical trials have proven effective interventions when reducing disease progression and complication in the context of diabetes[141]. Prediabetes was firstly discussed in 1979 by the National Diabetes Data Group

(NDDG) and conceptualized as a metabolic intermediate state between normal glucose homeostasis and T2D[142]. However, prediabetes, as a new clinical category, engenders costs, burden of disease and remains unclear whether it truly conveys a higher risk for CVD. In **Paper IV**, we hypothesized that prediabetes is associated with, macro- and micro-vascular complications, traditionally reserved for full-blown and long-term T2D. Using two epidemiological approaches to triangulate evidence, we increase statistical power (in the metanalysis) and tested genetically-proxied prediabetes, thus, our approach may lend credence to the notion that prediabetes should be intervened upon. To test this hypothesis, we conducted two specific analyses: (i) a systematic review and metanalysis of cohort-based studies and (ii) two-sample MR of published GWAS summary statistics.

The first stage of the analyses was led by me. We designed a search strategy to identify studies fulfilling the inclusion criteria, next we interrogated the PubMed repository by conducting a systematic literature review of published epidemiological studies (published through November 30th, 2017) focusing on 'prediabetes and diabetic complications' and extracted summary statistics that we, thereafter, combined through metanalysis. In brief, studies were included if participants were drawn from the general population, glycaemia was measured at baseline and the subsequent outcomes at follow-up were CHD, CKD or stroke, and were compared with a group of normoglycemic participants. Studies with individuals known to be diagnosed with diabetes or with diabetic values at baseline or follow-up were excluded from the analysis. Moreover, given prediabetes definition remains contested, we included studies using prediabetes identified by IGT, IFG per World Health Organization (WHO)[15] or American Diabetes Association (ADA) criteria and HbA1c per ADA criterion[16].

For MR (led by my co-author), we defined two sets of instruments that characterized excursions in fasting glucose and HbA1c without reaching the diabetic range. The genetically-proxied prediabetes exposure was built by SNPs associated with fasting glucose and HbA1c at a genome-wide level of statistical significance ($p < 5x10^{-8}$) within the MAGIC dataset[143, 144], but which are not associated with T1D or T2D ($p > 0.05$) in the most recent release of the DIAGRAM dataset[145, 146]. The IVs derived were then examined within the GWAS databases for any respective 'diabetic' complication. HbA1c (exposure) data were also obtained from the latest MAGIC dataset. CHD GWAS summary statistics were obtained from the latest metanalysis data repository[147]; stroke data was obtained from the most recent MEGASTROKE consortium metanalysis; data on kidney disease was obtained from the CKDGen GWAS summary data repository[148]. Lastly, selection of glucose-associated SNPs from MAGIC[149], resulted in 47 SNPs for FG and 10 for HbA1c.

The observational metanalysis results suggested that prediabetes is associated with increased risk of CHD and stroke (relative risk (RR) =1.16; 95%CI: 1.09, 1.23; Q= 52.5, p = 0.058; I^2=27.7%; and RR=1.11; 95%CI: 1.03, 1.18; Q=28.5, p = 0.23; I^2=16%, respectively) but not CKD (see Figure 15 and 16, respectively). In the MR

analysis, prediabetes conveyed a statistically significant increase in the odds of CHD by 26% but not stroke or CKD, without evidence of directional pleiotropy.

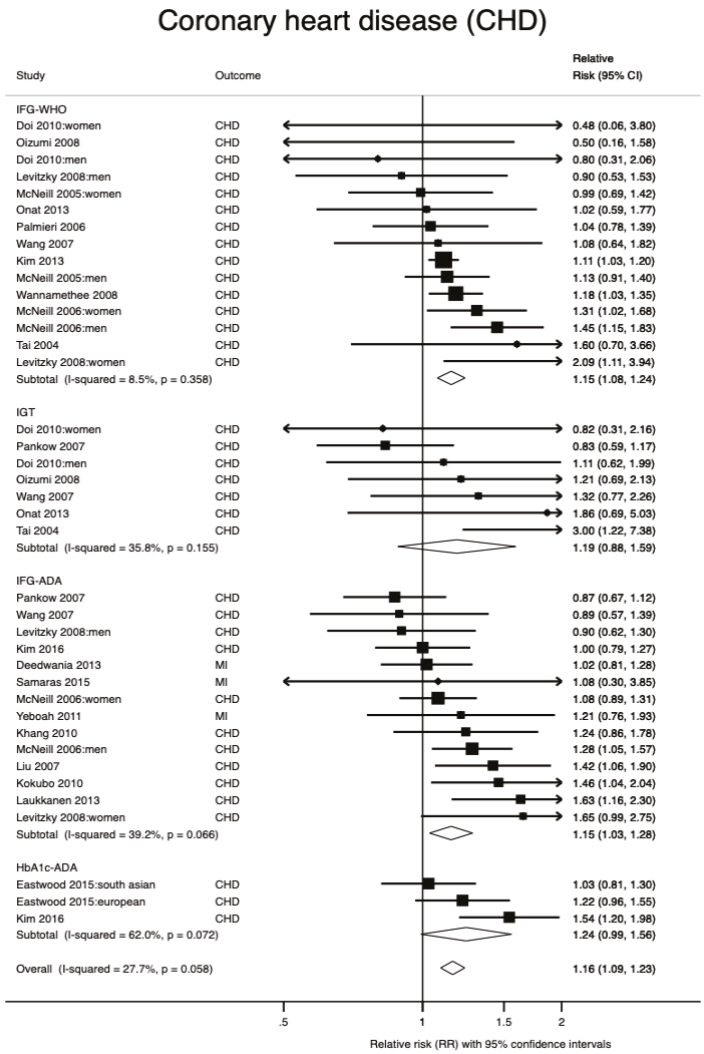


Figure 15. Metanalysis of the association between prediabetes and CHD. Data are presented as relative risks and their corresponding 95% CIs. The square and diamond shapes represent effect size whilst horizontal bars represent the 95% CIs. Source reference[136].

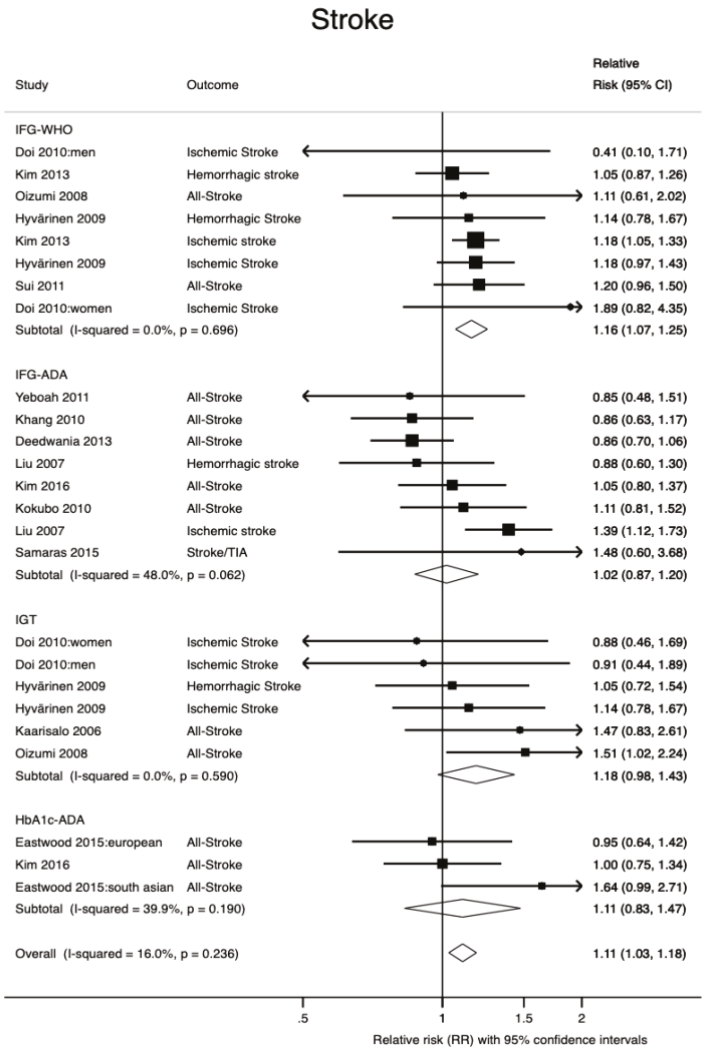


Figure 16. Metanalysis of the association between prediabetes and stroke. Data are presented as relative risks and their corresponding 95% CIs. The square and diamond shapes represent effect size whilst horizontal bars represent the 95% CIs. Source reference[136].

In summary, our findings highlight the likely causal link between prediabetes and CHD; thus, interventions for the prevention of diabetes-related CHD could be more effective if initiated early in people with prediabetes. Curiously, the last paper of this thesis was my first project of my Ph.D. studies and the consolidation of my Master's thesis. Indeed, I started working on this project during the last months of my Masters (2017) and being less experienced in research, yet, I received the most supervision of all papers during this project. I did most of the statistical analysis in a software that I learned only for this type of analysis and have not used anymore (STATA), nevertheless, this, like the other projects, have shown me the many different aspects to consider during the research process.

Summary and conclusions

The overarching goal of these four papers is to optimize the assessment of environmental exposures and its association with cardiometabolic risk, by doing so we have identified subgroups of the population that may be targeted for early interventions and we have elucidated causal associations. Often, epidemiological research investigates risk factors associated with disease without attempting to explain causal mechanisms. Here, using genetic epidemiology tools and data-driven methods, I focused on the study of putative causal links of environmental exposures with T2D and CVD. Both of these diseases, which are intertwined and increasing in prevalence, represent a large burden of disease worldwide, causing immense suffering, serious complications, and high treatment costs. Although we could not fully elucidate mechanistically the pathways linking exposures and disease, our findings might aid when elaborating prevention strategies and point to existing proven interventions for sensitive subpopulations. The findings in this thesis imply that intervening in those at-risk or before disease onset, it is likely to delay cardiometabolic disease progression and its complications.

In **Paper I**, we screened and prioritised variables that are modifiable and likely to be intervened upon, moreover, these factors can be used for further disease prediction modelling. By identifying a large number of exposures with environmental wide associations with the cardiometabolic traits, we evaluated whether the exposures conferred sensitivity to the environment and moderate the risk of T2D and CVD, using a machine learning-based approach in **Paper II**. The objective was to identify individuals at-risk yet often overlooked by conventional methods and assess their cardiometabolic risk. These individuals may be identified as populations at-risk, such as those with prediabetes or 'sensitive' to the environment, and potentially target them for more precise strategies. In **Paper III**, I focused on the environmental risk factors, specifically dietary items, and their direct impact on cardiometabolic traits and disease, where we conducted mediation analyses and an integrative genomic approach to identify the likely causal factor useful to inform dietary intervention studies. In **Paper IV**, the genetic instruments used therein demonstrated the likely causal role of prediabetes and CHD, suggesting, that intervening in those categorized as having prediabetes are likely to benefit from early interventions to prevent cardiovascular complications.

Future perspectives

Immediate extension and other considerations to the projects described in this thesis are discussed as following:

- From **Paper I**, as described in this thesis, the prioritised variables for this population can be used for other downstream analyses, for instance in **Paper II**, we used the variables for disease prediction and in **Paper III** for causal inference. Other potential venues include testing new epidemiological associations under more sophisticated techniques like deep learning to refine the identification of the most informative variables in disease prediction. Moreover, using other SEM configurations may allow us to test hypothesized associations between different environmental exposures and disease risk.
- Another consideration in **Paper I**, is that factors identified by EWAS provide a candidate shortlist of variables for further analyses, however, it also highlighted variables that were highly context-specific that may not extrapolate adequately to other populations; thus, I believe that the approaches in **Paper I** and **Paper II** must be applied in other populations before generalizing conclusions.
- One limitation in our studies, i.e. **Paper II**, is how exposure data was captured, differences amongst data collected from different studies may result in bias when aggregating and pooling results, yet more likely to pull the estimate towards the null. Moreover, it is recognized self-reported data (i.e. diet) may induce bias.
- For **Paper II** a potential extension, analogous to GWAS, is to derive an 'Environmental risk score (ERS)'. Although I did test a pilot ERS, when it was applied to an independent cohort did not show better predictive performance than conventional risk scores for cardiometabolic disease, perhaps, using ERS based upon novel ML algorithms may improve risk prediction. A potential extension is through the development and implementation of methods that can handle longitudinal data and incorporate time-varying exposures or through trajectory-based modelling.
- Other potential venues for **Paper I** and **Paper II**, is the enrichment of the dataset with genetics and multi-omics, initially, we discussed the plan to

perform genotyping *de novo* on VHU cohort, however, due to time and budget constraints we did not proceed. GWAS studies, with large sample sizes (~ millions of individuals) together with more recent exome and whole-genome sequencing studies have successfully identified novel genetic loci associated with cardiometabolic traits and diseases. Therefore, more research questions can be addressed using summary-level data in post-GWAS analyses. Similarly, despite having adequate power from GWAS data, the regular update of our analyses may aid to estimate with more precision the size of the effects here reported.

- Environmental exposures do not explain all variance in cardiometabolic risk; much of the remainder is likely attributable to genetic susceptibility, and their interaction with lifestyle. To identify individuals with specific susceptibility, I undertook a novel approach in **Paper II**, where sensitivity to the environment emphasizes and suggests a non-environmental component of susceptibility, thus, this may be tested with appropriate statistical analysis if genetic data is obtained. Moreover, the modelling of genetic interactions (if genotype data is available) with 'sensitivity' status, might help us to better understand the gene-environment interaction in the different population subgroups.

- Many analytical methods and data-driven approaches have successfully been used to identify heterogeneous subgroups of disease. Most of these approaches have identified these groups or clusters based upon phenotypic markers. In **Paper II** we identified a group of individuals based on their susceptibility to the environment in two different cohorts, the identified group had a distinct risk profile, however, the validation of our findings should be obtained in a diverse population. Thus, it is expected these subgroups have divergent multiomics signatures that may allow us to inspect putative mechanisms in greater detail in future research.

- Similar to genotyping, another venue to enrich data in our studies (i.e. VHU in **Paper II**) is through data mining of medical records. So far, a large number of epidemiological studies have primarily relied on self-reported data to quantify environmental exposures. However, lifestyles are highly complex to capture and seldomly remain the same through lifetime, thus, there may be opportunities to obtain objective biomarkers and clinical variables through linking health and medical records, conditioned upon the informed consent is given for these analyses.

- For **Paper III**, one of the limitations we had in the mediation analysis is the missing assessment of the role of host microbiome in carbohydrate metabolism and cardiometabolic disease. Whether this data is difficult to obtain, integrate and analyse, now is widely acknowledged the important role of microbiota in glucose metabolism and T2D risk.

- For **Paper IV**, a potential opportunity is to expand the causal investigation to other T2D microvascular complications such as retinopathy and neuropathy, which have also been associated with prediabetes, yet secondary data available for these conditions may be scarce.

Paper II was the final project of my Ph.D. studies, which I undertook during my research visit to the Christensen Lab at Geisel School of Medicine, Dartmouth College in New Hampshire, USA, where they focus on translational cancer research. A potential follow-up analysis would be to explore the same approach in an independent epidemiological cohort in the context of cancer risk (e.g. New Hampshire Birth cohort). In addition, for **Paper III** a potential research question for causal inference is to explore the role of established risk factors underlying metabolic and cancer processes, such as oxidized cholesterol derivatives (i.e. oxysterols) and insulin-mediated effects using genomic integrative approaches.

The findings in this thesis are informative for public health and nutritional guidelines that require causal evidence. Also the results presented may motivate new research questions. The methods applied in each paper represent a few of many possible approaches. Consequently, confirmatory analyses using alternative methods may yield new insights. Nevertheless, the work I have performed helps to assess the magnitude, directionality, and nature of associations, thereby improving our understanding of the relationships between environmental exposures and cardiometabolic traits and disease. This work also helps to ensure that when intervening upon targeted exposures or individuals, we anticipate it is likely to have a relevant clinical effect.

Finally, this thesis helps to inform the design of clinical trials by highlighting causal modifiable exposures to intervene or control upon. Within the context of precision medicine, to establish appropriate therapeutics (either pharmacotherapy or lifestyle modification) we need to target, and prioritise, subpopulations including individuals living with prediabetes or those sensitive to the environment to prevent cardiometabolic disease. In addition, interventions to delay T2D onset should consider the impact of dietary carbohydrate intake when formulating nutritional recommendations. As research advances, new tools and techniques will emerge in the future with the possibility to disentangle the underlying biological mechanisms and individualise care. The integration of new technologies and fields will pave the way to optimise the exposome-wide assessment to move forward from standard care to personalised medicine.

References

1. Turner MC*, et al.* Assessing the Exposome with External Measures: Commentary on the State of the Science and Research Recommendations. *Annu Rev Public Health* **38**, 215-239 (2017).
2. Roth GA*, et al.* Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet* **392**, 1736-1788 (2018).
3. Roth GA*, et al.* Trends and Patterns of Geographic Variation in Cardiovascular Mortality Among US Counties, 1980-2014. *JAMA* **317**, 1976-1992 (2017).
4. Chung WK*, et al.* 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 (2020).
5. Udler MS. Identifying subgroups of people at risk for type 2 diabetes. *Nature Medicine* **27**, 23-25 (2021).
6. Ahlqvist E*, et al.* Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The lancet Diabetes & endocrinology* **6**, 361-369 (2018).
7. Udler MS*, et al.* Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLoS Med* **15**, e1002654 (2018).
8. Wesolowska-Andersen A*, et al.* Four groups of type 2 diabetes contribute to the etiological and clinical heterogeneity in newly diagnosed individuals: An IMI DIRECT study. *Cell Rep Med* **3**, 100477 (2022).
9. Sharma A*, et al.* Cluster Analysis of Cardiovascular Phenotypes in Patients With Type 2 Diabetes and Established Atherosclerotic Cardiovascular Disease: A Potential Approach to Precision Medicine. *Diabetes care* **45**, 204-212 (2022).
10. Group DPPR. Long-term effects of lifestyle intervention or metformin on diabetes development and microvascular complications over 15-year follow-up: the Diabetes Prevention Program Outcomes Study. *The lancet Diabetes & endocrinology* **3**, 866-875 (2015).
11. Yu E*, et al.* Diet, Lifestyle, Biomarkers, Genetic Factors, and Risk of Cardiovascular Disease in the Nurses' Health Studies. *Am J Public Health* **106**, 1616-1623 (2016).
12. Franks PW, McCarthy MI. Exposing the exposures responsible for type 2 diabetes and obesity. *Science* **354**, 69-73 (2016).

13. Kahn R, Buse J, Ferrannini E, Stern M, American Diabetes A, European Association for the Study of D. The metabolic syndrome: time for a critical appraisal: joint statement from the American Diabetes Association and the European Association for the Study of Diabetes. *Diabetes care* **28**, 2289-2304 (2005).
14. Eckel RH, Kahn R, Robertson RM, Rizza RA. Preventing cardiovascular disease and diabetes: a call to action from the American Diabetes Association and the American Heart Association. *Circulation* **113**, 2943-2946 (2006).
15. Sasson C*, et al.* American Heart Association Diabetes and Cardiometabolic Health Summit: Summary and Recommendations. *J Am Heart Assoc* **7**, e009271 (2018).
16. Brunzell JD*, et al.* Lipoprotein management in patients with cardiometabolic risk: consensus conference report from the American Diabetes Association and the American College of Cardiology Foundation. *J Am Coll Cardiol* **51**, 1512-1524 (2008).
17. Cardiometabolic Risk Working Group: Executive C*, et al.* Cardiometabolic risk in Canada: a detailed analysis and position paper by the cardiometabolic risk working group. *Can J Cardiol* **27**, e1-e33 (2011).
18. International Diabetes Federation. IDF Diabetes Atlas, 9th edn. (2019).
19. World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF consultation. (2006).
20. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes care* **37 Suppl 1**, S81-90 (2014).
21. Tabak AG, Herder C, Rathmann W, Brunner EJ, Kivimaki M. Prediabetes: a high-risk state for diabetes development. *Lancet* **379**, 2279-2290 (2012).
22. Sattar N. Biomarkers for diabetes prediction, pathogenesis or pharmacotherapy guidance? Past, present and future possibilities. *Diabet Med* **29**, 5-13 (2012).
23. Acharjee S, Ghosh B, Al-Dhubiab BE, Nair AB. Understanding type 1 diabetes: etiology and models. *Can J Diabetes* **37**, 269-276 (2013).
24. Dudzik D*, et al.* Metabolic fingerprint of Gestational Diabetes Mellitus. *J Proteomics* **103**, 57-71 (2014).
25. Khan SR*, et al.* The discovery of novel predictive biomarkers and early-stage pathophysiology for the transition from gestational diabetes to type 2 diabetes. *Diabetologia* **62**, 687-703 (2019).
26. Gudmundsdottir V*, et al.* Whole blood co-expression modules associate with metabolic traits and type 2 diabetes: an IMI-DIRECT study. *Genome Med* **12**, 109 (2020).
27. Slieker RC*, et al.* Distinct Molecular Signatures of Clinical Clusters in People With Type 2 Diabetes: An IMI-RHAPSODY Study. *Diabetes* **70**, 2683-2693 (2021).
28. American Diabetes A. Diagnosis and classification of diabetes mellitus. *Diabetes care* **37 Suppl 1**, S81-90 (2014).
29. Chia CW, Egan JM, Ferrucci L. Age-Related Changes in Glucose Metabolism, Hyperglycemia, and Cardiovascular Risk. *Circ Res* **123**, 886-904 (2018).
30. Nakrani MN, Wineland RH, Anjum F. Physiology, Glucose Metabolism. *StatPearls*, (2021).

31. Roder PV, Wu B, Liu Y, Han W. Pancreatic regulation of glucose homeostasis. *Exp Mol Med* **48**, e219 (2016).
32. Olefsky JM. G protein-coupled receptors as targets for anti-diabetic therapeutics. *Nature reviews Drug discovery* **15**, 161-172 (2016).
33. da Silva AA, do Carmo JM, Li X, Wang Z, Mouton AJ, Hall JE. Role of Hyperinsulinemia and Insulin Resistance in Hypertension: Metabolic Syndrome Revisited. *Can J Cardiol* **36**, 671-682 (2020).
34. Jumpertz R, Thearle MS, Bunt JC, Krakoff J. Assessment of non-insulin-mediated glucose uptake: association with body fat and glycemic status. *Metabolism* **59**, 1396-1401 (2010).
35. Kim JK. Hyperinsulinemic-euglycemic clamp to assess insulin sensitivity in vivo. *Methods Mol Biol* **560**, 221-238 (2009).
36. Di Pino A, Urbano F, Piro S, Purrello F, Rabuazzo AM. Update on pre-diabetes: Focus on diagnostic criteria and cardiovascular risk. *World journal of diabetes* **7**, 423 (2016).
37. Nurmohamed NS, Navar AM, Kastelein JJP. New and Emerging Therapies for Reduction of LDL-Cholesterol and Apolipoprotein B: JACC Focus Seminar 1/4. *J Am Coll Cardiol* **77**, 1564-1575 (2021).
38. Ferrannini E, Natali A, Capaldo B, Lehtovirta M, Jacob S, Yki-Jarvinen H. Insulin resistance, hyperinsulinemia, and blood pressure: role of age and obesity. European Group for the Study of Insulin Resistance (EGIR). *Hypertension* **30**, 1144-1149 (1997).
39. Libby P*, et al.* Atherosclerosis. *Nat Rev Dis Primers* **5**, 57 (2019).
40. Benjamin EJ*, et al.* Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. *Circulation* **139**, e56-e528 (2019).
41. Collaborators GMaCoD. Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet* **385**, 117-171 (2015).
42. Piepoli MF*, et al.* Guidelines: Editor's choice: 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts) Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). *European heart journal* **37**, 2315 (2016).
43. Liu K*, et al.* Healthy lifestyle through young adulthood and the presence of low cardiovascular disease risk profile in middle age: the Coronary Artery Risk Development in (Young) Adults (CARDIA) study. *Circulation* **125**, 996-1004 (2012).
44. Stewart J, Manmathan G, Wilkinson P. Primary prevention of cardiovascular disease: A review of contemporary guidance and literature. *JRSM Cardiovasc Dis* **6**, 2048004016687211 (2017).

45. Patel SA, Winkel M, Ali MK, Narayan KM, Mehta NK. Cardiovascular mortality associated with 5 leading risk factors: national and state preventable fractions estimated from survey data. *Ann Intern Med* **163**, 245-253 (2015).
46. Urtamo A, Jyvakorpi SK, Kautiainen H, Pitkala KH, Strandberg TE. Major cardiovascular disease (CVD) risk factors in midlife and extreme longevity. *Aging Clin Exp Res* **32**, 299-304 (2020).
47. Bjornstad P, Donaghue KC, Maahs DM. Macrovascular disease and risk factors in youth with type 1 diabetes: time to be more attentive to treatment? *The lancet diabetes & endocrinology* **6**, 809-820 (2018).
48. Investigators WMPP. The World Health Organization MONICA Project (monitoring trends and determinants in cardiovascular disease): a major international collaboration. *Journal of clinical epidemiology* **41**, 105-114 (1988).
49. Yusuf S*, et al.* Investigators IS. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet* **364**, 937-952 (2004).
50. Wong ND. Epidemiological studies of CHD and the evolution of preventive cardiology. *Nat Rev Cardiol* **11**, 276-289 (2014).
51. Perreault L, Færch K. Approaching pre-diabetes. *Journal of Diabetes and its Complications* **28**, 226-233 (2014).
52. Buysschaert M, Medina JL, Bergman M, Shah A, Lonier J. Prediabetes and associated disorders. *Endocrine* **48**, 371-393 (2015).
53. Kubota Y, Evenson KR, Maclehose RF, Roetker NS, Joshu CE, Folsom AR. Physical Activity and Lifetime Risk of Cardiovascular Disease and Cancer. *Med Sci Sports Exerc* **49**, 1599-1605 (2017).
54. Vaccaro O*, et al.* Impact of diabetes and previous myocardial infarction on long-term survival: 25-year mortality follow-up of primary screenees of the Multiple Risk Factor Intervention Trial. *Arch Intern Med* **164**, 1438-1443 (2004).
55. Falk E. Pathogenesis of atherosclerosis. *J Am Coll Cardiol* **47**, C7-12 (2006).
56. Badimon L, Vilahur G. Thrombosis formation on atherosclerotic lesions and plaque rupture. *J Intern Med* **276**, 618-632 (2014).
57. Vergallo R, Crea F. Atherosclerotic plaque healing. *New England Journal of Medicine* **383**, 846-857 (2020).
58. Wernly B, Pernow J, Kelm M, Jung C. The role of arginase in the microcirculation in cardiovascular disease. *Clin Hemorheol Microcirc* **74**, 79-92 (2020).
59. Sorensen BM*, et al.* Prediabetes and Type 2 Diabetes Are Associated With Generalized Microvascular Dysfunction: The Maastricht Study. *Circulation* **134**, 1339-1352 (2016).
60. Maiese K, Chong ZZ, Shang YC, Hou J. Novel avenues of drug discovery and biomarkers for diabetes mellitus. *The Journal of Clinical Pharmacology* **51**, 128-152 (2011).
61. Backholer K, Chen L, Shaw J. Screening for diabetes. *Pathology* **44**, 110-114 (2012).
62. Visscher PM*, et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *American journal of human genetics* **101**, 5-22 (2017).

63. Mahajan A, *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet* **50**, 1505-1513 (2018).
64. Manning AK, *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature Genetics* **44**, 659-U681 (2012).
65. Saxena R, *et al.* Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet* **42**, 142-148 (2010).
66. Wheeler E, *et al.* Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS Med* **14**, e1002383 (2017).
67. Nikpay M, *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* **47**, 1121-1130 (2015).
68. Malik R, *et al.* Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet* **50**, 524-537 (2018).
69. Fuchsberger C, *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41-47 (2016).
70. Scott LJ, *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341-1345 (2007).
71. Larson MG, *et al.* Framingham Heart Study 100K project: genome-wide associations for cardiovascular disease outcomes. *BMC Med Genet* **8 Suppl 1**, S5 (2007).
72. Samani NJ, *et al.* Genomewide association analysis of coronary artery disease. *N Engl J Med* **357**, 443-453 (2007).
73. Dichgans M, Pulit SL, Rosand J. Stroke genetics: discovery, biology, and clinical applications. *The Lancet Neurology* **18**, 587-599 (2019).
74. Consortium CAD, *et al.* Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* **45**, 25-33 (2013).
75. Malik R, *et al.* Genome-wide meta-analysis identifies 3 novel loci associated with stroke. *Ann Neurol* **84**, 934-939 (2018).
76. Erdmann J, Kessler T, Munoz Venegas L, Schunkert H. A decade of genome-wide association studies for coronary artery disease: the challenges ahead. *Cardiovasc Res* **114**, 1241-1257 (2018).
77. Meigs JB. The Genetic Epidemiology of Type 2 Diabetes: Opportunities for Health Translation. *Curr Diab Rep* **19**, 62 (2019).
78. Udler MS, McCarthy MI, Florez JC, Mahajan A. Genetic Risk Scores for Diabetes Diagnosis and Precision Medicine. *Endocr Rev* **40**, 1500-1520 (2019).
79. Prasad RB, Groop L. Genetics of type 2 diabetes-pitfalls and possibilities. *Genes (Basel)* **6**, 87-123 (2015).
80. Bevan S, *et al.* Genetic heritability of ischemic stroke and the contribution of previously reported candidate gene and genomewide associations. *Stroke* **43**, 3161-3167 (2012).

81. McPherson R, Tybjaerg-Hansen A. Genetics of Coronary Artery Disease. *Circ Res* **118**, 564-578 (2016).
82. Manolio TA*, et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009).
83. Franks PW, Atabaki-Pasdar N. Causal inference in obesity research. *J Intern Med* **281**, 222-232 (2017).
84. Franks PW, Poveda A. Lifestyle and precision diabetes medicine: will genomics help optimise the prediction, prevention and treatment of type 2 diabetes through lifestyle therapy? *Diabetologia* **60**, 784-792 (2017).
85. Hallmans G*, et al.* 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 (2003).
86. Norberg M, Wall S, Boman K, Weinehall L. The Vasterbotten Intervention Programme: background, design and implications. *Glob Health Action* **3**, (2010).
87. Ahmad S*, et al.* Gene x physical activity interactions in obesity: combined analysis of 111,421 individuals of European ancestry. *PLoS Genet* **9**, e1003607 (2013).
88. Winkvist A, Hornell A, Hallmans G, Lindahl B, Weinehall L, Johansson I. More distinct food intake patterns among women than men in northern Sweden: a population-based survey. *Nutr J* **8**, 12 (2009).
89. Poveda A*, et al.* Exposome-wide ranking of modifiable risk factors for cardiometabolic disease traits. *Sci Rep* **12**, 4088 (2022).
90. Craig CL*, et al.* International physical activity questionnaire: 12-country reliability and validity. *Med Sci Sports Exerc* **35**, 1381-1395 (2003).
91. Hallal PC, Victora CG. Reliability and validity of the International Physical Activity Questionnaire (IPAQ). *Med Sci Sports Exerc* **36**, 556 (2004).
92. Johansson I, Hallmans G, Wikman A, Biessy C, Riboli E, Kaaks R. Validation and calibration of food-frequency questionnaire measurements in the Northern Sweden Health and Disease cohort. *Public Health Nutr* **5**, 487-496 (2002).
93. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem* **18**, 499-502 (1972).
94. Ng N, Carlberg B, Weinehall L, Norberg M. Trends of blood pressure levels and management in Vasterbotten County, Sweden, during 1990-2010. *Global Health Action* **5**, 1-12 (2012).
95. Wu J*, et al.* An investigation of the effects of lipid-lowering medications: genome-wide linkage analysis of lipids in the HyperGEN study. *BMC Genet* **8**, 60 (2007).
96. Tobin MD, Sheehan NA, Scurrah KJ, Burton PR. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Statistics in medicine* **24**, 2911-2935 (2005).
97. Berglund G, Elmstahl S, Janzon L, Larsson SA. The Malmo Diet and Cancer Study. Design and feasibility. *J Intern Med* **233**, 45-51 (1993).

98. Manjer J*, et al.* The Malmo Diet and Cancer Study: representativity, cancer incidence and mortality in participants and non-participants. *Eur J Cancer Prev* **10**, 489-499 (2001).
99. Manjer J, Elmståhl S, Janzon L, Berglund G. Invitation to a population-based cohort study: differences between subjects recruited using various strategies. *Scandinavian Journal of Public Health* **30**, 103-112 (2002).
100. Hedblad B, Nilsson P, Janzon L, Berglund G. Relation between insulin resistance and carotid intima-media thickness and stenosis in non-diabetic subjects. Results from a cross-sectional study in Malmö, Sweden. *Diabetic Medicine* **17**, 299-307 (2000).
101. Elmstahl S, Riboli E, Lindgarde F, Gullberg B, Saracci R. The Malmo Food Study: the relative validity of a modified diet history method and an extensive food frequency questionnaire for measuring food intake. *Eur J Clin Nutr* **50**, 143-151 (1996).
102. Riboli E, Elmstahl S, Saracci R, Gullberg B, Lindgarde F. The Malmo Food Study: validity of two dietary assessment methods for measuring nutrient intake. *Int J Epidemiol* **26 Suppl 1**, S161-173 (1997).
103. Callmer E, Riboli E, Saracci R, Akesson B, Lindgarde F. Dietary assessment methods evaluated in the Malmo food study. *J Intern Med* **233**, 53-57 (1993).
104. Wirfalt E, Mattisson I, Gullberg B, Berglund G. Food patterns defined by cluster analysis and their utility as dietary exposure variables: a report from the Malmo Diet and Cancer Study. *Public Health Nutr* **3**, 159-173 (2000).
105. Maclean W*, et al.* Food energy–Methods of analysis and conversion factors. In: *Food and agriculture organization of the united nations technical workshop report*) (2003).
106. Cleophas TJ, Zwinderman AH, Cleophas-Allers HI. *Machine learning in medicine*. Springer (2013).
107. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media (2009).
108. Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS One* **5**, e10746 (2010).
109. Lind PM, Riserus U, Salihovic S, Bavel B, Lind L. An environmental wide association study (EWAS) approach to the metabolic syndrome. *Environ Int* **55**, 1-8 (2013).
110. Jolliffe IT. *Principal component analysis for special types of data*. Springer (2002).
111. Ringnér M. What is principal component analysis? *Nature biotechnology* **26**, 303-304 (2008).
112. Breiman L. Random forests. *Machine Learning* **45**, 5-32 (2001).
113. Meinshausen N. Quantile regression forests. *Journal of Machine Learning Research* **7**, 983-999 (2006).
114. Sidey-Gibbons JA, Sidey-Gibbons CJ. *Machine learning in medicine: a practical introduction* (2019).

115. Fan J, Upadhye S, Worster A. Understanding receiver operating characteristic (ROC) curves. *CJEM* **8**, 19-20 (2006).
116. Lawless J, Fredette M. Frequentist prediction intervals and predictive distributions. *Biometrika* **92**, 529-542 (2005).
117. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychol Methods* **15**, 309-334 (2010).
118. Hayes AF. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications (2017).
119. Judea P. An introduction to causal inference. *The International Journal of Biostatistics* **6**, 1-62 (2010).
120. Pearl J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann (1988).
121. MacArthur J*, et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**, D896-D901 (2017).
122. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology* **32**, 1-22 (2003).
123. Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol* **33**, 30-42 (2004).
124. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine* **27**, 1133-1163 (2008).
125. Hemani G*, et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, (2018).
126. Verbanck M, Chen C-y, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature genetics* **50**, 693-698 (2018).
127. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512-525 (2015).
128. Hukku A, Pividori M, Luca F, Pique-Regi R, Im HK, Wen X. Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations. *American journal of human genetics* **108**, 25-35 (2021).
129. Foley CN*, et al.* A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nature communications* **12**, 1-18 (2021).
130. Giambartolomei C*, et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
131. Yengo L*, et al.* Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641-3649 (2018).
132. Higgins J, Green S. Cochrane handbook for systematic reviews of interventions.). The Cochrane Collaboration (2011).

133. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. *Introduction to meta-analysis*. John Wiley & Sons (2011).
134. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. *Contemp Clin Trials* **28**, 105-114 (2007).
135. Pomares-Millan H, Atabaki-Pasdar N, Coral D, Johansson I, Giordano GN, Franks PW. Estimating the Direct Effect between Dietary Macronutrients and Cardiometabolic Disease, Accounting for Mediation by Adiposity and Physical Activity. *Nutrients* **14**, 1218 (2022).
136. Mutie PM*, et al.* An investigation of causal relationships between prediabetes and vascular complications. *Nature communications* **11**, 4592 (2020).
137. Patel CJ, Ioannidis JP. Studying the elusive environment in large scale. *JAMA* **311**, 2173-2174 (2014).
138. Foley CN*, et al.* A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nature communications* **12**, 764 (2021).
139. Grant SF*, et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet* **38**, 320-323 (2006).
140. Zeggini E*, et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* **40**, 638-645 (2008).
141. Diabetes Prevention Program Research G. Long-term effects of lifestyle intervention or metformin on diabetes development and microvascular complications over 15-year follow-up: the Diabetes Prevention Program Outcomes Study. *Lancet Diabetes Endocrinol* **3**, 866-875 (2015).
142. Buysschaert M, Bergman M. Definition of prediabetes. *Med Clin North Am* **95**, 289-297, vii (2011).
143. Scott RA*, et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature genetics* **44**, 991-1005 (2012).
144. Wheeler E*, et al.* Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS Med* **14**, e1002383 (2017).
145. Mahajan A*, et al.* Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nature genetics* **50**, 559-571 (2018).
146. Morris AP*, et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* **44**, 981-990 (2012).
147. van der Harst P, Verweij N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ Res* **122**, 433-443 (2018).
148. Wuttke M*, et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat Genet* **51**, 957-972 (2019).
149. Morris AP*, et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics* **44**, 981-990 (2012).

Paper I

scientific reports

Check for updates

OPEN

Exposome-wide ranking of modifiable risk factors for cardiometabolic disease traits

Alaitz Poveda[1], Hugo Pomares-Millan[1,9], Yan Chen[1,2,3,9], Azra Kurbasic[1], Chirag J. Patel[4], Frida Renström[1,5,6], Göran Hallmans[5], Ingegerd Johansson[7] & Paul W. Franks[1,5,8]✉

The present study assessed the temporal associations of ~ 300 lifestyle exposures with nine cardiometabolic traits to identify exposures/exposure groups that might inform lifestyle interventions for the reduction of cardiometabolic disease risk. The analyses were undertaken in a longitudinal sample comprising > 31,000 adults living in northern Sweden. Linear mixed models were used to assess the average associations of lifestyle exposures and linear regression models were used to test associations with 10-year change in the cardiometabolic traits. 'Physical activity' and 'General Health' were the exposure categories containing the highest number of 'tentative signals' in analyses assessing the average association of lifestyle variables, while 'Tobacco use' was the top category for the 10-year change association analyses. Eleven modifiable variables showed a consistent average association among the majority of cardiometabolic traits. These variables belonged to the domains: (i) Smoking, (ii) Beverage (filtered coffee), (iii) physical activity, (iv) alcohol intake, and (v) specific variables related to Nordic lifestyle (hunting/fishing during leisure time and boiled coffee consumption). We used an agnostic, data-driven approach to assess a wide range of established and novel risk factors for cardiometabolic disease. Our findings highlight key variables, along with their respective effect estimates, that might be prioritised for subsequent prediction models and lifestyle interventions.

The majority of non-communicable diseases are caused by the complex interplay of genetic and environmental factors. In the last decades, major progress has been made in discovering genetic loci predisposing to these diseases, facilitated by genome-wide association studies (GWAS). These studies allow high-throughput and systematic screening of millions of variants against quantitative traits or hard disease endpoints. Unlike population genetics, there are no standard environment 'chips' that capture multiple environment exposures simultaneously. Therefore, environmental epidemiology typically involves approaches where hypothesized associations between specific environmental exposures and disease traits are separately tested. These studies are limited by the expectations and knowledge about the hypothesized relationships they seek to test, which may cause bias and inhibit discovery[1].

Environment-wide association studies (EWAS) represent an approach through which multiple environmental factors can be systematically screened for their associations with disease traits in a manner that is to a large degree agnostic to prior knowledge about disease associations; in this sense, the EWAS approach is similar to GWAS. EWAS was first described in the published literature in a 2010 paper reporting associations analyses between metabolites and type 2 diabetes[2]. Later, EWAS was used to identify nutrients, environmental contaminants, and prescribed drugs[3–9] associated with disease and disease complications. Almost all published EWAS have used cross-sectional epidemiological data to assess exposures at a fixed time point without consideration of the impact of exposures throughout an individual's lifetime. Longitudinal data analyses may help us understand the associations among exposures and changes in cardiometabolic traits over time.

[1]Department of Clinical Sciences, Genetic and Molecular Epidemiology Unit, Lund University, 214 28 Malmö, Sweden. [2]Cardiovascular Medicine Unit, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden. [3]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [5]Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden. [6]Division of Endocrinology and Diabetes, Cantonal Hospital St. Gallen, St. Gallen, Switzerland. [7]Department of Odontology, Umeå University, Umeå, Sweden. [8]Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [9]These authors contributed equally: Hugo Pomares-Millan and Yan Chen ✉email: paul.franks@med.lu.se

The present study sought to assess the temporal relationships of more than 300 lifestyle exposures (e.g. food items, sleep habits, physical activity, psychosocial factors) with nine cardiometabolic traits (i.e. BMI, blood lipids, blood glucose, and blood pressure) and use these results to identify target lifestyle exposures/exposure groups that could inform lifestyle interventions focused on controlling cardiometabolic diseases.

Methods

Participants. The analyses reported here were undertaken using data from the Västerbotten Health Survey (Västerbottens hälsoundersökning; VHU)[10]. VHU is a prospective, population-based cohort study originally designed as a long-term project intended for health promotion among the general population in Västerbotten county (approx. 254,000 inhabitants), northern Sweden. Since 1985, adults residing in Västerbotten have been invited to undergo a clinical examination and complete lifestyle questionnaires during the years of their 30th, 40th, 50th, and 60th birthdays.

A sub-cohort of VHU ($n = 88{,}614$) was used in the present analyses. Participants with non-Swedish origin ($n = 14{,}629$) were excluded from the analyses as the different cultural and lifestyle habits and disease predisposition of non-Swedish participants may cause confounding by population stratification in EWAS analyses. Participants with diagnosed diabetes and cardiovascular diseases ($n = 3025$) were also excluded to minimize bias attributable to diagnostic labelling and medications. The final dataset comprised 31,362 participants including 67,738 health examinations performed between 1990 and 2013. Written informed consent was obtained from all living participants as part of the VHU. The study was approved by the Regional Ethical Review Board in Umeå, and all research was conducted in accordance to this ethical approval and with the Declaration of Helsinki and other relevant guidelines and regulations.

Clinical measurements. Nine cardiometabolic traits were analysed in the study: body mass index (BMI), systolic and diastolic blood pressures (SBP and DBP, respectively), fasting and 2 h glucose, total cholesterol, triglycerides, HDL cholesterol and LDL cholesterol. Clinical measures in VHU are described in detail elsewhere[10]. In brief, participants' weight (in kg) and height (in cm) were measured using calibrated scale and stadiometer, with participants wearing light clothing and no shoes. BMI was calculated as body weight in kilograms divided by height in meters squared. SBP and DBP were measured once, after 5-min rest, with the participant in a recumbent position using either manual or automated sphygmomanometers. Capillary blood was drawn after overnight fasting and a second blood sample was drawn two hours after the administration of a 75-g oral glucose load. Blood glucose, total cholesterol and triacylglycerol levels were then measured using a Reflotron bench-top analyser (Roche Diagnostics Scandinavia AB). HDL cholesterol was measured in a subgroup of participants and LDL cholesterol was estimated using the Friedewald formula[11]. The measurement for lipids and blood pressure changed in September 2009. From this date onwards, blood pressure was measured twice in a sitting position and averaged, and total cholesterol and triglyceride levels were analysed using clinical chemical analysis in the laboratory. Thus, validated conversion equations were used to align the lipid and blood pressure measurements taken before and after September 2009[12]. For participants on lipid and/or blood pressure lowering medications, lipid and/or blood pressure levels were corrected by adding published constants (+0.208 mmol/l for triglycerides, +1.347 mmol/l for total cholesterol, −0.060 mmol/l for HDL cholesterol, +1.290 mmol/l for LDL cholesterol, +15 mmHg for SBP and +10 mmHg for DBP)[13,14]. Values of cardiometabolic traits located outside the normal range suggested by VHU data managers (see Supplementary Material) were considered outliers and excluded.

Lifestyle assessments. Participants were asked to complete a self-administered questionnaire during each visit that included questions about socio-economic factors, physical/mental health, quality of life, social network and support, working conditions, and alcohol/tobacco consumption. Physical activity was assessed through a modified version of the International Physical Activity Questionnaire[15,16]. A validated semi-quantitative food frequency questionnaire (FFQ) designed to capture habitual diet over the last year was used to capture information on various dietary factors[17]. Up to the mid-1990s, the FFQ consisted of 84 different foods items/groups, but it was reduced to 66 items in 1996 by combining similar line items and by removing items that provided minimal unique information. For the current analysis, matching food items from different FFQ versions were combined in new variables and all analyses including dietary variables were adjusted for FFQ version. In the FFQ, participants indicated how often they consumed foods and beverages on a nine-point frequency scale. Information on average portion size of meat and fish, vegetables, potatoes, rice and pasta was also gathered. Nutrient and energy content were calculated based on the Swedish Food Composition Database[18] based on meal frequency and portion size. Food intake level (FIL) was calculated as total energy intake divided by estimated basal metabolic rate. Participants with more than 10% FFQ data missing, one or more portion indication missing, or a seemingly implausible total energy intake (the top 2.5% and bottom 5% of FIL in the original VHU dataset) were excluded from the analyses. Implausible values for other lifestyle variables (see Supplementary Material) were also removed from the analyses. Lifestyle variables were grouped in 10 different categories to facilitate understanding of the results: (i) alcohol consumption, (ii) non-alcoholic beverage consumption, (iii) food, (iv) nutrients; (v) general health, (vi) physical activity and fitness, (vii) psychosocial, (viii) sleep, (ix) social conditions, (x) tobacco use.

Statistical analysis. The flowchart of the study is shown in Fig. 1. Lifestyle variables were treated either as continuous or as categorical variables; thus, ordinal variables were treated as continuous variables. For categorical variables with more than two levels dummy variables were created and dichotomized. All numeric lifestyle variables were inverse normalized in order to address skewness and scaled for comparability. Similarly, for cat-

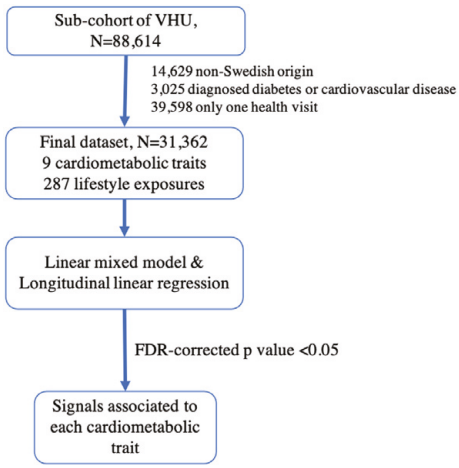


Figure 1. Flow chart of the method followed in the study.

egorical variables, levels were harmonized from low to high, using the lowest one as reference. Thirty-eight categorical variables that had 90% of the observations belonging to one category were excluded from the analyses. In total, the analyses included 242 numeric and 45 categorical lifestyle variables. Dietary variables were regressed on total energy intake and their residuals along with total energy intake were included in the analyses of these variables to account for potential confounding by total energy intake[19]. Models with glycaemic or lipid traits as the dependent variables were additionally adjusted for fasting status. All models (except models having BMI as outcome) were adjusted for BMI.

Average lifestyle associations. Linear mixed models were used to estimate an average linear effect of the lifestyle exposures on the cardiometabolic traits. The models were adjusted for age, age^2, sex, educational level, follow-up time, FFQ version (where appropriate), total energy intake (TEI; where appropriate), BMI (where appropriate) and fasting status (where appropriate).

$$\begin{aligned}\gamma_{ij} = {} & (\beta_{00} + \mu_{0j}) + \beta_{10}\text{age}_{ij} + \beta_{20}\text{age}_{ij}^2 + \beta_{30}\text{sex}_{ij} + \beta_{40}\text{follow} \\ & - \text{up time}_{ij} + \beta_{50}\text{FFQ version}_{ij} + \beta_{60}\text{TEI}_{ij} + \beta_{70}\text{BMI}_{ij} \\ & + \beta_{80}\text{fasting status}_{ij} + \beta_{90}\text{lifestyle variable}_{ij} + \varepsilon_{ij}\end{aligned} \quad (1)$$

where γ_{ij} represents a cardiometabolic trait value at visit i for participant j, β00 is the fixed intercept, μ_{0j} represents different random intercepts for each participant, the rest of the β estimates are the estimated fixed effect size parameters for each corresponding variable, and ε represents error.

Long-term lifestyle associations. Linear regression models were used to test if the lifestyle variables were associated with 10-year changes in the cardiometabolic traits:

$$\begin{aligned}\gamma_F = {} & \alpha + \beta_1\text{age}_B + \beta_2\text{age}_B^2 + \beta_3\text{sex} \\ & + \beta_4\text{follow up time}_+\beta_6\gamma_B + \beta_7\text{FFQ version}_B + \beta_8\text{TEI}_B + \beta_9\text{meanBMI} \\ & + \beta_{10}\text{fasting status}_B + \beta_{11}\text{fasting status}_F + \beta_{12}\text{lifestyle variable}_B + \varepsilon\end{aligned} \quad (2)$$

where γ_F represents the value of the cardiometabolic trait at follow-up and γ_B the value at baseline, α is the intercept, β_i represent the estimated effect size parameter for each corresponding variable. Age_B, FFQ version_B, TEI_B, fasting status_B and $\text{lifestyle variable}_B$ are the age, FFQ version, TEI, fasting status and lifestyle variable values at baseline; fasting status_F is the fasting status value at follow up; meanBMI is the average BMI of the baseline and follow-up BMI values, and ε represents error.

Tentative signals. The Benjamini and Hochberg[20] False Discovery Rate (FDR) was used to correct for multiple testing. Associations of lifestyle variables were considered "tentative signals" if they achieved significance at $P_{FDR} < 0.05$ after multiple testing correction. Overall estimates were used in the description of the results and effect estimates are reported in Supplementary material.

Variable	Number of observations	mean	SD
Age (years)	67,738	47.72	8.92
Height (cm)	67,476	172.04	9.21
BMI (kg/m2)	67,413	25.73	3.99
Waist circumference (cm)	28,621	92.27	12.13
Total cholesterol (mmol/L)	67,181	5.51	1.09
HDL-C (mmol/L)	27,803	1.40	0.47
LDL-C (mmol/L)	27,649	3.86	1.03
Triglycerides (mmol/L)	58,905	1.40	0.78
Fasting glucose (mmol/L)	67,339	5.39	0.75
2-h glucose (mmol/L)	64,951	6.57	1.50
SBP (mmHg)	67,193	126.42	17.54
DBP (mmHg)	67,160	78.83	11.29

Table 1. Summary of participant characteristics. *BMI* Body mass index, *HDL-C* High-density lipoprotein cholesterol, *LDL-C* Low-density lipoprotein cholesterol, *SBP* Systolic blood pressure, *DBP* Diastolic blood pressure, *SD* Standard deviation.

Correlation patterns. Correlations between 'tentative signals' on the linear mixed and/or longitudinal linear regression analyses were calculated and visualized using a heatmap. A hierarchical clustering algorithm was used to arrange lifestyle variables, so that the pair of variables with higher correlations appear closer in the heatmap.

Prioritization of modifiable lifestyle variables. Tentative signals for each of the cardiometabolic traits were gathered and prioritized to identify target lifestyle exposures and exposure groups in which lifestyle interventions aiming at controlling cardiometabolic diseases may focus. First, variance explained for each lifestyle variable (and covariates) was estimated and variables were rank-ordered within each lifestyle category for each of the nine outcome traits. In the linear mixed models, marginal (fixed terms) variance explained was used. The top-ranked variables (five per category per trait) were identified, and the topranked variables represented in the majority of the cardiometabolic traits (at least five traits) were prioritized. Target groups were evaluated using a hierarchical clustering algorithm based on correlations between the prioritized variables and visualized in a heatmap. Non-modifiable variables were excluded from the prioritization and clustering step as these variables could not be affected by a lifestyle intervention.

Statistical analyses and data visualization were performed using *R* software versions 3.5.2 and 3.6.1[21] (see Supplementary Material for the specific packages used for analyses).

Results

Descriptive characteristics of the study population are summarized in Tables 1, S1 and S2. Mean age of participants was 47.7 years and 50.6% were women.

Average lifestyle associations. 164 out of 286 lifestyle variables were considered tentative signals for BMI (S3), 37 for SBP (S4), 30 for DBP (S5), 84 for total cholesterol (S6), 96 for triglycerides (S7), 46 for HDL cholesterol (S8), 20 for LDL cholesterol (S9), 44 for fasting glucose (S10) and 43 for 2 h glucose (S11). 'Physical activity' and 'General health' were the top categories for BMI (Fig. 2) and 'General health' for blood pressure traits (Fig. 3). Regarding lipids, 'Beverage', 'Nutrients' and ' Physical activity' were the categories with the highest number of 'tentative signals' for total and LDL cholesterol (Figs. 4A and D), while 'Physical activity', 'Tobacco use' and 'General health' were the top categories for triglycerides (Fig. 4B), and 'Alcohol' for HDL cholesterol (Fig. 4C). For glucose traits, 'Physical activity', 'General health' and 'Tobacco use' were the top categories (Fig. 5).

Long-term lifestyle associations. After multiple testing correction, 35 lifestyle variables showed a tentative association with 10-year change in BMI (S12), 3 with change in SBP and DBP (S13-S14), 15 with change in total cholesterol (S15), 10 in triglycerides (S16), none in HDL and LDL cholesterol (S17-S18), 5 in fasting glucose (S19) and 8 in 2 h glucose (S20). The majority of the 'tentative signals' were in the 'Tobacco use' category for BMI, lipids and fasting glucose, while for blood pressure traits the top category was 'General health' and for 2 h glucose, 'Physical activity', 'Food', and 'General health' were the top categories. There were no material changes in key outcome variables during the 9-year follow-up period (see Supplementary Material).

Correlation patterns. Patterns of correlations were identified among lifestyle variables showing tentative association with any of the cardiometabolic traits based on the correlation heatmap (Fig. 6). Variables related to meat and fish consumption, sodium, calcium, vitamin B12, and total and animal based protein intake appeared in close proximity showing correlations around 0.5. Variables describing fat consumption and fatty acid intakes were grouped together showing a high positive correlation. Variables assessing vegetable, fibre and fruit intake, plant lignans, whole grain intake, and carbohydrates intake also appear near each other in the heatmap showing

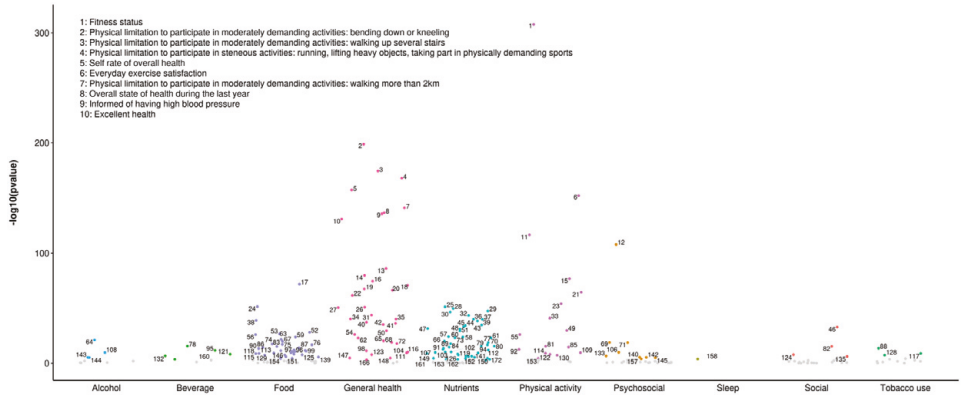


Figure 2. Manhattan plot representing the distribution of *P* values of the association of lifestyle variables and BMI by lifestyle category. Tentative signals are coloured, and number labelled in the figure and the top 10 variables are spelled out. See S25 for references to the labels.

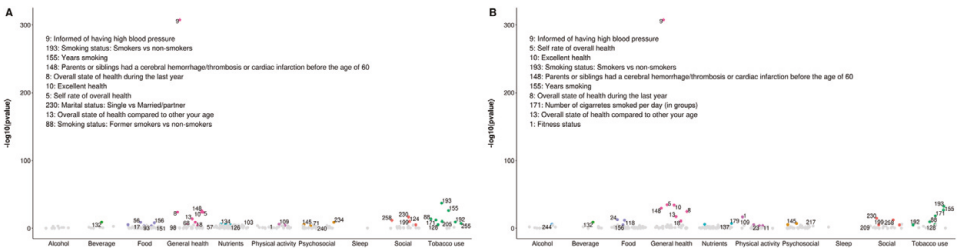


Figure 3. Manhattan plot representing the distribution of *P* values of the association of lifestyle variables and blood pressure traits by lifestyle category. (**A**) Systolic blood pressure and (**B**) Diastolic blood pressure. 'Tentative signals' are coloured, and number labelled in the figure and the top 10 variables are spelled out. See S25 for references to the labels.

high positive correlations between them and negative correlations with fat related variables. Variables in 'Psychosocial' category and 'General health' variables were grouped together.

Prioritization of modifiable lifestyle variables. *Average lifestyle associations.* Thirteen variables were prioritized among all the 'tentative signals' as they showed the most consistent associations across all the cardiometabolic traits (top-ranked in at least 5 out of 9 cardiometabolic traits) (S21). Two of these variables ('Informed of having a high blood pressure' and 'Overall state of health during the last year') were considered non-modifiable and excluded (S26 for modifiable and non-modifiable variables). The eleven remaining variables were included in a hierarchical clustering algorithm which identified four main targets suitable for interventions (Fig. 7). The first group included tobacco use/smoking related variables and were in general positively associated with BMI, fasting glucose, total cholesterol and triglycerides and negatively with blood pressure traits, HDL cholesterol and 2 h glucose (S21). The second included 'Brewed (filtered) coffee', which was negatively associated with BMI, blood pressure traits, triglycerides and 2 h glucose. The third group included physical activity related variables (e.g. 'Exercise during the last three months'). The fourth included the variable 'alcohol intake (g/day)'. These variables were in general negatively associated with all cardiometabolic traits except with HDL-C with which they showed a positive association. The fifth group was a composite of lifestyle variables which could be linked to the Swedish lifestyle (especially northern Swedish lifestyle), 'Frequency of hunting or fishing during leisure time' and 'Boiled coffee' (S26). These two variables did not show a clear common pattern of associations with cardiometabolic traits.

In general, BMI showed more shared tentative signals with 2 h glucose and HDL-cholesterol than with the rest of cardiometabolic traits and triglycerides, BMI and 2 h glucose were the cardiometabolic traits sharing the highest number of tentative signals with the rest of cardiometabolic traits (S22).

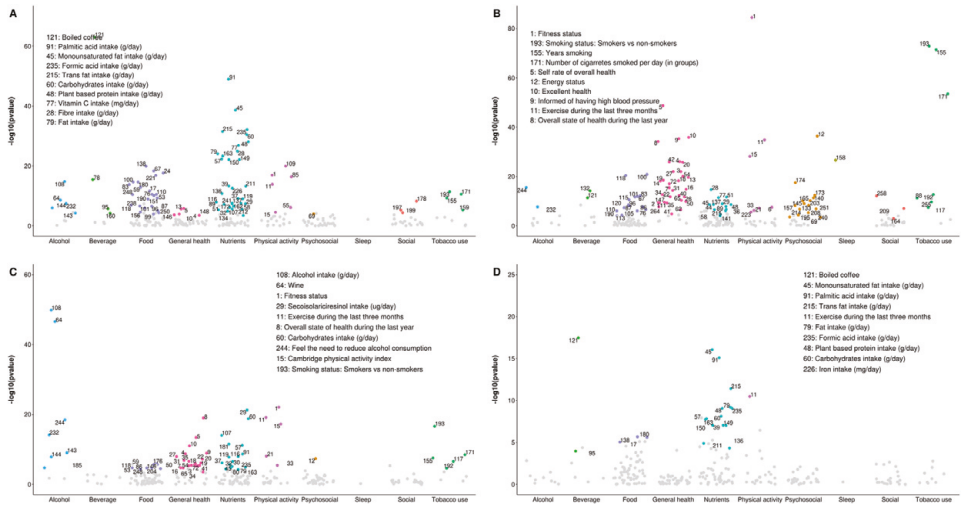


Figure 4. Manhattan plot representing the distribution of *P* values of the association of lifestyle variables and lipid traits by lifestyle category. (**A**) Total cholesterol, (**B**) Triglycerides, (**C**) HDL cholesterol and (**D**) LDL cholesterol. Tentative signals are coloured, and number labelled in the figure and the top 10 variables are spelled out. See S25 for references to the labels.

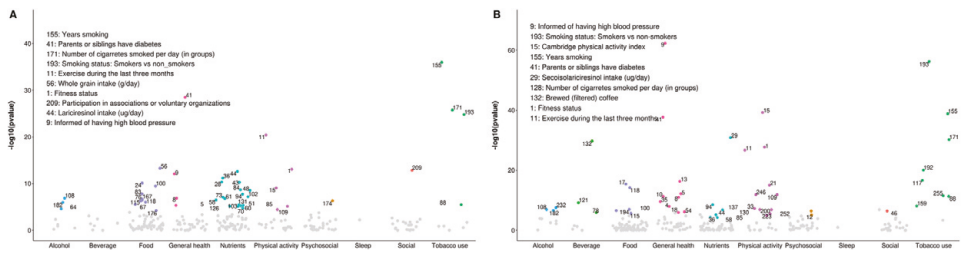


Figure 5. Manhattan plot representing the distribution of *P* values of the association of lifestyle variables and glucose traits by lifestyle category. (**A**) Fasting glucose and (**B**) 2 h glucose. Tentative signals are coloured, and number labelled in the figure and the top 10 variables are spelled out. See S25 for references to the labels.

Long-term lifestyle associations. None of the 'tentative signals' showed a consistent association with the majority of cardiometabolic traits (5 out 9 traits) (S23). However, four variables in the 'Tobacco use' category showed a consistent positive association with 10-year changes in at least three cardiometabolic traits (BMI, total cholesterol, triglycerides and/or fasting glucose).

Among all the cardiometabolic traits BMI and lipid traits shared the highest number of tentative signals (S24).

Discussion

Although EWAS analyses have been reported previously, this is the first study to integrate repeated exposures and outcome assessments, which allows inferences about long-term exposure to these risk factors to be made. Here, we systematically and agnostically assessed average (across the study's follow-up time) and ~ 10-year associations between 286 lifestyle variables and 9 cardiometabolic traits. In analyses assessing average association of lifestyle variables, 'Physical activity' and 'General Health' were the categories containing the highest number of tentative signals and 11 modifiable variables were prioritized for lifestyle interventions focused on controlling cardiometabolic diseases. A cluster analyses grouped these 11 variables into five main target groups: (i) Smoking, (ii) Beverage (filtered coffee), (iii) physical activity, (iv) alcohol intake, and (v) specific variables related to Swedish lifestyle (hunting/fishing during leisure time and boiled coffee).

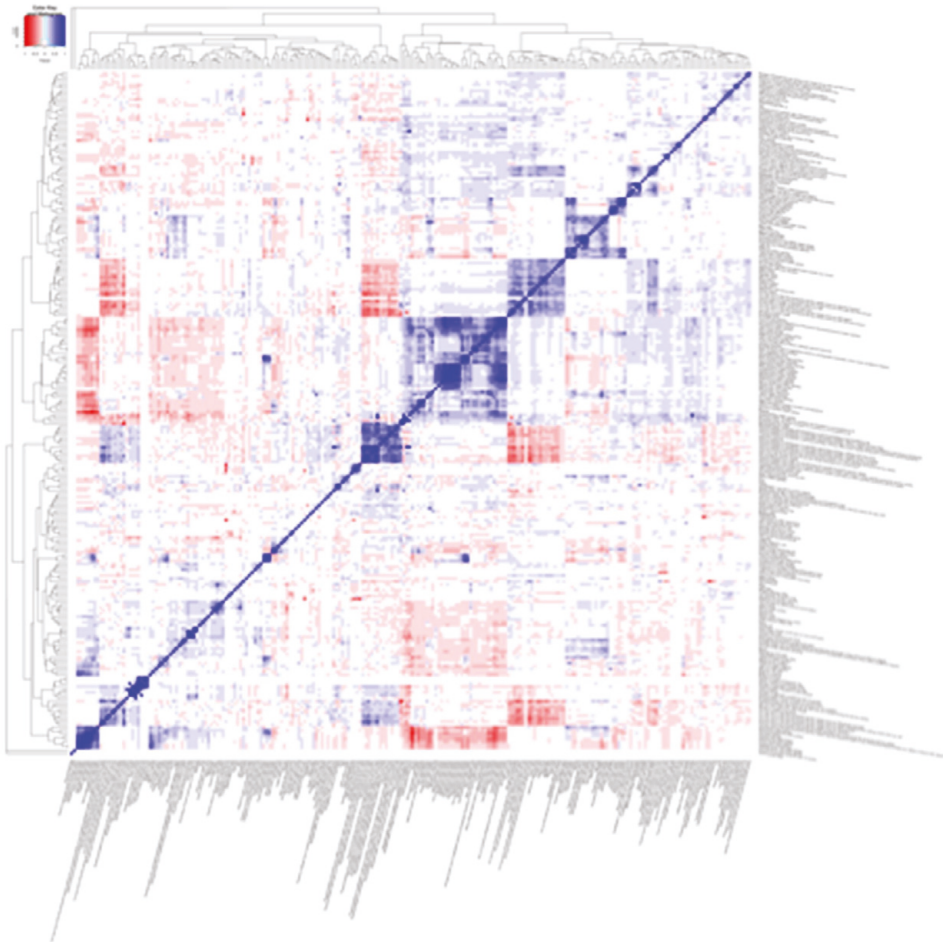


Figure 6. Heat map showing all the correlations for tentative signals. Pairs of factors where correlations could not be computed are shown in white. Figures were plotted using 'ggplot2', 'ggrepel', 'gridExtra', 'RColorBrewer' and 'gplots' packages in *R* software versions 3.5.2 and 3.6.1[21].

For 10-year associations, 'Tobacco use' was the category including the highest number of tentative signals for the majority of the cardiometabolic traits. No modifiable lifestyle variable was consistently associated with the majority of cardiometabolic traits but four variables in the 'Tobacco use' category were consistently associated with at least three of the analysed cardiometabolic traits (BMI, total cholesterol, triglycerides and/or fasting glucose).

Smoking and physical activity correspond to two of the most well-known modifiable risk factors for cardiometabolic diseases. According to a study analysing the burden of disease caused by physical inactivity, worldwide, 6% of the burden of coronary heart disease and 7% of type 2 diabetes was caused by physical inactivity[22]. On the other hand, smoking alters lipid metabolism and glucose homeostasis through the increase in lipolysis, insulin resistance and tissue lipotoxicity[23,24] and smoking cessation restores, at least in part, these metabolic alterations. However, in our study the association of smoking with cardiometabolic traits was not only restricted to the average effect across the studied period but we also found a remarkable association of variables included in the 'Tobacco use' category and cardiometabolic traits in the 10 years of follow-up.

Among the prioritized dietary variables, boiled (unfiltered) coffee but not brewed (filtered) coffee was found positively associated with lipid traits, specifically with total cholesterol, triglycerides, and LDL cholesterol.

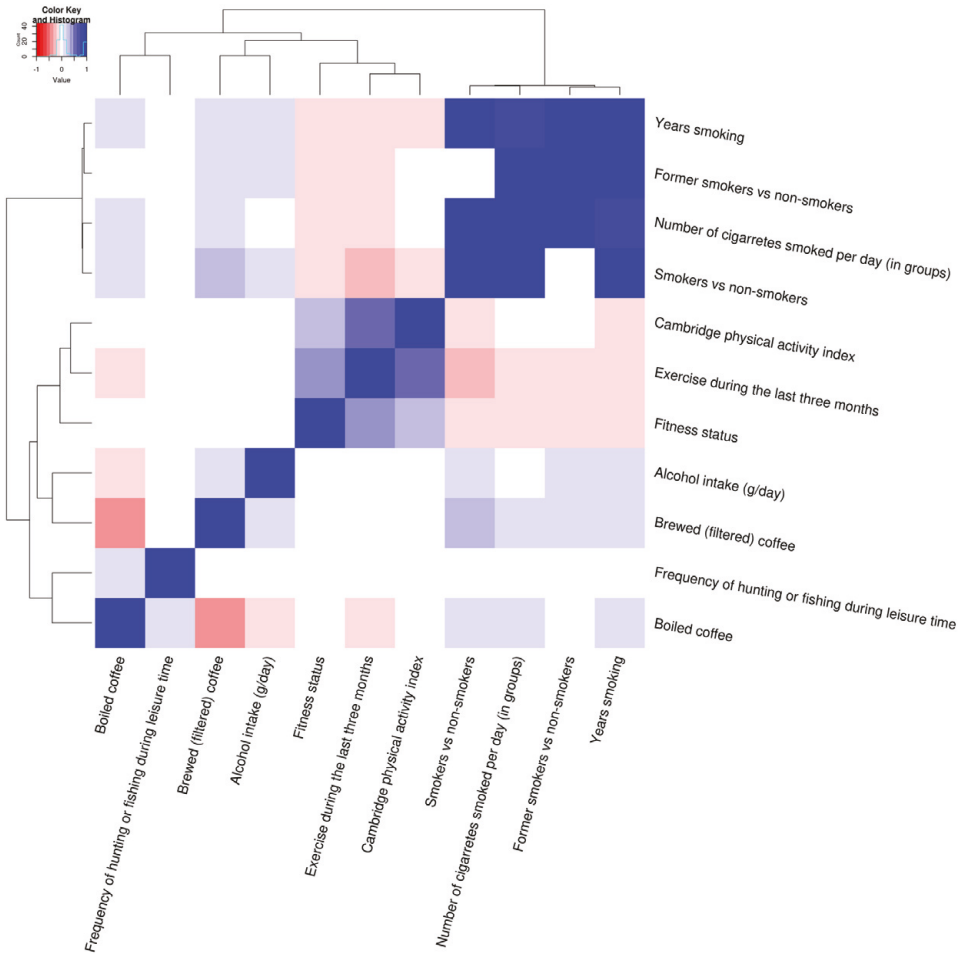


Figure 7. Heat map showing clusters of correlations between top-ranked modifiable lifestyle variables. Figures were plotted using 'ggplot2', 'ggrepel', 'gridExtra', 'RColorBrewer' and 'gplots' packages in *R* software versions 3.5.2 and 3.6.1[21].

Previous studies have also identified associations between unfiltered coffee and dose-dependent increase of plasma concentrations of total and LDL cholesterol[25,26]. The effects of coffee in the lipid profile are probably caused by two diterpenes (i.e. kahweol and cafestol), which sometimes get trapped in the filter used to make coffee which can explain the differential effects of filtered and unfiltered coffee[26]. On the other hand, brewed (filtered) coffee was found negatively associated with BMI, blood pressure, triglycerides, and 2 h glucose in the present study which is in agreement with previous studies showing an inverse association between habitual coffee intake and risk of several cardiometabolic diseases[27,28].

Plant lignans (biphenolic compounds found in tea, coffee, whole-grain products, berries, vegetables, fruit, nuts and seeds) were among the top tentative signals for fasting and 2 h glucose, showing a negative association with both traits. Previous studies have suggested that lignans and their metabolites may protect against cardiovascular disease and metabolic syndrome by reducing lipid concentrations, lowering blood pressure, and decreasing oxidative stress and inflammation[29]. A study conducted in Finland found that men with high serum concentrations of enterolactone (a lignan produced by the intestinal microflora) had a lower risk of acute coronary events than men with lower concentrations[30].

An interesting observation emerging from our analysis is that several variables that are featured in public health recommendations were not broadly associated with the cardiometabolic traits studied here. Recommended dietary patterns emphasize the importance of limiting the consumption of sugar-rich products, particularly sweet drinks[31]. However, variables related to sweets and sweet drink consumption (e.g. "Sodas, soft drinks, juice" and "Sweets") were not identified as tentative signals for any of the cardiometabolic traits. Salt content is also usually limited in diets recommended to lower risk of cardiometabolic diseases but "Sodium intake" was not consistently associated with cardiometabolic traits, being identified as a tentative signal only for BMI, total and HDL cholesterol. In the same way, fish and shellfish are frequently recommended in healthy dietary patterns but "Lean fish" and "Shellfish" variables were not tentative signals for any cardiometabolic traits, and "Fatty fish" was associated with lipid traits except for LDL cholesterol.

There are also limitations to the present study. EWAS and GWAS are not entirely analogous. However, both are experiment-wide association studies that adopt a so called 'agnostic' approach to consider a multitude of exposure-outcome relationships in parallel. This is hence a 'data-driven' approach that contrasts traditional association studies, where specific hypotheses are formulated and only those relationships consistent with the hypothesis are tested. The present sample is limited to a Swedish population between 30–70 years and thus caution should be used when extrapolating the findings to other countries and age groups, especially since lifestyle variables affecting cardiometabolic traits in Swedish population might differ from other populations. Dietary variables were characterized using an FFQ, which suffer from systematic and random measurement errors. However, to minimize this source of error the FFQ used in this study was validated against repeated 24 h recalls[17]. VHU cohort is exceptionally well-powered for analyses of the nature performed here and there were, consequently, a large number of associations that passed conventional statistical thresholds. Most of these statistically robust associations emerged due to the complex correlation structure (Fig. 6) found within the set of exposure variables. The EWAS analyses undertaken here, like those reported elsewhere, involve parallel tests of association with cardiometabolic traits for an array of variables, in this case modifiable lifestyle exposures. As with all observational analyses in free-living populations, including EWAS, there is a risk that the relationships observed are prone to confounding and reverse-causality. To mitigate these risks, we adjusted the regression models for putative confounding variables and assessed the key findings in both average and long-term models. Even with these attempts, it is important to highlight that one or more of the findings are false-positive owing to residual confounding. To assess this thoroughly requires appropriately designed experimental studies. Our findings highlight key variables, along with their respective effect estimates, that might be prioritised for subsequent prediction models and lifestyle interventions. However, it is important to keep in mind that epidemiological associations of this nature may not be causal. Thus, intervention studies are needed to test the causal nature of these associations.

In conclusion, using an EWAS approach in a large prospective Swedish cohort a large number of associations between lifestyle exposures and cardiometabolic traits were identified. Eleven modifiable exposures were consistently top-ranked among the majority of cardiometabolic traits and were identified as target lifestyle exposures that could inform lifestyle interventions aiming at controlling cardiometabolic diseases. These variables belonged to four target groups: (i) Smoking, (ii) Beverage (specifically brewed (filtered) coffee) and (iii) Leisure time physical activity and (iv) a group of lifestyles more specific to the Swedish lifestyle.

Received: 25 March 2021; Accepted: 28 February 2022
Published online: 08 March 2022

References

1. Patel, C. J. & Ioannidis, J. P. Studying the elusive environment in large scale. *JAMA* **311**(21), 2173–2174 (2014).
2. Patel, C. J., Bhattacharya, J. & Butte, A. J. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS ONE* **5**(5), e10746 (2010).
3. Tzoulaki, I. *et al.* A nutrient-wide association study on blood pressure. *Circulation* **126**(21), 2456–2464 (2012).
4. Lind, P. M. *et al.* An environmental wide association study (EWAS) approach to the metabolic syndrome. *Environ. Int.* **55**, 1–8 (2013).
5. Hall, M.A., Dudek, S.M., Goodloe, R. *et al.* Environment-wide association study (EWAS) for type 2 diabetes in the Marshfield personalized medicine research project biobank. *Pac. Symp. Biocomput.* 200–211 (2014).
6. McGinnis, D. P., Brownstein, J. S. & Patel, C. J. Environment-wide association study of blood pressure in the national health and nutrition examination survey (1999–2012). *Sci. Rep.* **6**, 30373 (2016).
7. Patel, C. J. *et al.* Systematic identification of correlates of HIV infection: an X-wide association study. *AIDS* **32**(7), 933–943 (2018).
8. Patel, C. J. *et al.* Systematic assessment of pharmaceutical prescriptions in association with cancer risk: a method to conduct a population-wide medication-wide longitudinal study. *Sci. Rep.* **6**, 31308 (2016).
9. Zhuang, X. *et al.* Environment-wide association study to identify novel factors associated with peripheral arterial disease: evidence from the national health and nutrition examination survey (1999–2004). *Atherosclerosis* **269**, 172–177 (2018).
10. Hallmans, G. *et al.* Cardiovascular disease and diabetes in the Northern Sweden Health and Disease Study Cohort - evaluation of risk factors and their interactions. *Scand. J. Publ. Health Suppl.* **61**, 18–24 (2003).
11. Friedewald, W. T., Levy, R. I. & Fredrickson, D. S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin. Chem.* **18**(6), 499–502 (1972).
12. Ng, N. *et al.* Trends of blood pressure levels and management in Vasterbotten County, Sweden, during 1990–2010. *Glob. Health Action* **5**, 1–12 (2012).
13. Wu, J. *et al.* An investigation of the effects of lipid-lowering medications: genome-wide linkage analysis of lipids in the HyperGEN study. *BMC Genet.* **8**, 60 (2007).
14. Tobin, M. D. *et al.* Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Stat. Med.* **24**(19), 2911–2935 (2005).
15. Hallal, P. C. & Victora, C. G. Reliability and validity of the International physical activity questionnaire (IPAQ). *Med. Sci. Sports Exerc.* **36**(3), 556 (2004).

16. Craig, C. L. *et al.* International physical activity questionnaire: 12-country reliability and validity. *Med. Sci. Sports Exerc.* **35**(8), 1381–1395 (2003).
17. Johansson, I. *et al.* Validation and calibration of food-frequency questionnaire measurements in the Northern Sweden Health and Disease cohort. *Public Health Nutr* **5**(3), 487–496 (2002).
18. Livsmedelsverket. The food database. (https://www.livsmedelsverket.se/en/food-and-content/naringsamnen/livsmedelsdatabasen). (Accessed 29th June 2018).
19. Willett, W. C., Howe, G. R. & Kushi, L. H. Adjustment for total energy intake in epidemiologic studies. *Am. J. Clin. Nutr.* **65**(4 Suppl), 1220S-1228S (1997) (**discussion 9S-31S**).
20. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Met.* **57**(1), 289–300 (1995).
21. R Development Core Team. R: A language and environment for statistic computing. Vienna, Austria: R Foundation for Statistic Computing (2015).
22. Lee, I. M. *et al.* Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *Lancet* **380**(9838), 219–229 (2012).
23. Chelland Campbell, S., Moffatt, R. J. & Stamford, B. A. Smoking and smoking cessation – the relationship between cardiovascular disease and lipoprotein metabolism: a review. *Atherosclerosis* **201**(2), 225–235 (2008).
24. Sliwinska-Mosson, M. & Milnerowicz, H. The impact of smoking on the development of diabetes and its complications. *Diab. Vasc. Dis. Res.* **14**(4), 265–276 (2017).
25. Jee, S. H. *et al.* Coffee consumption and serum lipids: a meta-analysis of randomized controlled clinical trials. *Am. J. Epidemiol.* **153**(4), 353–362 (2001).
26. Penson, P. *et al.* Does coffee consumption alter plasma lipoprotein(a) concentrations? A systematic review. *Crit. Rev. Food Sci. Nutr.* **58**(10), 1706–1714 (2018).
27. Carlstrom, M. & Larsson, S. C. Coffee consumption and reduced risk of developing type 2 diabetes: a systematic review with meta-analysis. *Nutr. Rev.* **76**(6), 395–417 (2018).
28. Crippa, A. *et al.* Coffee consumption and mortality from all causes, cardiovascular disease, and cancer: a dose-response meta-analysis. *Am. J. Epidemiol.* **180**(8), 763–775 (2014).
29. Adolphe, J. L. *et al.* Health effects with consumption of the flax lignan secoisolariciresinol diglucoside. *Br. J. Nutr.* **103**(7), 929–938 (2010).
30. Vanharanta, M. *et al.* Risk of acute coronary events according to serum concentrations of enterolactone: a prospective population-based case-control study. *Lancet* **354**(9196), 2112–2115 (1999).
31. Nordic Council. *Nordic Nutrition Recommendations 2012. Integrating Nutrition and Physical Activity.* 5th ed. (2012).

Acknowledgements

We thank the participants, health professionals and data managers involved in the Västerbotten Intervention Programme.

Author contributions

A.P., A.K., F.R. and P.W.F. designed the study and directed its implementation. A.P., H.P.M. and Y.C. analysed the data. H.P.M. and Y.C. authors contributed equally. C.J.P. helped with the EWAS scripts. G.H. and I.J. helped with the acquisition of the data. A.P., Y.C. and P.W.F. drafted the manuscript. All the authors reviewed the manuscript and comment on it.

Funding

Open access funding provided by Lund University. The work described in this paper was supported by the Innovative Medicines Initiative of the European Union (No. 875534 – SOPHIA), 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).

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/s41598-022-08050-1.

Correspondence and requests for materials should be addressed to P.W.F.

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.

Paper II

nutrients

MDPI

Article

Predicting Sensitivity to Adverse Lifestyle Risk Factors for Cardiometabolic Morbidity and Mortality

Hugo Pomares-Millan [1], Alaitz Poveda [1], Naemieh Atabaki-Pasdar [1], Ingegerd Johansson [2], Jonas Björk [3,4], Mattias Ohlsson [5,6], Giuseppe N. Giordano [1] and Paul W. Franks [1,2,7,*]

[1] Department of Clinical Sciences Malmö, Lund University Diabetes Centre, Lund University, 21428 Malmö, Sweden; hugo.pomares-millan@med.lu.se (H.P.-M.); alaitz.poveda@med.lu.se (A.P.); naeimeh.atabaki_pasdar@med.lu.se (N.A.-P.); giuseppe.giordano@med.lu.se (G.N.G.)
[2] Department of Public Health and Clinical Medicine, Umeå University, 90187 Umeå, Sweden; ingegerd.johansson@umu.se
[3] Division of Occupational and Environmental Medicine, Lund University, 22363 Lund, Sweden; jonas.bjork@med.lu.se
[4] Clinical Studies Sweden, Forum South, Skåne University Hospital, 22185 Lund, Sweden
[5] Computational Biology and Biological Physics Unit, Department of Astronomy and Theoretical Physics, Lund University, 22100 Lund, Sweden; mattias.ohlsson@thep.lu.se
[6] Center for Applied Intelligent Systems Research, Halmstad University, 30118 Halmstad, Sweden
[7] Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
* Correspondence: paul.franks@med.lu.se; Tel.: +45-2063-2350

Abstract: People appear to vary in their susceptibility to lifestyle risk factors for cardiometabolic disease; determining *a priori* who is most sensitive may help optimize the timing, design, and delivery of preventative interventions. We aimed to ascertain a person's degree of resilience or sensitivity to adverse lifestyle exposures and determine whether these classifications help predict cardiometabolic disease later in life; we pooled data from two population-based Swedish prospective cohort studies (n = 53,507), and we contrasted an individual's cardiometabolic biomarker profile with the profile predicted for them given their lifestyle exposure characteristics using a quantile random forest approach. People who were classed as 'sensitive' to hypertension- and dyslipidemia-related lifestyle exposures were at higher risk of developing cardiovascular disease (CVD, hazards ratio 1.6 (95% CI: 1.3, 1.91)), compared with the general population. No differences were observed for type 2 diabetes (T2D) risk. Here, we report a novel approach to identify individuals who are especially sensitive to adverse lifestyle exposures and who are at higher risk of subsequent cardiovascular events. Early preventive interventions may be needed in this subgroup.

Keywords: cardiometabolic risk factors; risk assessment; quantile random forests; prediction interval; sensitivity; lifestyle

Citation: Pomares-Millan, H.; Poveda, A.; Atabaki-Pasdar, N.; Johansson, I.; Björk, J.; Ohlsson, M.; Giordano, G.N.; Franks, P.W. Predicting Sensitivity to Adverse Lifestyle Risk Factors for Cardiometabolic Morbidity and Mortality. *Nutrients* **2022**, *14*, 3171. https://doi.org/10.3390/nu14153171

Academic Editor: Arrigo Cicero

Received: 15 July 2022
Accepted: 29 July 2022
Published: 1 August 2022

1. 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 therapies (both lifestyle and pharmacologic) exist that can help prevent these diseases [1]. Moreover, once manifest, T2D and CVD often cause life-threatening health complications that are often 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 either viewed as a consequence of measurement error [3] or ignored. However, some of this variability is likely to reflect between-person differences in the effects of unhealthful

lifestyle exposures, with some people more susceptible to the adverse effects of these exposures than others.

Predictive modeling often provides a point estimate that represents a response to be anticipated for; yet, in precision medicine a range of values where an effect would be expected to fall may prove more informative for the design of preventive measures rather than a single estimate. Thus, prediction intervals (PIs) allow examining a future series of values for each individual with a given probability, making them potentially useful for identifying where the future value is likely to appear.

Identifying subpopulations who are especially sensitive to adverse lifestyle exposures may help optimize the delivery of cardiometabolic disease prevention programs, especially when resources are lacking [1,4]. In aging and diseased individuals, conditions such as frailty syndrome and nutritional deficiencies often coexist with cardiometabolic disease (i.e., T2D and hypertension) [5]; however, it remains unclear whether vulnerability status associated with adverse environments can be present in disease-free individuals. Here, we used a machine learning approach [6] to differentiate error from true between-individual variability in susceptibility to lifestyle risk factors for T2D and CVD. Accordingly, we identified the subgroup of sensitive individuals and assessed the degree to which this classification aids the prediction of incident disease and premature mortality.

2. Materials and Methods

2.1. Study Design and Participants

The Västerbotten Health Survey (Västerbottens hälsoundersökning; VHU) [7,8] 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 are invited to attend their primary care center to undertake a baseline clinical examination and complete detailed lifestyle questionnaires during the calendar years of their 40th, 50th, and 60th birthdays. We used data derived from VHU (n = 42,887) in our analyses. A total of 7039 of these participants were born outside Sweden, and the current analysis focused only on the Swedish-born contingent of VHU. Participants in whom diabetes or cardiovascular disease were 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 behaviors. Participants with two health examinations between 1985 through 2016 (with ~10 years between each visit) were included in the final dataset, which comprised 35,440 participants.

2.2. MDCS

The Malmö Diet and Cancer Study (MDCS) is a prospective, population-based cohort study conducted between 1991 and 1996. All men and women residing in the city of Malmö, southern Sweden born between 1923 to 1950 were invited to participate. Up to 30,446 participants (~40% men) completed the baseline assessment [9–11]. Glycemic and lipid traits were assessed in a subset of participants, the *MDCS Cardiovascular Cohort* (MDCS-CC; n = 6103), who were randomly selected for assessment of cardiometabolic risk markers between 1991 and 1994 [12]. As with the VHU cohort, data from non-Swedish participants and those with prevalent diabetes or CVD were removed prior to analysis. In total, a maximum of 18,067 CC participants were included in the analysis from MDCS or MDCS-CC (see flowchart in Figure 1).

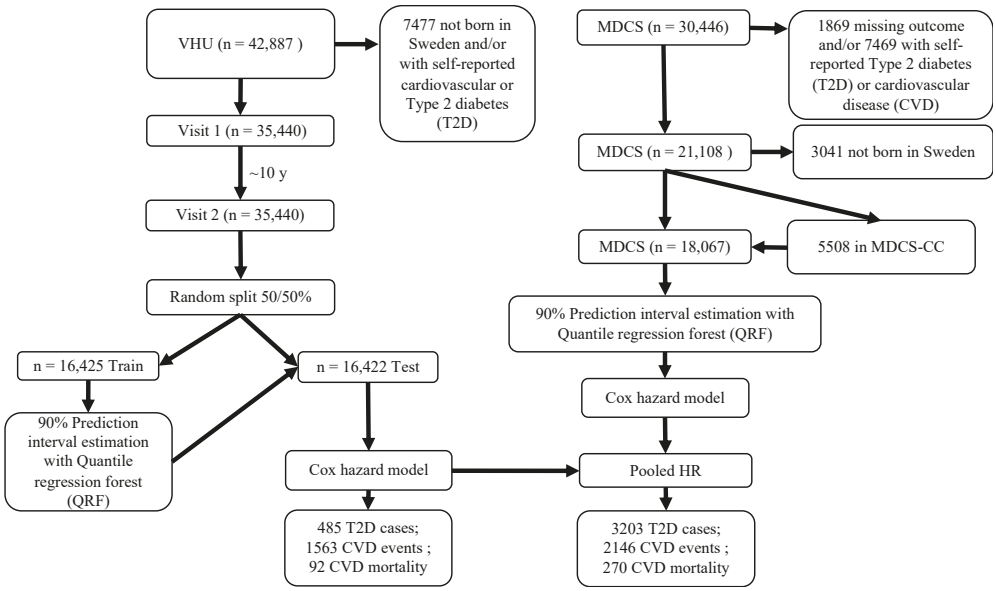


Figure 1. Study flowchart of VHU and MDC studies, data processing, and model training. VHU: Västerbotten Health Survey; MDCS: Malmö Diet and Cancer Study; MDCS-CC: MDCS Cardiovascular Cohort.

2.3. Cardiometabolic Risk Markers

Clinical assessment methods in VHU [7] and MDCS are reported elsewhere [9,12]. Briefly, height and weight were measured with calibrated stadiometer and weighing scales respectively, with participants wearing light clothing and no shoes. Body mass index (BMI) was calculated as the body weight in kilograms divided by height in meters squared. Systolic and diastolic blood pressures were measured with participants resting supine, using either manual or automated sphygmomanometers. Peripheral blood was drawn after overnight fasting, and a venous blood sample was drawn two hours after the administration of a 75 g oral glucose load (the latter only in VHU). Blood glucose (i.e., fasting and 2 h glucose), total cholesterol, and triglyceride levels were then measured using a Reflotron bench-top analyzer (Roche Diagnostics Scandinavia AB); HbA1c was measured only in MDCS-CC using standard procedures at the Department of Clinical Chemistry, University Hospital Malmö. High-density lipoprotein cholesterol (HDL-C) was also measured, and low-density lipoprotein cholesterol (LDL-C) was calculated using the Friedewald formula [13]. In September 2009, blood lipids and blood pressure measurements in VHU changed; thereafter, blood pressure was measured twice in a sitting position and averaged. Triglycerides and total cholesterol levels were analyzed 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 [14]. 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 triglycerides, +1.347 mmol/L for total cholesterol, −0.060 mmol/L for HDL-C, +1.290 mmol/L for LDL-C, +15 mm Hg for systolic, and +10 mm Hg for diastolic blood pressure) suggested in the literature [15,16]. Cardiometabolic trait values outside the thresholds for plausible values suggested by VHU data managers were considered outliers and removed in all datasets (Supplementary Table S1 in Supplementary Materials).

2.4. Lifestyle and Dietary Assessments

For both Swedish cohorts, all participants were requested to complete a self-administered, validated, comprehensive lifestyle questionnaire during each visit, which queried socioeconomic factors, physical/mental health, quality of life, social network and support, working conditions, and alcohol/tobacco use. In VHU, physical activity was assessed using the modified version of the International Physical Activity Questionnaire [17,18], and a validated semiquantitative food frequency questionnaire (FFQ), designed to capture habitual diet over the last year, was used to obtain information on various dietary factors [19]. In 1996, the FFQ was reduced from 84 to 66 items by merging similar items and removing those deemed redundant. For MDCS, a modified diet history method consisting of a 7-day food diary covering all cooked meals and a 168-item FFQ covering the noncooked meals for the previous year were administered. Moreover, a 1 h interview was used to determine portion sizes, cooking methods and food choices. Nutrient and energy contents were calculated using the Swedish Food Composition Database (https://www.livsmedelsverket.se/en/food-and-content/naringsamnen/livsmedelsdatabasen; accessed on 16 February 2021), which is based on meal frequency and portion size. In VHU, food intake level (FIL) was calculated as total energy intake (TEI) divided by estimated basal metabolic rate; individuals with extreme TEI (below the fifth and above the 97.5th percentile of food intake level) were excluded from the analyses [20]. Observations with lifestyle values considered biologically implausible were removed (Supplementary Table S2). Written, informed consent was obtained from all living participants at enrolment into VHU and MDCS. VHU study was approved by the Region Ethical Review Board in Umeå and MDCS by the Ethical Committee at Lund University (LU 51-90).

2.5. Outcome Ascertainment

Data pertaining to medical diagnoses and mortality were retrieved through record linkage from the National Board of Health and Welfare in Sweden until 31 December 2019. Using each participant's unique personal identification number, the following diagnosis 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.

2.6. Statistical Analysis

All numeric predictors were inverse-normalized to correct skewness, and the derived ordinal variables were treated as continuous variables in subsequent analyses. From an environment-wide association study (EWAS) described elsewhere [21], we prioritized (~300) environmental risk factors that were statistically significant at the corrected p-value threshold after multiple testing. We retrieved 167 predictors for BMI, 49 for systolic blood pressure, 47 for diastolic blood pressure, 87 for total cholesterol, 108 for triglycerides, 50 for HDL-C, 21 for LDL-C, 43 for fasting glucose, and 58 for 2 h glucose [22]. Categorical exposure variables with more than two levels were dichotomized into dummy variables. Nutrient data were adjusted for TEI with the residual method [23] to minimize confounding by energy intake and basal energy requirement. We removed correlated (>80%) and zero-variance predictors to minimize the multiple testing burden [24] (Supplementary Tables S3 and S4). For all datasets, we assumed missingness at random [25], and environmental predictor variables with <50% missingness were imputed with the missForest package from R software using a nonparametric approach for mixed data type, to allow a complete case analysis suitable for the random forest algorithm; continuous predictor variables were verified by the mean squared error (MSE) and categorical predictors were verified by the proportion falsely classified (PFC) [26].

We randomly partitioned each dataset into training (50%) and testing (50%) sets to ensure a sufficient number of events per category for the time-to-event analysis in the

testing set. The training set was used to fit quantile regression forest (QRF) models for predictors associated with the cardiometabolic traits, and the testing set was used to predict future intervals. Multicollinearity of the variables within these models was assessed using the variance inflation factor, with variables with values > 10 removed [27]. All models were adjusted for age, age^2, sex, FFQ version, BMI (when not as response variable), follow-up time, and fasting status (for glycemic and lipid models). We utilized QRFs [6], an extension of the supervised machine learning technique *random forest*, which is an ensemble of simultaneous decision trees derived from bootstrapped samples [28]. Furthermore, we set PIs at 90% probability (fifth and 95th quantiles) to minimize false positives ($(1 - \alpha) \times 100\%$). 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 $I(x) = [q\ \alpha/2\ (Y \mid X = x),\ q\ 1 - \alpha/2\ (Y \mid X = x)]$. Thus, the 90% prediction interval for the trait value was estimated using Equation (1).

$$I(x) = [q\ 0.05\ (Y|X = x),\ q\ 0.95\ (Y|X = x)], \quad (1)$$

where, for a given x, the trait response lies within the interval $I(x)$ with high probability. For VHU, on the basis of the obtained PIs per trait, we defined two groups of persistence: those above the 90% PI ('sensitive') and below 90% PI ('resilient'). However, in MDCS, it was not possible to consider two consecutive measures. Instead, QRFs were obtained only for the baseline visit. In addition, when obtaining the quantiles, variable importance was estimated as the percentage in mean square error (%IncMSE), calculated by permuting sample values of the out-of-bag (OOB) in the test dataset, and increase in node purity (incNodepurity), calculated on the basis of the reduction in sum of squared errors for each decision tree; we rank-ordered the most important variable across all models in Supplementary Table S5 and Supplementary Figures S1–S9) [29].

2.7. Predictive Performance

We estimated two CVD risk scores, (i) the Framingham risk score laboratory- and nonlaboratory-based [30], and (ii) the 2013 American College of Cardiology/American Heart Association Task Force [31]. Overall, both algorithms comprise data on age, sex, smoking, diabetes diagnosis, systolic blood pressure and its treatment, total cholesterol, and HDL-C. For the nonlaboratory-based risk model, BMI was used instead of lipids. We further compared the predictive ability (i.e., area under the receiver operating characteristic curve; ROC AUC) of two logistic regression models, one with the generated risk scores and one with risk score plus a variable indicating risk factor 'sensitivity' (Supplementary Table S6).

2.8. Time-to-Event Analysis

Cox proportional hazards regression models were used to estimate hazard ratios (HRs) and corresponding 95% confidence intervals (CIs) between sensitivity categories for each cardiometabolic trait derived from the QRF approach and the risk of diabetes and CVD-incidence and mortality. The proportional hazards assumption was tested with Schoenfeld residuals. The 'neutral' category was used as the reference group. Statistical significance (p-value) was set at the 5% level. Per cardiometabolic trait, a model including age and sex (and BMI, where this was not the outcome), fasting status, FFQ version, TEI, educational level (education was previously used as a proxy of socioeconomic status in this population [32]), smoking status, physical activity, and alcohol consumption. The covariates were selected a priori owing to their previously established associations with cardiovascular mortality in the Swedish population [33]; if a covariate was already in the environmental QRF model, it was not included. The timescale was the elapsed time from baseline in years until an event occurred or the study ended, whichever came first. HRs and 95% CIs were pooled for each cardiometabolic trait by sensitivity category to obtain an overall estimate under a random-effects model [34]; heterogeneity was assessed with Cochran's Q statistic [35,36]. All statistical analyses were performed using *R* software version 3.6.1 [37]; statistical packages are listed in Supplementary Table S7.

3. Results

Baseline characteristics for each cohort are shown in Table 1. Median follow-up time (interquartile range (IQR)) for VHU was 9.7 (5.8) years and 21.1 (4.9) years for MDCS. In both cohorts, individuals classified as being 'sensitive' to lifestyle exposures affecting blood pressure and lipids had more cardiovascular events and deaths compared with the remainder of the population (all hazard ratios (HRs) and 95% CIs for CVD events, T2D, and CVD-mortality are in Supplementary Table S8).

Table 1. Baseline characteristics of study cohorts.

	VHU	MDCS
n	35,440	18,067
Male (%)	15,599 (46.8)	6772 (37.5)
Age	42.96 (7.02)	57.72 (7.71)
BMI (kg/m^2)	25.10 (3.71)	25.30 (3.78)
Total cholesterol (mmol/L)	5.47 (1.14)	6.20 (1.11)
HDL-C (mmol/L)	1.32 (0.57)	1.40 (0.37)
LDL-C (mmol/L)	3.92 (1.16)	4.19 (1.02)
Triglycerides (mmol/L)	1.32 (0.76)	1.47 (0.75)
Fasting glucose (mmol/L)	5.31 (0.63)	5.02 (0.83)
2 h glucose (mmol/L)	6.39 (1.30)	-
HbA1c (mmol/mol) [a]	-	31.4 (5.05)
Systolic blood pressure (mm Hg)	123.27 (15.77)	138.58 (18.97)
Diastolic blood pressure (mm Hg)	77.25 (10.86)	84.02 (9.53)

All values are the mean (SD) unless otherwise stated. VHU: Västerbotten intervention program; MDCS: Malmö Diet and Cancer; BMI: body mass index; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; HbA1c: glycated hemoglobin; 2 h glucose: 2 h glucose tolerance. [a] Raw value collected in DCCT (Diabetes Control and Complications Trial) units, transformed to mmol/mol units using formula HbA1c (mmol/mol) = 10.929 × (HbA1c (%) − 2.15) [38]. *Note*: To convert to mg/dL multiply cholesterol by 38.67, blood glucose by 18.0182, and triglycerides by 38.67.

3.1. Cardiovascular Events

In VHU, the risk of CVD in those who were classified as 'sensitive' to the lifestyle exposures affecting diastolic blood pressure was doubled, whereas, in MDCS, the risk in this same subgroup was increased by 32%, compared to the reference group. The risk of nonfatal and fatal CVD in people classified as sensitive to the lifestyle exposures affecting systolic blood pressure was ~60% and ~50% higher than the reference population for MDCS and VHU, respectively. When hazard estimates were pooled, the overall systolic and diastolic blood pressure 'sensitive' HRs were statistically significant under a random-effects model. In addition, the pooled groups of 'sensitive' individuals for systolic and diastolic blood pressure were also at higher risk for early death (Figure 2).

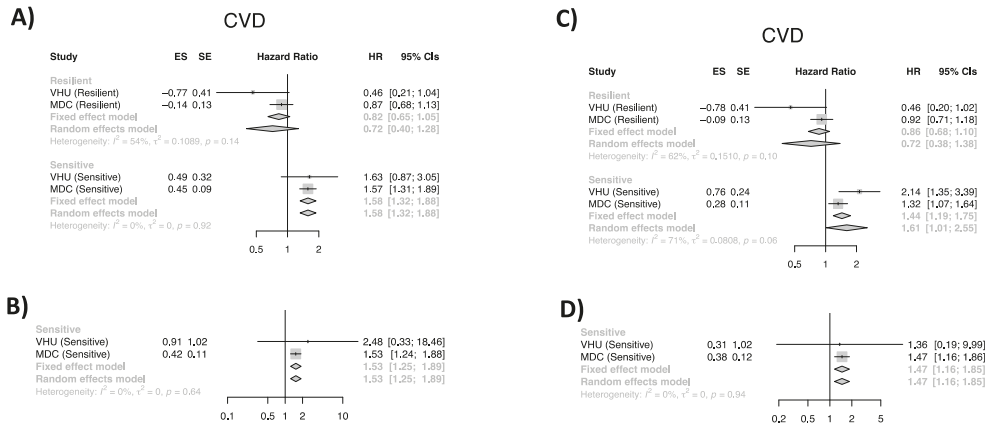


Figure 2. Forest plots of pooled studies by persistence category and CVD event. (**A**,**C**) Systolic blood pressure (SBP (mm/Hg)); (**B**,**D**) diastolic blood pressure (DBP (mm/Hg)). Random- and fixed-effects meta-analysis of the association between trait-persistence category and CVD and CVD mortality. For (**C**,**D**), the number of events did not allow to obtain pooled estimates for the 'resilient' group. 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; CVD: cardiovascular disease.

The risk of CVD in people classified as sensitive to LDL-C-related risk exposures was doubled in VHU and ~60% higher in MDCS, with the pooled estimate being statistically significant. In MDCS, those who were sensitive to lifestyle exposures lowering HDL-C were at higher risk of CVD, but this was not the case in VHU.

3.2. T2D Incidence

For glycemic traits, those classified as 'sensitive' in MDCS to the lifestyle risk factors for elevated fasting glucose had a fourfold increased risk of T2D. However, when risk estimates from MDCS were pooled with those from VHU, this result was not statistically significant (Table 2).

Table 2. Pooled hazard ratios (HR) and 95% CI and outcomes from VHU and MDCS.

	CVD			Test between Groups [a]		T2D			Test between Groups [a]		CVD Mortality			Test between Groups [a]	
Trait	**HR**	**95% (CIs)**		**Q**	***p***	**HR**	**95% (CIs)**		**Q**	***p***	**HR**	**95% (CIs)**		**Q**	***p***
Fasting glucose															
Pooled neutrality	1.00					1.00					1.00				
Pooled resilient	0.77	0.31	1.90	0.25	0.62	0.73	0.46	1.16	0.75	0.39	1.04	0.61	1.75	0.12	0.73
Pooled sensitive	1.01	0.55	1.86			1.69	0.26	10.87			1.18	0.69	2.03		
[b] 2 h Glucose/HbA1c															
Pooled neutrality	1.00					1.00					1.00				
Pooled resilient	0.77	0.54	1.12	5.41	0.02	0.62	0.08	4.55	0.36	0.55	0.75	0.39	1.47	0.74	0.39
Pooled sensitive	1.46	0.99	2.17			1.23	0.46	3.31			1.11	0.62	2.00		
Diastolic blood pressure															
Pooled neutrality	1.00					1.00					1.00				
Pooled resilient	0.72	0.38	1.38	3.88	0.05	0.64	0.26	1.55	0.15	0.70	1.05	0.81	1.37	3.45	0.06
Pooled sensitive	1.61	1.01	2.55			0.81	0.36	1.82			1.47	1.16	1.85		
HDL-C															
Pooled neutrality	1.00					1.00					1.00				
Pooled resilient	1.21	0.50	2.98	0.03	0.87	2.22	0.96	5.12	1.12	0.29	1.39	0.79	2.44	0.01	0.94
Pooled sensitive	1.12	0.67	1.84			0.69	0.10	5.03			1.47	0.38	5.62		
BMI															
Pooled neutrality	1.00					1.00					1.00				
Pooled resilient	1.07	0.84	1.37	0.59	0.44	1.37	0.30	6.24	0.98	0.32	1.57	1.20	2.06	1.70	0.19
Pooled sensitive	0.86	0.51	1.44			0.59	0.31	1.13			1.22	0.93	1.60		
LDL-C															
Pooled neutrality	1.00					1.00					1.00				
Pooled resilient	1.34	0.91	1.98	0.99	0.32	0.59	0.24	1.44	0.03	0.87	1.31	0.80	2.15	0.21	0.65
Pooled sensitive	1.75	1.24	2.46			0.65	0.29	1.48			1.72	0.60	4.97		
Total Cholesterol															
Pooled neutrality	1.00					1.00					1.00				
Pooled resilient	1.17	0.55	2.51	0.32	0.57	1.07	0.62	1.85	0.20	0.66	1.58	0.99	2.53	0.23	0.63
Pooled sensitive	1.58	0.78	3.19			1.30	0.67	2.53			1.25	0.53	2.92		

Table 2. *Cont.*

	CVD			Test between Groups [a]		T2D			Test between Groups [a]		CVD Mortality			Test between Groups [a]	
Trait	**HR**	**95% (CIs)**		**Q**	***p***	**HR**	**95% (CIs)**		**Q**	***p***	**HR**	**95% (CIs)**		**Q**	***p***
Triglycerides															
Pooled neutrality	1.00					1.00					1.00				
Pooled resilient	1.09	0.66	1.78	0.01	0.94	-	-	-	-	-	0.84	0.44	1.59	1.53	0.22
Pooled sensitive	1.06	0.74	1.52			1.04	0.48	2.25			1.39	0.85	2.29		
Systolic blood pressure															
Pooled neutrality	1.00					1.00					1.00				
Pooled resilient	0.72	0.40	1.28	6.55	0.01	0.74	0.38	1.47	3.17	0.07	1.01	0.77	1.32	5.74	0.02
Pooled sensitive	1.58	1.32	1.88			1.65	0.95	2.84			1.53	1.25	1.89		

[a] Test for subgroup differences between resilient and sensitive groups; [b] VHU; '-' indicates that it was not possible to estimate the number. Pooled estimates were obtained with inverse variance method and DerSimonian–Laird estimator for random-effects models; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; BMI: body mass index; HbA1c: glycated hemoglobin CVD: cardiovascular disease. T2D: type 2 diabetes. Adjustment for each cohort model included age, sex, BMI, fasting status, FFQ version, TEI, educational level, smoking status, physical activity, and alcohol intake.

4. Discussion

Overall, a 50% to 60% higher risk of CVD and fatal CVD was observed in those individuals sensitive to the environments associated with blood pressure traits. Similarly, those with sensitivity to the environment related to LDL-C had 74% higher risk of CVD incidence. These findings are in line with others where higher blood pressure and dyslipidemia were shown to be associated with cardiovascular risk [39].

Public health guidelines to reduce disease risk rely on population-averaged estimates of risk factor susceptibility, often focusing on intermediate markers of cardiometabolic risk such as blood pressure or serum cholesterol levels. This strategy assumes that broad recommendations work well for most people, yet risk factor susceptibility and treatment response are highly heterogeneous [40], justifying public health interventions that are tailored to subgroups of the population. To explore whether doing so might be of clinical value, we used machine learning to identify, avoiding distributional assumptions, a population subgroup that is especially sensitive to modifiable lifestyle exposures for cardiometabolic disease. We showed that those who are especially sensitive to these risk exposures tended to develop CVD more rapidly. This type of risk classification is important, as it highlights individuals with 'normal' or 'low' levels of intermediate cardiometabolic markers, who are at relatively high risk of clinical events overlooked by conventional screening and risk classification approaches.

The approach we used focuses on sensitivity to modifiable risk factors trained on intermediate biomarkers of clinical disease. Not all of these intermediate marker sets proved informative. For example, sensitivity to obesogenic lifestyle factors did not raise the risk of T2D or CVD. Indeed, we found no clear evidence that sensitivity to lifestyle exposures in any biomarker set raised the risk of T2D. This may be because diagnosis of T2D is one of exclusion, where all known causes of chronically elevated blood glucose are eliminated, leaving the idiopathic label of T2D to be applied. Thus, T2D is highly heterogeneous in etiology and clinical presentation, making it harder to predict than more precisely defined diagnoses such as CVD. Nevertheless, as the wide confidence intervals around some of the risk estimates reported here indicate, it is likely that these analyses are underpowered, and some negative findings may be false positive.

Although these analyses benefited from comprehensive assessments of lifestyle exposures in these cohorts, a limitation is that they are predominantly self-reported data. Such data are prone to reporting biases, and some lifestyle factors are likely to have been assessed more precisely than others. Moreover, many variables prioritized from VHU were unavailable or captured differently in MDCS, which makes it difficult to isolate biological from statistical heterogeneity when pooled. The observational nature of the studies makes causal inference challenging, and one cannot rule out the possibility that some associations are confounded. There is little one can do to mitigate this common limitation of epidemiological studies. It might also be argued that to be classified as *sensitive* to adverse lifestyle exposures is a function of regression dilution, as this subgroup lies at the extreme of the prediction distributions, where measurement error will be greatest. However, this is unlikely in this setting, as sensitivity to lifestyle exposures persists across many years of follow-up. Nevertheless, trials are needed that assess whether people defined as *sensitive*, yet with apparently healthy biomarker profiles, are more susceptible to cardiovascular events than those who are not defined as *sensitive* and also benefit from intensive lifestyle interventions.

Most current clinical guidelines for T2D and CVD discuss the importance of personalized care, yet include generic lifestyle recommendations [41,42], overlooking between-person 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 risks are minimized [1]. The approach described here is aligned with the objectives of precision prevention, by identifying people at high risk of cardiometabolic disease and helping determine which modifiable exposures to intervene in. Strategies to prevent disease in this subpopulation may include nutritional support [43],

lifestyle modification, and pharmacotherapy [44]; however, further investigation from randomized clinical trials is needed to discern which modality is more appropriate.

5. Conclusions

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 [45].

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/nu14153171/s1, Table S1: VHU Criteria for exclusions on cardiometabolic traits; Table S2: VHU Criteria for implausible values for lifestyle variables; Table S3: Variables removed during data processing; Table S4: VHU variable meaning; Table S5: Rank-ordered most important variables among 9 cardiometabolic traits in VHU; Table S6: AUCs for each trait in VHU; Table S7: R packages used for the analyses in the current study; Table S8: Hazard ratios and 95%CI of prediction interval categories and clinical outcomes; Figure S1: Variable importance plot of fasting glucose (FG) model in VHU per visit; Figure S2: Variable importance plot of 2-hour glucose (2hr G) model in VHU per visit; Figure S3: Variable importance plot of body mass index (BMI) model in VHU per visit; Figure S4: Variable importance plot of Cholesterol (total cholesterol) model in VHU per visit; Figure S5: Variable importance plot of diastolic blood pressure (DBP) model in VHU per visit; Figure S6: Variable importance plot of high-density cholesterol (HDL-C) model in VHU per visit; Figure S7: Variable importance plot of low-density cholesterol (LDL-C) model in VHU per visit; Figure S8: Variable importance plot of systolic blood pressure (SBP) model in VHU per visit; Figure S9: Variable importance plot of triglycerides model in VHU per visit.

Author Contributions: Conceptualization, P.W.F., A.P. and H.P.-M.; formal analysis, investigation, data curation, writing—original draft preparation, and visualization, H.P.-M.; writing—review and editing, N.A.-P., A.P., I.J., J.B., M.O., G.N.G. and P.W.F.; supervision, G.N.G. and P.W.F.; project administration, G.N.G.; funding acquisition, P.W.F. All authors have read and agreed to the published version of the manuscript.

Funding: This project received funding from the Swedish Research Council, Strategic Research Area Exodiab, (Dnr 2009-1039), the Swedish Foundation for Strategic Research (IRC15-0067), the Swedish Research Council, Linnaeus grant (Dnr 349-2006-237), and the European Research Council (CoG-2015_681742_NASCENT). J.B received funding from the Swedish Research Council (Artificially Intelligent use of Registers (AIR Lund), Dnr 2019-00198). N.A.-P. received funding from the Swedish Research Council (Avtals-ID: 2021-06714_3) and the Henning och Johan Throne-Holsts Foundation.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Regional Ethical Review Board of Northern Sweden, Umeå and Ethical Committee at Lund University approval: LU 51-90.

Informed Consent Statement: Written informed consent was obtained from all study participants.

Data Availability Statement: The individual-level data from VHU and MDCS are not publicly available due to privacy and consenting constraints. However, applications for data access can be submitted to the Department of Biobank Research, Umeå University (https://www.umu.se/en/biobank-research-unit/; accessed on 14 February 2020) and Lund university (https://www.malmo-kohorter.lu.se/malmo-cohorts; accessed on 20 September 2021) for the VHU and MDCS cohorts, respectively.

Acknowledgments: We extend our gratitude to all participants involved in the Västerbotten Health Survey and Malmö Diet Cancer study.

Conflicts of Interest: P.W.F. has received research grants from numerous diabetes drug companies and fees as consultant from Novo Nordisk, Lilly, and Zoe Ltd., London, UK; He is currently the Scientific Director in Medical Science at the Novo Nordisk Foundation. Other authors declare no conflict of interests.

References

1. Chung, W.K.; Erion, K.; Florez, J.C.; Hattersley, A.T.; Hivert, M.-F.; Lee, C.G.; McCarthy, M.I.; Nolan, J.J.; Norris, J.M.; Pearson, E.R. Precision medicine in diabetes: A Consensus Report from the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetologia* **2020**, *63*, 1–23. [CrossRef] [PubMed]
2. Zhou, X.; Siegel, K.R.; Ng, B.P.; Jawanda, S.; Proia, K.K.; Zhang, X.; Albright, A.L.; Zhang, P. Cost-effectiveness of diabetes prevention interventions targeting high-risk individuals and whole populations: A systematic review. *Diabetes Care* **2020**, *43*, 1593–1616. [CrossRef]
3. Henley, S.S.; Golden, R.M.; Kashner, T.M. Statistical modeling methods: Challenges and strategies. *Biostat. Epidemiol.* **2020**, *4*, 105–139. [CrossRef]
4. Franks, P.W.; Poveda, A. Lifestyle and precision diabetes medicine: Will genomics help optimise the prediction, prevention and treatment of type 2 diabetes through lifestyle therapy? *Diabetologia* **2017**, *60*, 784–792. [CrossRef] [PubMed]
5. Mone, P.; Gambardella, J.; Lombardi, A.; Pansini, A.; De Gennaro, S.; Leo, A.L.; Famiglietti, M.; Marro, A.; Morgante, M.; Frullone, S.; et al. Correlation of physical and cognitive impairment in diabetic and hypertensive frail older adults. *Cardiovasc. Diabetol.* **2022**, *21*, 10. [CrossRef] [PubMed]
6. Meinshausen, N. Quantile regression forests. *J. Mach. Learn. Res.* **2006**, *7*, 983–999.
7. Hallmans, G.; Agren, A.; Johansson, G.; Johansson, A.; Stegmayr, B.; Jansson, J.H.; Lindahl, B.; Rolandsson, O.; Soderberg, S.; Nilsson, M.; et al. Cardiovascular disease and diabetes in the Northern Sweden Health and Disease Study Cohort-evaluation of risk factors and their interactions. *Scand. J. Public Health Suppl.* **2003**, *61*, 18–24. [CrossRef]
8. Norberg, M.; Wall, S.; Boman, K.; Weinehall, L. The Vasterbotten Intervention Programme: Background, design and implications. *Glob. Health Action* **2010**, *3*, 4643. [CrossRef]
9. Manjer, J.; Carlsson, S.; Elmstahl, S.; Gullberg, B.; Janzon, L.; Lindstrom, M.; Mattisson, I.; Berglund, G. The Malmo Diet and Cancer Study: Representativity, cancer incidence and mortality in participants and non-participants. *Eur. J. Cancer Prev.* **2001**, *10*, 489–499. [CrossRef] [PubMed]
10. Berglund, G.; Elmstahl, S.; Janzon, L.; Larsson, S.A. The Malmo Diet and Cancer Study. Design and feasibility. *J. Intern. Med.* **1993**, *233*, 45–51. [CrossRef]
11. Manjer, J.; Elmståhl, S.; Janzon, L.; Berglund, G. Invitation to a population-based cohort study: Differences between subjects recruited using various strategies. *Scand. J. Public Health* **2002**, *30*, 103–112. [CrossRef]
12. Hedblad, B.; Nilsson, P.; Janzon, L.; Berglund, G. Relation between insulin resistance and carotid intima-media thickness and stenosis in non-diabetic subjects. Results from a cross-sectional study in Malmö, Sweden. *Diabet. Med.* **2000**, *17*, 299–307. [CrossRef] [PubMed]
13. Friedewald, W.T.; Levy, R.I.; Fredrickson, D.S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin. Chem.* **1972**, *18*, 499–502. [CrossRef] [PubMed]
14. Ng, N.; Carlberg, B.; Weinehall, L.; Norberg, M. Trends of blood pressure levels and management in Vasterbotten County, Sweden, during 1990–2010. *Glob. Health Action* **2012**, *5*, 18195. [CrossRef] [PubMed]
15. Wu, J.; Province, M.A.; Coon, H.; Hunt, S.C.; Eckfeldt, J.H.; Arnett, D.K.; Heiss, G.; Lewis, C.E.; Ellison, R.C.; Rao, D.C.; et al. An investigation of the effects of lipid-lowering medications: Genome-wide linkage analysis of lipids in the HyperGEN study. *BMC Genet.* **2007**, *8*, 60. [CrossRef]
16. Tobin, M.D.; Sheehan, N.A.; Scurrah, K.J.; Burton, P.R. Adjusting for treatment effects in studies of quantitative traits: Antihypertensive therapy and systolic blood pressure. *Stat. Med.* **2005**, *24*, 2911–2935. [CrossRef] [PubMed]
17. Hallal, P.C.; Victora, C.G. Reliability and validity of the International Physical Activity Questionnaire (IPAQ). *Med. Sci. Sports Exerc.* **2004**, *36*, 556. [CrossRef]
18. Craig, C.L.; Marshall, A.L.; Sjostrom, M.; Bauman, A.E.; Booth, M.L.; Ainsworth, B.E.; Pratt, M.; Ekelund, U.; Yngve, A.; Sallis, J.F.; et al. International physical activity questionnaire: 12-country reliability and validity. *Med. Sci. Sports Exerc.* **2003**, *35*, 1381–1395. [CrossRef]
19. Johansson, I.; Hallmans, G.; Wikman, A.; Biessy, C.; Riboli, E.; Kaaks, R. Validation and calibration of food-frequency questionnaire measurements in the Northern Sweden Health and Disease cohort. *Public Health Nutr.* **2002**, *5*, 487–496. [CrossRef]
20. Winkvist, A.; Hornell, A.; Hallmans, G.; Lindahl, B.; Weinehall, L.; Johansson, I. More distinct food intake patterns among women than men in northern Sweden: A population-based survey. *Nutr. J.* **2009**, *8*, 12. [CrossRef] [PubMed]
21. Patel, C.J.; Bhattacharya, J.; Butte, A.J. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS ONE* **2010**, *5*, e10746. [CrossRef] [PubMed]
22. Poveda, A.; Pomares-Millan, H.; Chen, Y.; Kurbasic, A.; Patel, C.J.; Renstrom, F.; Hallmans, G.; Johansson, I.; Franks, P.W. Exposome-wide ranking of modifiable risk factors for cardiometabolic disease traits. *Sci. Rep.* **2022**, *12*, 4088. [CrossRef] [PubMed]
23. Willett, W.C.; Howe, G.R.; Kushi, L.H. Adjustment for total energy intake in epidemiologic studies. *Am. J. Clin. Nutr.* **1997**, *65*, 1220S–1228S, discussion 1229S–1231S. [CrossRef] [PubMed]

24. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 26.
25. Zhang, Z. Missing data exploration: Highlighting graphical presentation of missing pattern. *Ann. Transl. Med.* **2015**, *3*, 356. [CrossRef]
26. Stekhoven, D.J.; Buhlmann, P. MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics* **2012**, *28*, 112–118. [CrossRef]
27. Zuur, A.F.; Ieno, E.N.; Elphick, C.S. A protocol for data exploration to avoid common statistical problems. *Methods Ecol. Evol.* **2010**, *1*, 3–14. [CrossRef]
28. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
29. Strobl, C.; Boulesteix, A.L.; Zeileis, A.; Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* **2007**, *8*, 25. [CrossRef]
30. Goff, D.C.; Lloyd-Jones, D.M.; Bennett, G.; Coady, S.; D'agostino, R.B.; Gibbons, R.; Greenland, P.; Lackland, D.T.; Levy, D.; O'donnell, C.J. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J. Am. Coll. Cardiol.* **2014**, *63*, 2935–2959. [CrossRef]
31. D'Agostino Sr, R.B.; Vasan, R.S.; Pencina, M.J.; Wolf, P.A.; Cobain, M.; Massaro, J.M.; Kannel, W.B. General cardiovascular risk profile for use in primary care: The Framingham Heart Study. *Circulation* **2008**, *117*, 743–753. [CrossRef] [PubMed]
32. Norberg, M.; Lundqvist, G.; Nilsson, M.; Gilljam, H.; Weinehall, L. Changing patterns of tobacco use in a middle-aged population: The role of snus, gender, age, and education. *Glob. Health Action* **2011**, *4*, 5613. [CrossRef] [PubMed]
33. Padyab, M.; Blomstedt, Y.; Norberg, M. No association found between cardiovascular mortality, and job demands and decision latitude: Experience from the Vasterbotten Intervention Programme in Sweden. *Soc. Sci. Med.* **2014**, *117*, 58–66. [CrossRef] [PubMed]
34. DerSimonian, R.; Kacker, R. Random-effects model for meta-analysis of clinical trials: An update. *Contemp. Clin. Trials* **2007**, *28*, 105–114. [CrossRef] [PubMed]
35. Borenstein, M.; Hedges, L.V.; Higgins, J.P.; Rothstein, H.R. *Introduction to Meta-Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
36. Borenstein, M.; Higgins, J.P. Meta-analysis and subgroups. *Prev. Sci.* **2013**, *14*, 134–143. [CrossRef] [PubMed]
37. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017; Volume 888. Available online: https://www.R-project.org/ (accessed on 16 February 2022).
38. Nathan, D.M.; Kuenen, J.; Borg, R.; Zheng, H.; Schoenfeld, D.; Heine, R.J.; A1c-Derived Average Glucose Study Group. Translating the A1C assay into estimated average glucose values. *Diabetes Care* **2008**, *31*, 1473–1478. [CrossRef]
39. Huang, Y.; Wang, S.; Cai, X.; Mai, W.; Hu, Y.; Tang, H.; Xu, D. Prehypertension and incidence of cardiovascular disease: A meta-analysis. *BMC Med.* **2013**, *11*, 177. [CrossRef]
40. Gewandter, J.S.; McDermott, M.P.; He, H.; Gao, S.; Cai, X.; Farrar, J.T.; Katz, N.P.; Markman, J.D.; Senn, S.; Turk, D.C. Demonstrating heterogeneity of treatment effects among patients: An overlooked but important step toward precision medicine. *Clin. Pharmacol. Ther.* **2019**, *106*, 204–210. [CrossRef]
41. Arnett, D.K.; Blumenthal, R.S.; Albert, M.A.; Buroker, A.B.; Goldberger, Z.D.; Hahn, E.J.; Himmelfarb, C.D.; Khera, A.; Lloyd-Jones, D.; McEvoy, J.W. 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. *J. Am. Coll. Cardiol.* **2019**, *74*, 1376–1414. [CrossRef]
42. American Diabetes Association. 15. Diabetes Care in the Hospital: Standards of Medical Care in Diabetes—2021. *Diabetes Care* **2021**, *44*, S211–S220. [CrossRef]
43. Bruins, M.J.; Van Dael, P.; Eggersdorfer, M. The Role of Nutrients in Reducing the Risk for Noncommunicable Diseases during Aging. *Nutrients* **2019**, *11*, 85. [CrossRef]
44. Braunwald, E. SGLT2 inhibitors: The statins of the 21st century. *Eur. Heart J.* **2022**, *43*, 1029–1030. [CrossRef] [PubMed]
45. Berry, S.E.; Valdes, A.M.; Drew, D.A.; Asnicar, F.; Mazidi, M.; Wolf, J.; Capdevila, J.; Hadjigeorgiou, G.; Davies, R.; Al Khatib, H. Human postprandial responses to food and potential for precision nutrition. *Nat. Med.* **2020**, *26*, 964–973. [CrossRef] [PubMed]

Paper III

nutrients

MDPI

Article

Estimating the Direct Effect between Dietary Macronutrients and Cardiometabolic Disease, Accounting for Mediation by Adiposity and Physical Activity

Hugo Pomares-Millan [1], Naeimeh Atabaki-Pasdar [1], Daniel Coral [1], Ingegerd Johansson [2], Giuseppe N. Giordano [1] and Paul W. Franks [1,2,3,*]

[1] Lund University Diabetes Centre, Department of Clinical Sciences, Lund University, Skåne University Hospital, 21428 Malmö, Sweden; hugo.pomares-millan@med.lu.se (H.P.-M.); naeimeh.atabaki_pasdar@med.lu.se (N.A.-P.); daniel.coral@med.lu.se (D.C.); giuseppe.giordano@med.lu.se (G.N.G.)
[2] Department of Public Health and Clinical Medicine, Umeå University, 90187 Umeå, Sweden; ingegerd.johansson@umu.se
[3] Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
* Correspondence: paul.franks@med.lu.se; Tel.: +45-20632350

Abstract: Assessing the causal effects of individual dietary macronutrients and cardiometabolic disease is challenging because distinguish direct effects from those mediated or confounded by other factors is difficult. To estimate these effects, intake of protein, carbohydrate, sugar, fat, and its subtypes were obtained using food frequency data derived from a Swedish population-based cohort (n~60,000). Data on clinical outcomes (i.e., type 2 diabetes (T2D) and cardiovascular disease (CVD) incidence) were obtained by linking health registry data. We assessed the magnitude of direct and mediated effects of diet, adiposity and physical activity on T2D and CVD using structural equation modelling (SEM). To strengthen causal inference, we used Mendelian randomization (MR) to model macronutrient intake exposures against clinical outcomes. We identified likely causal effects of genetically predicted carbohydrate intake (including sugar intake) and T2D, independent of adiposity and physical activity. Pairwise, serial- and parallel-mediational configurations yielded similar results. In the integrative genomic analyses, the candidate causal variant localized to the established T2D gene *TCF7L2*. These findings may be informative when considering which dietary modifications included in nutritional guidelines are most likely to elicit health-promoting effects.

Keywords: macronutrient intake; mediation; causal inference; cardiometabolic risk; cardiovascular disease; adiposity; physical activity

Citation: Pomares-Millan, H.; Atabaki-Pasdar, N.; Coral, D.; Johansson, I.; Giordano, G.N.; Franks, P.W. Estimating the Direct Effect between Dietary Macronutrients and Cardiometabolic Disease, Accounting for Mediation by Adiposity and Physical Activity. *Nutrients* **2022**, *14*, 1218. https://doi.org/10.3390/nu14061218

Academic Editor: Shanon L. Casperson

Received: 9 February 2022
Accepted: 9 March 2022
Published: 13 March 2022

1. Introduction

Global patterns of food consumption and energy expenditure have changed drastically in recent decades. Increased sedentary behavior, coupled with the availability of cheap, energy-dense foods, has led to the rapid rise in overweight and obesity worldwide [1]. Excess weight (i.e., body mass index (BMI) > 25 kg/m^2) is one precursor to type 2 diabetes (T2D) and cardiovascular disease (CVD). Hence, an imbalance between energy intake, physical activity and lifestyle behaviors has a major impact on BMI, CVD and T2D risk. Indeed, the Global Burden of Disease Study 2017 reported that dietary risk accounted for 11 million deaths and 255 million disability-adjusted life year (DALYs) in adults [2].

Recent studies have revealed genetic variants associated with food preferences, dietary patterns and food intake [3–7]. Among those macronutrients ingested, total or specific fats and carbohydrates have been associated with obesity, CVD and T2D, yet controversy remains about whether it is energy density that mediates such associations or if single nutrients (e.g., saturated fat, fructose) increase risk of disease [8]. Clinical trials have

indicated that macronutrients might influence glucose metabolism; for example, as part of a lifestyle intervention, a low-energy, low-carbohydrate diet reduced T2D risk [9,10]. Whilst it is plausible that each nutrient could affect disease risk, some might be of greater relevance.

Understanding the causal role of each macronutrient, therefore, could elucidate pathways for more precise dietary intervention strategies [11]. We sought to disentangle the causal role of macronutrients through an integrative analysis using Mendelian randomization (MR) and colocalization obtained through published genome-wide association studies (GWAS) of T2D and CVD. Moreover, we characterize the direct and indirect effects of mediators (i.e., adiposity and physical activity (PA)) on metabolic traits, such as plasma lipids, blood sugar and cardiometabolic disease.

2. Materials and Methods

2.1. Study Design and Population

The Northern Sweden Diet Database (NSDD) contains data from participants collected within the Västerbotten Health Survey (VHU) [12]. Briefly, VHU is an ongoing, prospective, population-based cohort study started in 1985, where adult residents in the county of Västerbotten in Northern Sweden have been invited to a health examination at 40, 50 and 60 years of age (<1% of 30-year-olds were included initially, then discontinued). For this study, participants screened between 1991 and 2016 were eligible, as they had undergone an extensive health examination by trained nurses and family physicians at their local primary care center, including anthropometry, blood lipids and glucose levels before and after a 75 g oral glucose load, and completed surveys, i.e., food frequency questionnaire (FFQ), socio-economic and lifestyle conditions. Values outside normal ranges suggested by VHU data managers were considered outliers and excluded (see Tables S1 and S2). The study protocol and data handling procedures were approved by the Regional Ethical Review Board of Northern Sweden, Umeå, and written informed consent was obtained from all study participants.

2.2. Exposure, Mediator and Outcome Measures

Exposure data were derived for participants who completed the FFQ. Two versions were used during the study: a long version (84 items) and a shortened version (64–66 items). The FFQs have been validated against repeated 24 h dietary records and/or biological markers [13]. Daily energy intake and macronutrient subtypes were calculated for each participant from the food composition database provided by the National Food Agency of Sweden (www.livsmedelsverket.se/en/foodand-content/naringsamnen/livsmedelsdatabasen/; accessed 25 June 2021). This included proteins (animal- and plant-based), carbohydrates and added sugar, the latter being estimated by adding all sucrose and monosaccharides intake minus sugars from fruits and vegetables. Total sugar was further calculated as the sum of all monosaccharides and disaccharides in diet. Saturated, trans-and total fat were also obtained per participant. The macronutrient percentage of energy intake (E%) was calculated by multiplying intake by the metabolizable energy conversion factors and dividing this by total energy intake (TEI) [14]. Those that reported taking dietary supplements or vitamins in the last 14 days were not included.

Since fats, proteins and carbohydrates are rarely consumed in isolation, we added the micronutrients queried from the FFQ and obtained nutrient patterns through principal component (PC) analysis to represent a comprehensive characterization of diet in a real-world setting.

As mediators, adiposity was defined as body mass index (BMI), calculated as body weight in kg (using a calibrated weighing scale) divided by height in m^2, obtained from participants wearing light clothes and no shoes. For physical activity (PA), we calculated a PA index, ranging from 1 = inactive to 4 = active, as described elsewhere [15]. We further included the 'exercise in leisure time' variable, reported in five different ordered categories ranging from (1 = never exercise to 5 = more than three times/week). Both were treated as continuous in analyses.

The primary outcomes (T2D and CVD), expressed as binary variables, were obtained through record linkage to the health databases of the National Board of Health and Welfare in Sweden (www.socialstyrelsen.se/register; accessed 25 June 2021). Clinical endpoints were retrieved using 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 applied for MI. For stroke cases, ICD-9 codes 430, 431 and 433–436 and ICD-10 codes I60, I61, I63 and I64 were used. Secondary outcomes were lipid traits (i.e., high- and low-density lipoprotein (HDL-C, LDL-C, respectively), total cholesterol (TC) and triglycerides (TG)). Glycemic traits included fasting glucose (FG) and two-hour glucose (2 h glucose). For FG, blood was drawn after overnight or 4 h fasting; for 2 h glucose, a blood sample was drawn two hours after the administration of a 75 g oral glucose load, then measured using a Reflotron bench-top analyzer (Roche Diagnostics Scandinavia AB). HDL-C was only measured in a subgroup of participants (n = 23,581) and LDL-C was obtained using the Friedewald formula [16]. TG and TC levels were analyzed using standardized chemical analysis [12]. Validated conversion equations were used to adjust blood lipid measurements taken before and after September 2009 [17]. For participants on lipid lowering medication, lipid 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), as recommended elsewhere [18].

2.3. Statistical Analysis

The distribution of all continuous explanatory variables was assessed for normality. A constant (0.1) was added to all dietary variables prior to log-transforming to correct skewness. We retrieved complete cases for glycemic (n = 55,613) and lipid models (n = 23,581). Mediation models were employed to decompose total effects into direct and indirect effects [19]. We used structural equation modelling (SEM) to study the extent to which PA and BMI influenced associations between macronutrient intake and changes in T2D and CVD status, as well as lipid and glycemic traits. In mediation analysis, a pathway of relationships between variables (i.e., exposure, mediator and outcome) can be modelled using generalized linear regression equations according to a prespecified configuration [20]; these analyses also allow covariance between variables to be determined (see below). For indirect pathways, the two hypothesized mediators of macronutrient intake (PA and BMI) were fitted into pairwise models (Figure 1A) [21].

Next, we fitted parallel mediation models (i.e., exposure → PA → outcome and exposure → BMI → outcome) (Figure 1B) [20] and, given PA and BMI are often correlated, serial mediation models were also tested (exposure → PA → BMI → outcome in Figure 1C). Estimates and standard errors (SE) were obtained through bootstrapping (5000 draws), as recommended elsewhere [22]. To represent real-world dietary habits, all raw nutrient variables were adjusted for TEI using the residual method [23], then centered and scaled to obtain PCs of dietary patterns.

2.3.1. Mediation Analysis

Overall, the mediation analysis is constructed using three linear equations:

$$Y = i1 + cX + \epsilon 1 \quad (1)$$

$$Y = i2 + c'X + bM + \epsilon 2 \quad (2)$$

$$M = i3 + aX + \epsilon 3 \quad (3)$$

where $i1$, $i2$ and $i3$ are intercepts, Y is the outcome, X is the explanatory variable, M is the mediator and ϵ represents the error term. Thus, under the sequential ignorability assumption [24], the model equation can be expressed as:

$$Y = i2 + bi3 + (c' + ab)\, X + \epsilon 2 + b\epsilon 3 \quad (4)$$

For pathways *a*, *b*, *c*′, the following models were fitted: (i) a linear regression assessing the association between each macronutrient (or PC) and the mediators BMI and/or PA, either in serial or in parallel form (pathway *a*); (ii) a linear or logistic regression between mediators BMI and/or PA and the outcome, adjusted for changes in macronutrient intake (pathway *b*); (iii) linear or logistic regression assessing associations between macronutrient intake (or PC) and outcomes, having adjusted for mediators (pathway *c*′). The indirect effect ($a \times b$) was quantified as the effect of the mediators (BMI and PA), and the total effect by the sum of indirect and direct effects ($c' + ab$ in Equation (4)). To assess multicollinearity between variables, the variance inflation factor (VIF) was calculated (variables > 10 were removed). All models were adjusted for putative confounders for each outcome (i.e., age, sex, education, TEI, portion size of potatoes, meat and vegetables, fiber intake (g/day), and alcohol intake (g/day)). For the CVD composite outcome, we further adjusted for tobacco use. Statistical significance was $p < 0.05$ (two-tailed test); in pairwise analyses, a false discovery rate (FDR) correction was set at $P_{FDR} < 0.05$ under the Benjamini–Hochberg procedure [25].

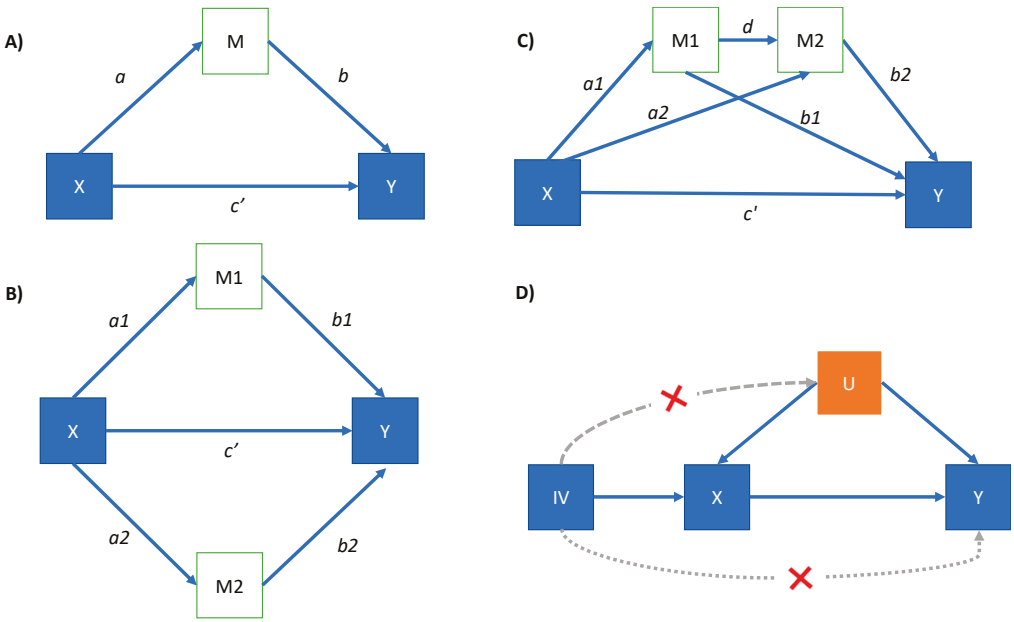


Figure 1. Hypothetical directed acyclic graph models. (**A**) Pairwise mediation model; (**B**) Parallel mediation model; (**C**) Serial mediation model; (**D**) Mendelian randomization model. X: independent variable; M: mediator; Y: outcome; IV: instrumental variable; U: confounding. SEM Pathways: *a* is the coefficient of the effect of X on M; *a*1 and *a*2 are coefficient effects between X and mediators 1 (M1) and 2 (M2), respectively. *b* is the effect of M on Y adjusting for the explanatory variable; *b*1 and *b*2 are coefficient effects between mediators 1 (M1) and 2 (M2), and Y, respectively; *c*′ is the coefficient of the effect of X on Y adjusting for M (direct effect), and *d* is the coefficient effect between mediators. For (**D**) in MR, the IV must not be related to confounders (dotted line) of the exposure–outcome association and affect the outcome only via the exposure and not through another via (dotted lines).

2.3.2. Two-Sample Mendelian Randomization and Bayesian Colocalization

Genetic variants, used here as instrumental variables (IVs) for dietary intake, are randomly assorted during conception [26] and, thus, can be employed for causal inference. For IVs to be valid, they should be associated with the exposure, unrelated to confounders of the exposure–outcome association; they should also affect the outcome only via the exposure (Figure 1D). We assessed the causal impact of dietary carbohydrates, sugars, fat and protein intake with glycemic and lipid traits, T2D and CVD (i.e., stroke and CHD), in a two-sample MR framework (2SMR). The SNPs for exposure data were retrieved from public GWAS summary data from Meddens et al. [5], which were derived from the Social Science Genetic Association Consortium (SSGAC) in 268,922 European ancestry participants. A more detailed description of the dataset is available in their website (https://www.thessgac.org/data; accessed 1 July 2021). Briefly, all dietary intake data were obtained through self-reported food frequency questionnaires and single 24 h diet recalls (only for UK Biobank), and macronutrients were reported as % of energy intake (E%). Owing to the low number of GWAS-significant SNPs in the exposures (6 for fat, 7 for protein, 13 and 10 for carbohydrate and sugar intake, respectively), we relaxed the GWAS threshold to p-value $< 5 \times 10^{-6}$. Further, proxies were used if genetic variants were in linkage disequilibrium (LD) at $r^2 \geq 0.8$ in any of the two-samples. To minimize correlations between the IVs, we performed LD-clumping (where SNPs with lowest p-value are retained) restricted to $r^2 < 0.2$ in a 1000 kb window for the final sets. To disentangle the effect of carbohydrates from sugar (considered a subcomponent in the original GWAS [5]), we combined the significant sugar- and carbohydrate-associated SNPs (n = 79) at the set threshold (p-value $< 5 \times 10^{-6}$). Those overlapping (n = 4) were removed to avoid pleiotropy. To construct the IVs for the outcome variables, we used GWAS available in European ancestry populations. CAD GWAS summary statistics were derived from the Coronary Artery Disease (C4D) Genetics consortium (CARDIoGRAMplusC4D) [27], which included 60,801 cases of CAD and 123,504 controls. For stroke, summary statistics were obtained from the MEGASTROKE consortium, which includes 40,585 cases and 406,111 controls [28]. For T2D, we obtained the unadjusted and BMI-adjusted summary statistics, which include 48,286 cases and 250,671 controls from the DIAGRAM consortium [29]. We used data derived from the MAGIC consortium for fasting [30] and 2 h glucose [31]. For lipid traits, we used data derived from a recent secondary analysis in UK Biobank for TG, HDL-C, and LDL-C [32]. For TC, we used data from a recent GWAS [33]. Characteristics of all GWAS utilized in this study are in Table S3.

We used the inverse variance weighted (IVW) method for our main analysis to estimate the effects of the IVs. Moreover, we used MR-Egger and weighted median estimators to address consistency. As the number of instruments was expected to be low, we used the median F-statistic to measure the IV strength. We further employed the robust adjusted profile score (MR-RAPS) method, by weighing each variant for the effect and precision of the SNP-exposure association, as recommended when using weaker instruments (i.e., below the conventional GWAS threshold [34]). To quantify heterogeneity, bias from horizontal pleiotropy and outliers, we estimated the Cochran's Q statistic for MR-Egger and IVW, and the MR Pleiotropy Residual Sum and Outlier (MR-PRESSO) global test at p level of >0.05 [35]. Exposure and outcome data were harmonized to ensure alleles were aligned, with ambiguous and/or palindromic variants being removed. In addition, we estimated the potential of sample overlap according to Burgess et al. [36] (Table S22). We also performed a leave-one-out sensitivity analysis to assess the impact of each SNP (Figure S5). To identify shared causal pathways among traits, we employed the Hypothesis Prioritization for multi-trait Colocalization (HyPrColoc) algorithm [37], which identifies genome-wide regions with evidence of shared variants (putative of a causal pathway) across traits (Figure S5). All statistical analyses were performed with R version 3.6.2. Mediation analyses were performed with the 'mediation' [21] and 'lavaan' R packages [38]. Two-sample MR analysis was conducted using 'TwoSampleMR' [39] and 'MendelianRandomization' [40].

Colocalization was performed with the 'HyPrColoc' [37] and 'coloc' R packages [41], and PC analysis was visualized with 'PCATools', 'ComplexHeatmap'.

3. Results

Data from a total of 63,862 participants were analyzed. The mean (SD) age of the cohort was 46.5 (8.37) years and 50.3% were female. The means (SD) of glycemic and lipid traits were FG 5.44 (0.93) mmol/L; 2 h glucose 6.55 (1.53) mmol/L; TC 5.39 (1.09) mmol/L, LDL-C 3.59 (1.06) mmol/L and HDL-C 1.37 (0.46) mmol/L, and the median TG was 1.40 (0.81) mmol/L (see Table S4). Genetic correlations were computed using LD Score Regression [42] for traits for which GWAS summary statistics were available, and Pearson's pairwise correlations among mediators and outcomes are shown in Figure S1. For PC analysis, we selected the top three PCs that explained >52% of the total variance (Figure S2) to maintain distinctive dietary patterns. The ten variables contributing the most to the top three PCs are plotted in Figure S3. From these, 'polyunsaturated fat' and 'total fat' were observed in PC1 and PC3. The variable with the largest loading value for PC 1 was 'fiber', for PC 2 it was 'sucrose' and for PC 3 'polyunsaturated fat'. The correlation among traits, nutrients and PCs are shown in Figure 2.

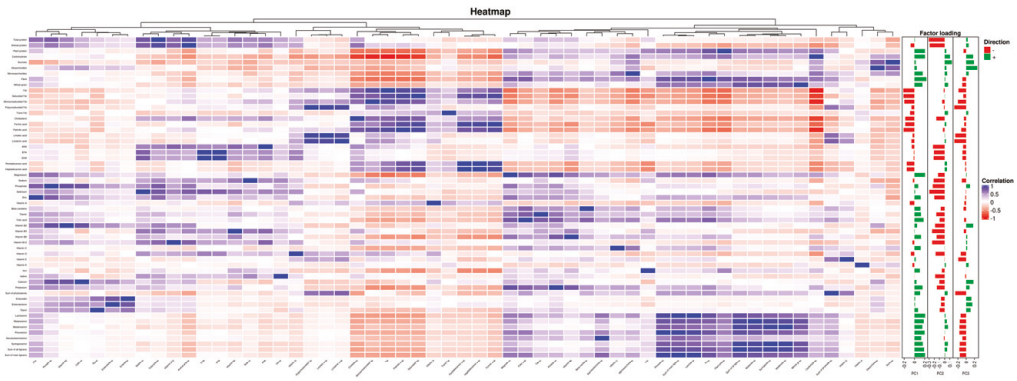


Figure 2. Heatmap of FFQ with 57 items and 3 PCA factor loadings. Correlation key: blue represents positive Pearson's correlations and red represents negative Pearson's correlations. Direction key: red represents a negative direction and green represents a positive direction, large loadings (bars) mean that a variable has greater effects on the principal component.

3.1. Mediation Analysis

The direct and indirect effects for each macronutrient (or PC)–mediator associations are depicted in Figure 3 and summarized in Tables S5–S12. In parallel and serial mediation models, given that we were mainly interested in the direct effect of our exposures, we compared partially and fully mediated nested models (i.e., Figure 1B,C with and without pathway c', respectively) using the chi-squared difference test [43]. The bootstrapped direct and indirect effect estimates, standard errors, and fit indices for parallel and serial mediation models are summarized in Tables S13–S20.

For those macronutrients that remained significant after correction ($P_{FDR} < 0.05$) with glycaemic traits, i.e., FG, we identified nine direct effects (Table S5)—these included added sugar, total sugar, trans-fat, total carbohydrates with positive direction, and with negative effects—saturated and total fat. For 2 h glucose, negative direct effects were observed for saturated, trans-, and total fat (Table S6). Moreover, either in serial or in parallel form, the fully mediated models were not statistically different from the partially mediated model (Tables S13 and S14).

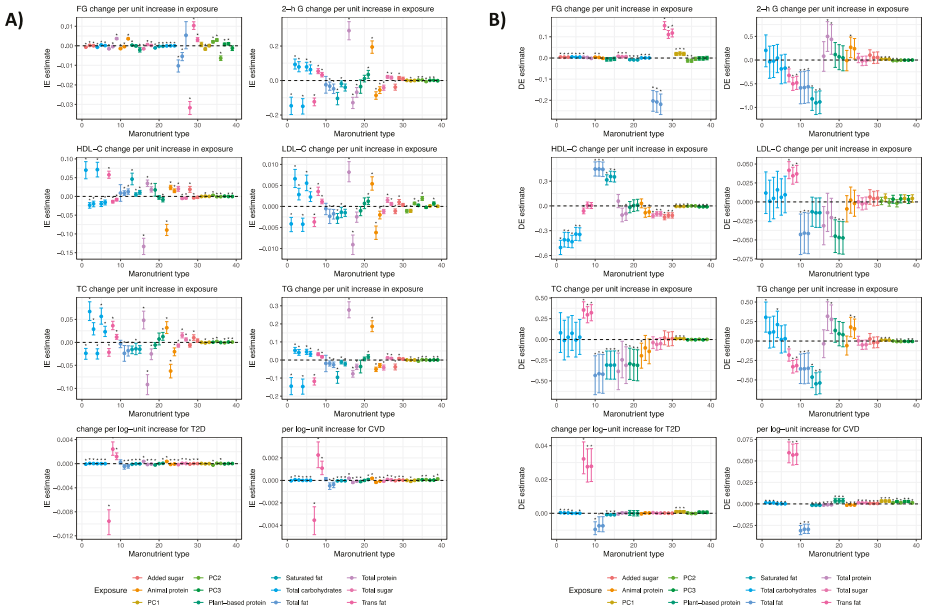


Figure 3. Direct and indirect estimates between macronutrients and outcome in pairwise mediation analysis with body mass index, 5-level physical activity and physical activity index as mediators. Macronutrients are organized on the x-axis in colour codes, ordered consecutively (from left to right) for body mass index, 5-level physical activity and physical activity index. Data are presented as (**A**) indirect and (**B**) direct estimates and 95% confidence intervals; Indirect effect is the estimated average increase in the dependent variable as a result of the mediators; (*) significant after FDR correction at $p < 0.05$; HDL-C, LDL-C: high- and low-density lipoprotein, respectively; TC: total cholesterol; TG: triglycerides; FG: fasting glucose; 2-h G: two-hour glucose; Units: FG mmol/L; 2-h G mmol/L; TC mmol/L; LDL-C mmol/L; HDL-C mmol/L; TG mmol/L; For T2D and CVD, the unit increase corresponds to the probability.

With respect to lipids, there were four direct effects for HDL-C, these consisted of total carbohydrates, added and total sugar with negative direction, and total fat with positive effects; all macronutrients in their fully mediated models were statistically different from the partially mediated model, favoring the latter. Three direct effects from total fat and plant-based protein (negative) and trans-fat (positive) were observed for LDL-C; For TC, plant-based proteins and total fat (negative), trans-and saturated fat (positive) had evidence of direct effect. Only total fat and its subtypes had negative direct effects on TG (Tables S7–S10 and S15–S18).

For T2D, total carbohydrates and trans-fat had positive significant effects, whilst saturated and total fat had an opposite effect; the partially mediated models were significantly different from the fully mediated models, favoring the former (Tables S11 and S19). With respect to CVD, total protein intake was the only macronutrient without significant direct and total effects, irrespective of mediational configuration (Tables S12 and S20).

3.2. MR Causal Effects

In MR analyses, carbohydrate intake was associated with T2D per E% unit increase: OR_{IVW} 0.1 (95% CI: 0.013, 0.71; $p = 0.02$); however, the MR-Egger estimate was not significant, yet when using T2D adjusted for BMI (T2DadjBMI), the effect decreased to OR_{IVW}

0.47 (95% CI: 0.3, 0.75; $p = 0.001$) with $\beta_{MR\text{-}RAPS}$ −0.82 (se 0.3; $p = 0.004$) and no evidence of pleiotropy $P_{MR\text{-}PRESSO} = 0.43$ (Figure 4 and Table 1).

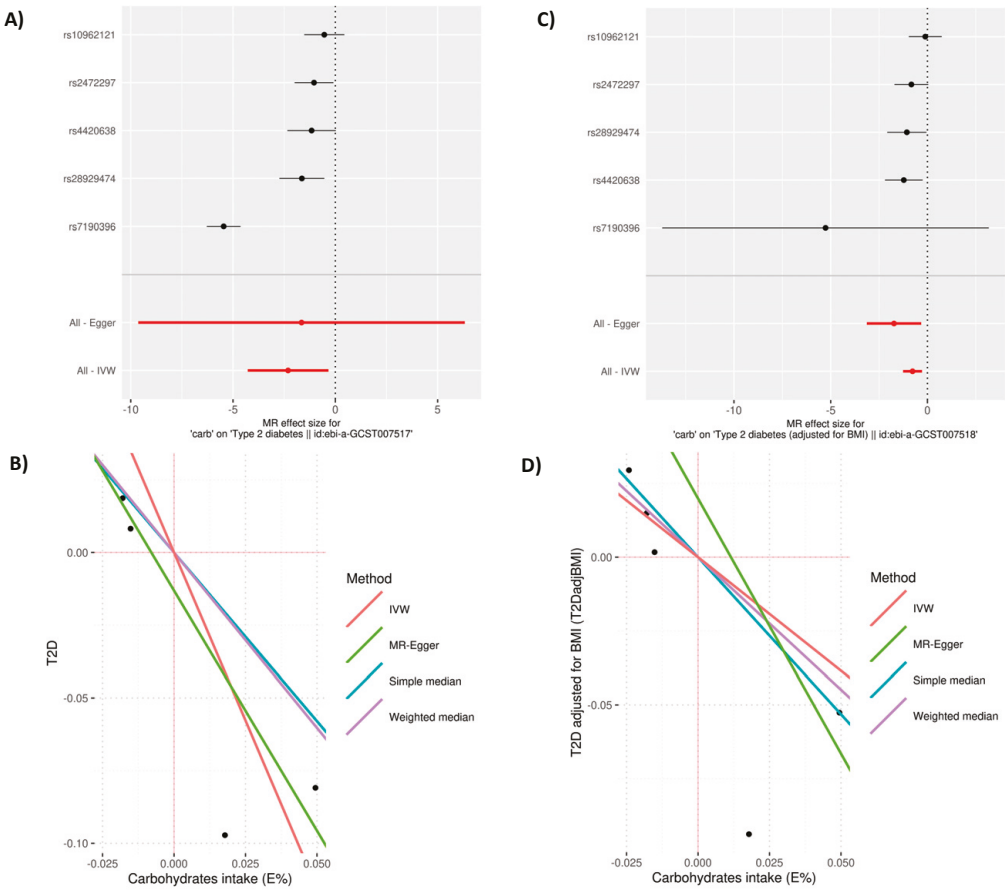


Figure 4. Forest plot of 5-SNP instrument and scatter plot of SNP effects on exposures versus outcomes using different MR methods. For the forest plot, effect size and 95% confidence intervals (standard deviation (SD) change) of the impact of carbohydrates intake SNPs. (**A**,**B**) correspond to carbohydrates (E%)→T2D; (**C**,**D**) correspond to carbohydrates (E%)→T2D adjusted for BMI (T2DadjBMI).

Table 1. Two-sample MR exposure–outcome associations per macronutrient type.

			IVW						MR-Egger						MR-PRESSO		MR-RAPS		
Exposure	**Outcome**	**Number of SNPs**	**F**	**β**	**95% CI**		***p*-Value**	**Q Statistic**	***p*-Value**	**β**	**95% CI**		***p*-Value**	**Q Statistic**	***p*-Value**	**Global Test *p*-Value**	**Distortion Test *p*-Value**	**β**	**β SE**	***p*-Value**
Sugar	FG	26	5	−0.09	−0.18	0.01	0.07	24.48	0.18	−0.11	−0.5	0.29	0.6	24.39	0.14	0.17	-	−0.08	0.05	0.15
	2 h glucose	24	5	−0.07	−0.54	0.41	0.79	16.02	0.85	−0.93	−2.9	1.05	0.36	15.24	0.85	-	-	−0.04	0.25	0.86
	HDL-C	40	4	−0.05	−0.25	0.16	0.66	1265.66	3.62×10^{-240}	−0.21	−0.89	0.48	0.55	1253.76	2.02×10^{-238}	$<1 \times 10^{-4}$	0.6	−0.04	0.06	0.49
	LDL-C	40	4	0.43	0.05	0.82	0.03	3787.17	-	1.06	−0.21	2.33	0.1	3695.86	-	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$	0.1	0.08	0.16
	TC	40	4	0.32	0.004	0.64	0.05	673.38	1.29×10^{-116}	0.88	−0.17	1.91	0.1	657.34	6.04×10^{-114}	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$	0.12	0.09	0.17
	TG	40	4	0.17	0.02	0.32	0.03	619.14	1.65×10^{-105}	0.32	−0.17	0.82	0.2	611	1.87×10^{-104}	$<1 \times 10^{-4}$	9.00×10^{-4}	0.05	0.05	0.28
	T2D	2	7	0.04	0.002	0.84	0.04	22.32	2.30×10^{-6}	-	-	-	-	-	-	-	-	−2.42	1.3	0.06
	Stroke	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	* T2D	2	6	3.9	0.02	969.07	0.02	0.24	0.63	-	-	-	-	-	-	-	-	1.36	2.91	0.64
	CHD	40	4	1.15	0.85	1.56	0.47	89.69	7.20×10^{-6}	2.02	0.64	6.32	0.32	87.4	9.20×10^{-6}	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$	0.09	0.14	0.51
Fat	FG	22	5	0.02	−0.07	0.11	0.68	28.16	0.14	−0.23	−0.49	0.04	0.09	24.9	0.21	0.15	-	0.01	0.06	0.92
	2h glucose	22	5	0.09	−0.43	0.62	0.73	18.73	0.6	−0.31	−2.27	1.65	0.76	18.37	0.56	0.62	-	0.17	0.29	0.56
	HDL-C	34	5	−0.16	−0.34	0.02	0.09	696.01	3.65×10^{-125}	−0.07	−0.53	0.38	0.75	691.01	8.50×10^{-125}	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$	−0.01	0.05	0.76
	LDL-C	34	5	−0.38	−0.79	0.04	0.08	3012.02	-	−0.82	−1.84	0.19	0.11	2940.35	-	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$	−0.19	0.09	0.05
	TC	34	5	−0.28	−0.62	0.06	0.11	550.56	3.74×10^{-95}	−0.62	−1.44	0.21	0.15	538.59	2.56×10^{-93}	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$	−0.16	0.1	0.11
	TG	34	5	0.11	−0.22	0.43	0.51	2018.29	-	−0.13	−0.92	0.67	0.75	1995.28	-	$<1 \times 10^{-4}$	3.00×10^{-4}	0.04	0.06	0.57
	T2D	5	13	2.91	0.47	17.81	0.25	90	-	0.05	1.17×10^{-4}	22.07	0.34	55.78	-	2.00×10^{-4}	$<1 \times 10^{-4}$	−0.06	0.67	0.93
	** Stroke	1	-	0.92	0.52	1.63	0.78	-	-	-	-	-	-	-	-	-	-	−0.08	0.3	0.79
	* T2D	5	13	0.94	0.59	1.51	0.81	5.11	0.28	0.77	0.1	5.83	0.82	5.04	0.17	0.55	-	−0.06	0.22	0.77
	CHD	31	5	0.81	0.58	1.12	0.29	81.71	1.10×10^{-6}	0.69	0.36	1.32	0.21	80.83	9.00×10^{-7}	$<1 \times 10^{-4}$	0.82	−0.21	0.14	0.12
Carbohydrates	FG	28	5	−0.07	−0.17	0.03	0.16	44.25	0.02	−0.12	−0.59	0.35	0.61	44.16	0.01	0.02	-	−0.13	0.06	0.02
	2h glucose	31	5	−0.08	−0.6	0.44	0.76	36.6	0.16	−0.16	−2.83	2.5	0.91	36.58	0.13	0.16	-	−0.09	0.27	0.74
	HDL-C	44	4	−0.12	−0.32	0.09	0.27	1272.13	1.59×10^{-238}	−0.35	−0.98	0.28	0.28	1254.88	1.22×10^{-235}	$<1 \times 10^{-4}$	0.7563	−0.13	0.05	0.02
	LDL-C	44	4	0.44	0.05	0.82	0.03	3784.41	-	1	−0.17	2.18	0.1	3698.32	-	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$	0.08	0.07	0.25
	TC	45	4	0.33	0.03	0.63	0.03	652	3.30×10^{-109}	0.76	−0.19	1.7	0.12	638.86	3.95×10^{-107}	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$	0.11	0.08	0.16
	TG	44	4	0.19	0.03	0.34	0.02	663.34	4.11×10^{-112}	0.37	−0.1	0.84	0.13	653.47	1.06×10^{-110}	$<1 \times 10^{-4}$	0.1338	0.15	0.02	0
	T2D	6	5	0.1	0.01	0.71	0.02	80.3	-	0.19	6.56×10^{-5}	560.9	0.69	79.55	-	2.00×10^{-4}	$<1 \times 10^{-4}$	−1.68	0.63	0.01
	Stroke	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	* T2D	5	5	0.47	0.3	0.75	0.001	3.51	0.48	0.18	0.04	0.72	0.02	1.42	0.7	0.4343	-	−0.82	0.29	0.004
	CHD	44	4	1.23	0.92	1.64	0.2	95.94	6.50×10^{-6}	1.12	0.44	2.87	0.76	95.85	4.30×10^{-6}	$<1 \times 10^{-4}$	0.2603	0.17	0.12	0.16
Proteins	FG	24	5	−0.12	−0.32	0.09	0.26	156.55	-	0.24	−0.44	0.92	0.49	148.81	-	$<1 \times 10^{-4}$	0.8797	−0.1	0.08	0.24
	2h glucose	24	5	0.14	−0.58	0.86	0.7	52.9	3.78×10^{-4}	−0.98	-3.43	1.47	0.43	51.59	3.56×10^{-4}	3.00×10^{-4}	0.0805	−0.07	0.32	0.83
	HDL-C	38	5	−0.18	−0.34	−0.01	0.03	656.82	1.81×10^{-114}	−0.28	−0.7	0.15	0.2	653.96	1.63×10^{-114}	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$	−0.07	0.05	0.14
	LDL-C	38	5	−0.19	−0.36	−0.03	0.02	564.08	1.75×10^{-95}	−0.47	−0.9	−0.05	0.03	540.64	2.63×10^{-91}	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$	0.1	0.06	0.09
	TC	38	5	−0.19	−0.41	0.03	0.09	275.49	8.63×10^{-38}	−0.53	−1.06	0.01	0.05	262.63	8.45×10^{-36}	$<1 \times 10^{-4}$	0.0876	−0.14	0.08	0.09
	TG	38	5	0.04	−0.26	0.34	0.78	1927.21	-	−0.37	−1.13	0.38	0.33	1993.84	-	$<1 \times 10^{-4}$	0.023	−0.07	0.06	0.24
	T2D	4	6	1.78	0.03	105.08	0.78	219.15	-	0.01	4.81×10^{-27}	1.57×10^{22}	0.88	215.3	-	$<1 \times 10^{-4}$	-	−0.76	1.63	0.64
	Stroke	38	5	0.93	0.72	1.19	0.68	58.95	0.01	0.84	0.43	1.66	0.51	58.79	0.01	0.01	0.13	−0.03	0.12	0.81
	* T2D	4	6	0.61	0.07	5.38	0.66	87.4	-	9.77	1.19×10^{-14}	7.99×10^{15}	0.91	86.4	-	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$	−0.44	1.07	0.68
	CHD	38	5	1.09	0.83	1.43	0.46	66.19	2.20×10^{-3}	0.95	0.46	1.96	0.82			1.90×10^{-3}	-	0.07	0.11	0.53

For T2D, Stroke, * T2D and CHD outcomes the effect estimate correspond to Odds ratio (OR); * adjusted for BMI; ** Wald ratio method for single SNP; (-) Not possible to estimate; We considered significant if the directions of the estimates by IVW, weighted median (Table S21) and MR-Egger were directionally consistent with $p < 0.05$, and no significant evidence of pleiotropy tested by MR-PRESSO ($p > 0.05$). F statistics (median) for the strength of correlation between instrument and exposure. IVW: inverse variance weighted; MR-RAPS: Robust adjusted profile score; MR-PRESSO: Pleiotropy residual sum and outlier.T2D: Type 2 diabetes; CHD: Coronary heart disease; FG: fasting glucose; 2 h glucose: two-hour glucose; HDL-C: high-density lipoprotein; LDL-C: low-density lipoprotein; TG: triglycerides; TC: total cholesterol. F-statistic corresponds to the median.

Regarding the effect of carbohydrate intake on lipid levels, TC, LDL-C and TG per E% unit change 9β_{IVW} 0.32 (95% CI: 0.02, 0.63; $p = 0.03$, β_{IVW} 0.44 (95% CI: 0.05, 0.82; $p = 0.03$), and β_{IVW} 0.1 (95% CI: 0.01, 0.2; $p = 0.03$), respectively), yet there was evidence of pleiotropy. For the carbohydrate adjusted for sugar intake instrument (6 SNPs instrumentalized) per E% unit change and T2D, the effect estimate was β_{IVW} 0.09 (95% CI: −7.7, 7.9; $p = 0.9$), and not significant MR-Egger and MR-RAPS models (Tables S26–S28). Moreover, for fat when undertaking MR-Egger, there were no significant associations with any outcome (Table S21).

4. Discussion

We report a comprehensive analysis investigating mediational and causal effects of macronutrient intake and cardiometabolic traits and diseases in >60,000 Swedish participants. To our knowledge, this is the first study reporting the likely causal role of macronutrient intake and the risk of cardiometabolic disease, triangulating evidence from observational and genetic studies. Implications of our findings indicate carbohydrate intake (with predominance of fiber) is likely followed by reduction in T2D risk. By contrast, sugar intake likely raises T2D risk. Due to the modest magnitude of observed effects, it is unlikely to prove a useful target when intervening only through diet for disease prevention. These findings reinforce the notion that complex carbohydrates may be recommended in dietary modifications, alongside other lifestyle changes, to lower individuals' risk of T2D.

The apparent protective effects of dietary carbohydrates in T2D suggests that the quality of carbohydrate is key in T2D prevention. Previous observational studies indicate that associations with T2D can vary according to the carbohydrate type [44], i.e., fiber (sourced from fruits, vegetables or cereals) had a protective effect [45], whereas starch had deleterious effects [46]. In our MR analyses, it was not possible to interrogate carbohydrate or sugar subtypes. Mechanistic studies show that carbohydrate metabolism is heavily dependent on insulin action. However, the fiber effect is believed to be secondary to the transformation to β-glucans, a water-soluble gel-forming substance that decreases surface of exposure in the small intestine, delaying the gut absorption of glucose and reducing postprandial plasma glucose [47]. Moreover, dietary fiber has been associated with lower energy intake and increased satiety [48]. The most probable causal locus, *TCF7L2*, is an established T2D-associated gene [49] which appears to interact with intake of dietary fiber [50], fat [51] and whole grains [52]. Nevertheless, *TCF7L2*'s mechanisms of action, especially in the context of interactions with dietary factors, remains poorly defined. Recent evidence suggests a key role of glucagon-like peptide 1 (GLP-1), secreted after meal ingestion [53], or serotonin [54]. More recent findings from pooled clinical trials in T2D have emphasized the role of gut microbiome in the transformation of fiber-rich foods and glycemic markers [55]. With respect to lipid markers, our observational findings are in line with those reported in previous studies [56], where carbohydrate intake has been linked to LDL-C, HDL-C, TC and TG. Yet, in our MR findings, there was no evidence of causality. For protein intake, studies evaluating protein subtypes have shown a protective effect of plant-based proteins against CVD [57]; conversely, proteins from animal sources increased CVD risk [58]. It was not possible to interrogate protein subtypes with MR; yet this source of heterogeneity may explain the observed pleiotropy.

Our study had limitations. Firstly, although SEM allows direct effect modelling, and despite the multiple configurations explored, our hypothesized models do not cover all possible pathways. Moreover, conditioning on a potential mediator or a shared outcome can induce bias. Secondly, inconsistent mediation (positive direct and negative indirect effects or vice versa) was observed for some of the pairwise associations between the independent and mediating variable, suggesting the mediator was not a significant predictor of the outcome when including both. Thirdly, in MR analysis, horizontal pleiotropy and population stratification were addressed using conventional statistical solutions, yet bias cannot be completely ruled out given the paucity of variants available to construct the IVs and other genetically driven individual features (e.g., microbiome composition) [59] may influence the

observed associations, moreover, evidence of weak instrument bias may still be present, as indicated by the F-statistics. Fourth, not all macro- and micronutrients (including subtypes) had corresponding genetic instruments; thus, we cannot assess with sufficient granularity the causal effect of single-nutrient intake. Further caveats are that dietary patterns seldomly remain the same over the life course, in contrast to a person's nuclear DNA variation, which is fixed at conception. Moreover, observational FFQ data were self-reported and estimated effects may be larger than those observed in a real-world setting. Thus, we cannot rule out residual confounding. Another consideration is the generalizability of our findings. Given that the populations included for mediation analysis and MR were predominantly of European ancestry, our findings may not generalize to other ethnicities. Nevertheless, consistent findings across and within methods help ensure detected relationships are robust to confounding and bias, thereby minimizing false positive association, and support the contemporary view that carbohydrates play a causal role in T2D beyond PA and adiposity.

5. Conclusions

Our analyses highlight the direct effect of carbohydrate intake in T2D risk, helping to quantify the role of higher-quality carbohydrates (which lower risk). These findings warrant confirmation through clinical trials; however, they may enhance current nutritional guidelines by helping distinguish the dietary factors that are likely to be causal from those that are mostly mediated.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/nu14061218/s1 which contains Figures S1–S5 and Tables S1–S30. *Code availability*: Two-sample MR and colocalization analyses *R* scripts are available in https://github.com/hpomares/, accessed 3 September 2021. (References [29,41,42,49–52,60–64] are cited in the Supplementary Materials)

Author Contributions: Conceptualization, P.W.F. and H.P.-M.; formal analysis; investigation; data curation; writing—original draft preparation; visualization, H.P.-M.; writing—review and editing, N.A.-P., D.C., I.J., G.N.G. and P.W.F.; supervision, G.N.G. and P.W.F.; project administration, G.N.G.; funding acquisition, P.W.F. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from the Swedish Research Council, Strategic Research Area Exodiab, (Dnr 2009-1039), the Swedish Foundation for Strategic Research (IRC15-0067), the Swedish Research Council, Linnaeus grant (Dnr 349-2006-237), and the European Research Council (CoG-2015_681742_NASCENT). N.A.-P., was supported in part by Henning och Johan Throne-Holsts, and Hans Werthe´n Foundations.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki, and approved by the Regional Ethical Review Board of Northern Sweden, Umeå.

Informed Consent Statement: Written informed consent was obtained from all study participants.

Data Availability Statement: The exposures and outcomes GWAS summary statistics are available in: CAD (URL: http://www.cardiogramplusc4d.org/data-downloads/, accessed 1 July 2021) [27]. Stroke (URL: https://megastroke.org/download.html/, accessed 1 July 2021) [28]. T2D (URL: https://www.diagram-consortium.org/downloads.html/, accessed 1 July 2021 [29]). Fasting and 2h glucose (URL: https://www.magicinvestigators.org/downloads/ accessed 1 July 2021 [30,31]). HDL-C, LDL-C, TG and TC (URL: https://gwas.mrcieu.ac.uk/datasets/, accessed 1 July 2021 [39]). The individual level data from VHU are not publicly available due privacy and confidentiality constraints of Swedish regulation, but data are available from the Department of Biobank Research, Umeå University, upon reasonable request.

Acknowledgments: We extend our gratitude to all participants involved in the VHU study.

Conflicts of Interest: P.W.F. 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 no conflict of interests.

References

1. Mozaffarian, D. Diverging global trends in heart disease and type 2 diabetes: The role of carbohydrates and saturated fats. *Lancet Diabetes Endocrinol.* **2015**, *3*, 586–588. [CrossRef]
2. Afshin, A.; Sur, P.J.; Fay, K.A.; Cornaby, L.; Ferrara, G.; Salama, J.S.; Mullany, E.C.; Abate, K.H.; Abbafati, C.; Abebe, Z.; et al. Health effects of dietary risks in 195 countries, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **2019**, *393*, 1958–1972. [CrossRef]
3. Cornelis, M.C.; Flint, A.; Field, A.E.; Kraft, P.; Han, J.; Rimm, E.B.; van Dam, R.M. A genome-wide investigation of food addiction. *Obesity* **2016**, *24*, 1336–1341. [CrossRef] [PubMed]
4. McRae, J.F.; Jaeger, S.R.; Bava, C.M.; Beresford, M.K.; Hunter, D.; Jia, Y.; Chheang, S.L.; Jin, D.; Peng, M.; Gamble, J.C.; et al. Identification of regions associated with variation in sensitivity to food-related odors in the human genome. *Curr. Biol.* **2013**, *23*, 1596–1600. [CrossRef] [PubMed]
5. Meddens, S.F.W.; de Vlaming, R.; Bowers, P.; Burik, C.A.P.; Linner, R.K.; Lee, C.; Okbay, A.; Turley, P.; Rietveld, C.A.; Fontana, M.A.; et al. Genomic analysis of diet composition finds novel loci and associations with health and lifestyle. *Mol. Psychiatry* **2021**, *26*, 2056–2069. [CrossRef] [PubMed]
6. Hwang, L.D.; Lin, C.; Gharahkhani, P.; Cuellar-Partida, G.; Ong, J.S.; An, J.; Gordon, S.D.; Zhu, G.; MacGregor, S.; Lawlor, D.A.; et al. New insight into human sweet taste: A genome-wide association study of the perception and intake of sweet substances. *Am. J. Clin. Nutr.* **2019**, *109*, 1724–1737. [CrossRef] [PubMed]
7. Eriksson, L.; Esberg, A.; Haworth, S.; Holgerson, P.L.; Johansson, I. Allelic Variation in Taste Genes Is Associated with Taste and Diet Preferences and Dental Caries. *Nutrients* **2019**, *11*, 1491. [CrossRef] [PubMed]
8. Stanhope, K.L. Sugar consumption, metabolic disease and obesity: The state of the controversy. *Crit. Rev. Clin. Lab. Sci.* **2016**, *53*, 52–67. [CrossRef] [PubMed]
9. Group, D.P.P.R. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N. Engl. J. Med.* **2002**, *346*, 393–403. [CrossRef]
10. Pan, X.-R.; Li, G.-w.; Hu, Y.-H.; Wang, J.-X.; Yang, W.-Y.; An, Z.-X.; Hu, Z.-X.; Xiao, J.-Z.; Cao, H.-B.; Liu, P.-A.J.D.c. Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance: The Da Qing IGT and Diabetes Study. *Diabetes Care* **1997**, *20*, 537–544. [CrossRef]
11. Wang, D.D.; Hu, F.B. Precision nutrition for prevention and management of type 2 diabetes. *Lancet Diabetes Endocrinol.* **2018**, *6*, 416–426. [CrossRef]
12. Norberg, M.; Wall, S.; Boman, K.; Weinehall, L. The Vasterbotten Intervention Programme: Background, design and implications. *Glob. Health Action* **2010**, *3*, 6343. [CrossRef] [PubMed]
13. Johansson, I.; Hallmans, G.; Wikman, A.; Biessy, C.; Riboli, E.; Kaaks, R. Validation and calibration of food-frequency questionnaire measurements in the Northern Sweden Health and Disease cohort. *Public Health Nutr.* **2002**, *5*, 487–496. [CrossRef]
14. Ramne, S.; Alves Dias, J.; Gonzalez-Padilla, E.; Olsson, K.; Lindahl, B.; Engstrom, G.; Ericson, U.; Johansson, I.; Sonestedt, E. Association between added sugar intake and mortality is nonlinear and dependent on sugar source in 2 Swedish population-based prospective cohorts. *Am. J. Clin. Nutr.* **2019**, *109*, 411–423. [CrossRef] [PubMed]
15. Consortium, I. Validity of a short questionnaire to assess physical activity in 10 European countries. *Eur. J. Epidemiol.* **2012**, *27*, 15–25. [CrossRef]
16. Friedewald, W.T.; Levy, R.I.; Fredrickson, D.S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin. Chem.* **1972**, *18*, 499–502. [CrossRef]
17. Ng, N.; Carlberg, B.; Weinehall, L.; Norberg, M. Trends of blood pressure levels and management in Vasterbotten County, Sweden, during 1990-2010. *Glob. Health Action* **2012**, *5*, 499–502. [CrossRef] [PubMed]
18. Wu, J.; Province, M.A.; Coon, H.; Hunt, S.C.; Eckfeldt, J.H.; Arnett, D.K.; Heiss, G.; Lewis, C.E.; Ellison, R.C.; Rao, D.C.; et al. An investigation of the effects of lipid-lowering medications: Genome-wide linkage analysis of lipids in the HyperGEN study. *BMC Genet.* **2007**, *8*, 60. [CrossRef]
19. Imai, K.; Keele, L.; Tingley, D. A general approach to causal mediation analysis. *Psychol. Methods* **2010**, *15*, 309–334. [CrossRef] [PubMed]
20. Hayes, A.F. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*; Guilford Publications: New York, NY, USA, 2017.
21. Tingley, D.; Yamamoto, T.; Hirose, K.; Keele, L.; Imai, K. Mediation: R Package for Causal Mediation Analysis. *J. Stat. Softw.* **2014**, *59*. [CrossRef]
22. Leth-Steensen, C.; Gallitto, E. Testing Mediation in Structural Equation Modeling: The Effectiveness of the Test of Joint Significance. *Educ. Psychol. Meas.* **2016**, *76*, 339–351. [CrossRef] [PubMed]
23. Willett, W.C.; Howe, G.R.; Kushi, L.H. Adjustment for total energy intake in epidemiologic studies. *Am. J. Clin. Nutr.* **1997**, *65*, 1220S–1228S; discussion 1229S–1231S. [CrossRef] [PubMed]
24. Forastiere, L.; Mattei, A.; Ding, P. Principal ignorability in mediation analysis: Through and beyond sequential ignorability. *Biometrika* **2018**, *105*, 979–986. [CrossRef]
25. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate—A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **1995**, *57*, 289–300. [CrossRef]

26. Davey Smith, G.; Ebrahim, S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **2003**, *32*, 1–22. [CrossRef] [PubMed]
27. Nikpay, M.; Goel, A.; Won, H.H.; Hall, L.M.; Willenborg, C.; Kanoni, S.; Saleheen, D.; Kyriakou, T.; Nelson, C.P.; Hopewell, J.C.; et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **2015**, *47*, 1121–1130. [CrossRef]
28. Malik, R.; Chauhan, G.; Traylor, M.; Sargurupremraj, M.; Okada, Y.; Mishra, A.; Rutten-Jacobs, L.; Giese, A.K.; van der Laan, S.W.; Gretarsdottir, S.; et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat. Genet.* **2018**, *50*, 524–537. [CrossRef]
29. Mahajan, A.; Wessel, J.; Willems, S.M.; Zhao, W.; Robertson, N.R.; Chu, A.Y.; Gan, W.; Kitajima, H.; Taliun, D.; Rayner, N.W.; et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* **2018**, *50*, 559–571. [CrossRef]
30. Manning, A.K.; Hivert, M.F.; Scott, R.A.; Grimsby, J.L.; Bouatia-Naji, N.; Chen, H.; Rybin, D.; Liu, C.T.; Bielak, L.F.; Prokopenko, I.; et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **2012**, *44*, 659–669. [CrossRef]
31. Saxena, R.; Hivert, M.F.; Langenberg, C.; Tanaka, T.; Pankow, J.S.; Vollenweider, P.; Lyssenko, V.; Bouatia-Naji, N.; Dupuis, J.; Jackson, A.U.; et al. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat. Genet.* **2010**, *42*, 142–148. [CrossRef]
32. Richardson, T.G.; Sanderson, E.; Palmer, T.M.; Ala-Korpela, M.; Ference, B.A.; Davey Smith, G.; Holmes, M.V. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. *PLoS Med.* **2020**, *17*, e1003062. [CrossRef]
33. Borges, M.C.; Schmidt, A.F.; Jefferis, B.; Wannamethee, S.G.; Lawlor, D.A.; Kivimaki, M.; Kumari, M.; Gaunt, T.R.; Ben-Shlomo, Y.; Tillin, T.; et al. Circulating Fatty Acids and Risk of Coronary Heart Disease and Stroke: Individual Participant Data Meta-Analysis in Up to 16 126 Participants. *J. Am. Heart Assoc.* **2020**, *9*, e013131. [CrossRef] [PubMed]
34. Bowden, J.; Davey Smith, G.; Burgess, S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **2015**, *44*, 512–525. [CrossRef]
35. Verbanck, M.; Chen, C.-Y.; Neale, B.; Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **2018**, *50*, 693–698. [CrossRef] [PubMed]
36. Burgess, S.; Davies, N.M.; Thompson, S.G. Bias due to participant overlap in two-sample Mendelian randomization. *Genet. Epidemiol.* **2016**, *40*, 597–608. [CrossRef] [PubMed]
37. Foley, C.N.; Staley, J.R.; Breen, P.G.; Sun, B.B.; Kirk, P.D.W.; Burgess, S.; Howson, J.M.M. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat. Commun.* **2021**, *12*, 764. [CrossRef] [PubMed]
38. Rosseel, Y. Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *J. Stat. Softw.* **2012**, *48*, 1–36. [CrossRef]
39. Hemani, G.; Zheng, J.; Elsworth, B.; Wade, K.H.; Haberland, V.; Baird, D.; Laurin, C.; Burgess, S.; Bowden, J.; Langdon, R.; et al. The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **2018**, *7*, e34408. [CrossRef] [PubMed]
40. Yavorska, O.O.; Burgess, S. MendelianRandomization: An R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **2017**, *46*, 1734–1739. [CrossRef]
41. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet* **2020**, *16*, e1008720. [CrossRef] [PubMed]
42. Bulik-Sullivan, B.K.; Loh, P.-R.; Finucane, H.K.; Ripke, S.; Yang, J.; Patterson, N.; Daly, M.J.; Price, A.L.; Neale, B.M. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **2015**, *47*, 291–295. [CrossRef] [PubMed]
43. Pavlov, G.; Shi, D.X.; Maydeu-Olivares, A. Chi-square Difference Tests for Comparing Nested Models: An Evaluation with Non-normal Data. *Struct. Equ. Model.-A Multidiscip. J.* **2020**, *27*, 908–917. [CrossRef]
44. Weickert, M.O.; Pfeiffer, A.F.H. Impact of Dietary Fiber Consumption on Insulin Resistance and the Prevention of Type 2 Diabetes. *J. Nutr.* **2018**, *148*, 7–12. [CrossRef] [PubMed]
45. McRae, M.P. Dietary Fiber Intake and Type 2 Diabetes Mellitus: An Umbrella Review of Meta-analyses. *J. Chiropr. Med.* **2018**, *17*, 44–53. [CrossRef] [PubMed]
46. Reynolds, A.; Mann, J.; Cummings, J.; Winter, N.; Mete, E.; Te Morenga, L. Carbohydrate quality and human health: A series of systematic reviews and meta-analyses. *Lancet* **2019**, *393*, 434–445. [CrossRef]
47. Bernstein, A.M.; Titgemeier, B.; Kirkpatrick, K.; Golubic, M.; Roizen, M.F. Major cereal grain fibers and psyllium in relation to cardiovascular health. *Nutrients* **2013**, *5*, 1471–1487. [CrossRef] [PubMed]
48. Runchey, S.S.; Valsta, L.M.; Schwarz, Y.; Wang, C.; Song, X.; Lampe, J.W.; Neuhouser, M.L. Effect of low- and high-glycemic load on circulating incretins in a randomized clinical trial. *Metabolism* **2013**, *62*, 188–195. [CrossRef]
49. Grant, S.F.; Thorleifsson, G.; Reynisdottir, I.; Benediktsson, R.; Manolescu, A.; Sainz, J.; Helgason, A.; Stefansson, H.; Emilsson, V.; Helgadottir, A.; et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* **2006**, *38*, 320–323. [CrossRef]

50. Hindy, G.; Sonestedt, E.; Ericson, U.; Jing, X.J.; Zhou, Y.; Hansson, O.; Renstrom, E.; Wirfalt, E.; Orho-Melander, M. Role of TCF7L2 risk variant and dietary fibre intake on incident type 2 diabetes. *Diabetologia* **2012**, *55*, 2646–2654. [CrossRef]
51. Grau, K.; Cauchi, S.; Holst, C.; Astrup, A.; Martinez, J.A.; Saris, W.H.; Blaak, E.E.; Oppert, J.M.; Arner, P.; Rossner, S.; et al. TCF7L2 rs7903146-macronutrient interaction in obese individuals' responses to a 10-wk randomized hypoenergetic diet. *Am. J. Clin. Nutr.* **2010**, *91*, 472–479. [CrossRef] [PubMed]
52. Fisher, E.; Boeing, H.; Fritsche, A.; Doering, F.; Joost, H.G.; Schulze, M.B. Whole-grain consumption and transcription factor-7-like 2 (TCF7L2) rs7903146: Gene-diet interaction in modulating type 2 diabetes risk. *Br. J. Nutr.* **2009**, *101*, 478–481. [CrossRef] [PubMed]
53. Lyssenko, V.; Lupi, R.; Marchetti, P.; Del Guerra, S.; Orho-Melander, M.; Almgren, P.; Sjogren, M.; Ling, C.; Eriksson, K.F.; Lethagen, A.L.; et al. Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. *J. Clin. Invest.* **2007**, *117*, 2155–2163. [CrossRef] [PubMed]
54. Leiherer, A.; Muendlein, A.; Saely, C.H.; Fraunberger, P.; Drexel, H. Serotonin is elevated in risk-genotype carriers of TCF7L2 - rs7903146. *Sci. Rep.* **2019**, *9*, 12863. [CrossRef] [PubMed]
55. Ojo, O.; Feng, Q.Q.; Ojo, O.O.; Wang, X.H. The Role of Dietary Fibre in Modulating Gut Microbiota Dysbiosis in Patients with Type 2 Diabetes: A Systematic Review and Meta-Analysis of Randomised Controlled Trials. *Nutrients* **2020**, *12*, 3239. [CrossRef] [PubMed]
56. McKeown, N.M.; Meigs, J.B.; Liu, S.; Rogers, G.; Yoshida, M.; Saltzman, E.; Jacques, P.F. Dietary carbohydrates and cardiovascular disease risk factors in the Framingham offspring cohort. *J. Am. Coll. Nutr.* **2009**, *28*, 150–158. [CrossRef] [PubMed]
57. Qi, X.X.; Shen, P. Associations of dietary protein intake with all-cause, cardiovascular disease, and cancer mortality: A systematic review and meta-analysis of cohort studies. *Nutr. Metab. Cardiovasc. Dis.* **2020**, *30*, 1094–1105. [CrossRef]
58. Guasch-Ferré, M.; Satija, A.; Blondin, S.A.; Janiszewski, M.; Emlen, E.; O'Connor, L.E.; Campbell, W.W.; Hu, F.B.; Willett, W.C.; Stampfer, M.J. Meta-analysis of randomized controlled trials of red meat consumption in comparison with various comparison diets on cardiovascular risk factors. *Circulation* **2019**, *139*, 1828–1845. [CrossRef]
59. Kurilshikov, A.; Medina-Gomez, C.; Bacigalupe, R.; Radjabzadeh, D.; Wang, J.; Demirkan, A.; Le Roy, C.I.; Raygoza Garay, J.A.; Finnicum, C.T.; Liu, X.; et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nat. Genet.* **2021**, *53*, 156–165. [CrossRef] [PubMed]
60. Merino, J.; Dashti, H.S.; Li, S.X.; Sarnowski, C.; Justice, A.E.; Graff, M.; Papoutsakis, C.; Smith, C.E.; Dedoussis, G.V.; Lemaitre, R.N.; et al. Genome-wide meta-analysis of macronutrient intake of 91,114 European ancestry participants from the cohorts for heart and aging research in genomic epidemiology consortium. *Mol. Psychiatry* **2018**, *24*, 1920–1932. [CrossRef]
61. Zeggini, E.; Scott, L.J.; Saxena, R.; Voight, B.F.; Marchini, J.L.; Hu, T.; de Bakker, P.I.; Abecasis, G.R.; Almgren, P.; Andersen, G.; et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **2008**, *40*, 638–645. [CrossRef] [PubMed]
62. Florez, J.C.; Jablonski, K.A.; Bayley, N.; Pollin, T.I.; De Bakker, P.I.; Shuldiner, A.; Knowler, W.C.; Nathan, D.M.; Altshuler, D. TCF7L2 Polymorphisms and Progression to Diabetes in the Diabetes Prevention Program. *New Engl. J. Med.* **2006**, *355*, 241–250. [CrossRef] [PubMed]
63. Garver, W.S.; Newman, S.B.; Gonzales-Pacheco, D.M.; Castillo, J.J.; Jelinek, D.; Heidenreich, R.A.; Orlando, R.A. The genetics of childhood obesity and interaction with dietary macronutrients. *Genes Nutr.* **2013**, *8*, 271–287. [CrossRef]
64. Giambartolomei, C.; Vukcevic, D.; Schadt, E.E.; Franke, L.; Hingorani, A.; Wallace, C.; Plagnol, V. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **2014**, *10*, e1004383. [CrossRef]

Paper IV

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-18386-9 OPEN

An investigation of causal relationships between prediabetes and vascular complications

Pascal M. Mutie[1,6], Hugo Pomares-Millan[1,6], Naeimeh Atabaki-Pasdar[1], Nina Jordan[2], Rachel Adams[3], Nicole L. Daly[3], Juan Fernandes Tajes[1], Giuseppe N. Giordano[1] & Paul W. Franks[1,4,5]✉

Prediabetes is a state of glycaemic dysregulation below the diagnostic threshold of type 2 diabetes (T2D). Globally, ~352 million people have prediabetes, of which 35–50% develop full-blown diabetes within five years. T2D and its complications are costly to treat, causing considerable morbidity and early mortality. Whether prediabetes is causally related to diabetes complications is unclear. Here we report a causal inference analysis investigating the effects of prediabetes in coronary artery disease, stroke and chronic kidney disease, complemented by a systematic review of relevant observational studies. Although the observational studies suggest that prediabetes is broadly associated with diabetes complications, the causal inference analysis revealed that prediabetes is only causally related with coronary artery disease, with no evidence of causal effects on other diabetes complications. In conclusion, prediabetes likely causes coronary artery disease and its prevention is likely to be most effective if initiated prior to the onset of diabetes.

[1] Genetic and Molecular Epidemiology Unit, Lund University Diabetes Centre, Department of Clinical Sciences, Clinical Research Centre, Lund University, Skåne University Hospital, Jan Waldenströms gata 35, Malmö SE-20502, Sweden. [2] Regulatory Affairs Intelligence, Novo Nordisk A/S, Copenhagen, Denmark. [3] Regulatory Affairs—Neuroscience and Cardiovascular Metabolism, Janssen, High Wycombe, UK. [4] Department of Public Health and Clinical Medicine, Section for Medicine, Umeå University, Umeå, Sweden. [5] Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [6] These authors contributed equally: Pascal M. Mutie, Hugo Pomares-Millan. ✉email: paul.franks@med.lu.se

Prediabetes is an impaired state of glucose metabolism defined by elevated but not yet diabetic levels of fasting or 2-h glucose, or HbA1c. The specific cutoffs used to define prediabetes vary but the widely adopted American Diabetes Association (ADA) definitions are: impaired fasting glucose (IFG) = fasting glucose 5.6–6.9 mmol L^{-1}; impaired glucose tolerance (IGT) = 2-h glucose 7.8–11.0 mmol L^{-1}; HbA1c = 39–46 mmol mol^{-1} (or 5.7–6.4%). The cooccurrence of IFG and IGT is termed "impaired glucose regulation".

Whilst the global prevalence of prediabetes in adults is about 7.3% ($n = 352$ million people), in Europe and the US, roughly 4.6% ($n = 36$ million people) and 33.9% ($n = 84.1$ million people) of the adult populations, respectively, are estimated to have prediabetes[1]. In the short term, a relatively small proportion (5–10% annually) of those with prediabetes will progress to full-blown diabetes; however, after 5 years, about half will have developed the disease[2].

As diabetes progresses, it becomes increasingly difficult to treat, as the capacity to endogenously produce insulin diminishes and life-threatening complications arise. About five million people died from diabetes-related complications in 2015, of which more than 50% of the deaths were cardiovascular in nature, with costs attributable to diabetes amounting to about one trillion USD globally as of 2017[1].

Many observational studies have shown that prediabetes is a risk factor for cardiovascular disease (CVD), suggesting that the pathogenic effects of dysregulated glucose metabolism have already begun even before diabetes is manifest[3]. However, these observations cannot be directly interpreted as causal effects owing to the limitations of observational epidemiology. Nevertheless, if prediabetic blood glucose variation was known to cause micro- and/or macro-vascular disease, this could profoundly impact clinical guidelines for the prevention of micro- and macro-vascular disease.

Following a cohort of participants who remain in the prediabetic state for many years would help determine if blood glucose variations within the prediabetic range are associated with CVD; however, such a study is probably unfeasible and would (owing to its observational nature) be prone to confounding and reverse causality. In theory, one could design a clinical trial in which people with prediabetes are randomized to interventions that either (i) maintain blood glucose at the prediabetic level (e.g., by clamping blood glucose and insulin concentrations), or (ii) cause blood glucose control to deteriorate through diabetes and thereafter assess the impact of these interventions on the development of complications. However, for ethical and other pragmatic reasons, such trials are unlikely to be conducted.

Mendelian randomization (MR) is a recently popularized adjunct to randomized controlled trials (RCTs) that makes use of epidemiological data for causal inference. The approach leverages the strengths (stability and random assortment of alleles) of germline DNA variation to generate so-called "instrumental variables" that serve as proxies for environmental exposures[4]. Whilst not without limitations[5], MR is less prone to confounding and reverse causality than observational epidemiology and has been used extensively to validate causal relationships indicated by observational studies.

For the purpose of the current analysis, we have designed an instrumental variable that isolates the exposure of prediabetes from diabetes by selecting single nucleotide polymorphisms (SNPs) with robust signals for variation in nondiabetic glycaemic traits only, with no signal for risk of type 2 diabetes (T2D). We use these instrumental variables to test whether nondiabetic variations in fasting blood glucose (FG) and glycated hemoglobin (HbA1c) are causally related with the most common micro- and macro-vascular complications of diabetes: heart disease, occlusive and hemorrhagic stroke, and renal disease.

Results

Observational and MR results. Thirty-seven articles were included in the meta-analysis of observational studies. The pooled sample size was 1,326,915 participants, with mean (±SD) age 53.2 ± 10.2 years and follow-up duration of 9.6 ± 4.8 years.

In the observational data meta-analysis, prediabetes was associated with a 16% elevated risk of coronary artery disease (CAD) (RR = 1.16; 95% CI: 1.09, 1.23; $Q = 52.5$, $P_{Qstat} = 0.058$; $I^2 = 27.7\%$; Fig. 1). In the MR analysis, nondiabetic fasting glucose variation was also significantly associated with CAD, such that 1 mmol L^{-1} higher fasting glucose conveyed an OR of 1.26 (95% CI: 1.16, 1.38) for CAD, with no evidence of directional horizontal pleiotropy (Egger intercept = 1, $P = 0.76$) (Table 1 and Fig. 2). Sensitivity analyses (MR-Egger and weighted median regression) yielded consistent results. Hba1c yielded eight SNPs, which were not classifiable as erythrocytic or glycemic. The association between HbA1c and risk of CAD was not statistically significant (OR = 1.03; 95% CI: 0.64, 1.64) and there was evidence of directional horizontal pleiotropy (Egger intercept = 1.03, $P = 0.01$; Table 1).

In observational analyses, prediabetes conveyed a RR of 1.11 (95% CI: 1.03, 1.18; $Q = 28.5$, $P_{Qstat} = 0.23$; $I^2 = 16\%$) for stroke (Fig. 3), these remained virtually unchanged in the subgroup analysis (Supplementary Data 2); however, in the MR analysis, prediabetes was not causally associated with overall stroke (any stroke (AS), OR = 0.88, 95% CI: 0.69, 1.13) or any of the subtypes of stroke (Table 1). Prediabetes was not associated with chronic kidney disease (CKD) in the observational analysis (RR = 1.05; 95% CI: 0.98, 1.12; $Q = 27.2$, $P_{Qstat} = 0.002$; $I^2 = 63.3\%$), Fig. 4, or in the MR analyses (OR = 1.04; 95% CI: 0.87, 1.25), see below. In the latter, there was no evidence of horizontal pleiotropy.

Sensitivity analyses. In further sensitvity and validation analyses of the prediabetes-only instrument, as defined in our study, prediabetes-only SNPs were not significantly associated with T2D risk across all MR methods used, $P > 0.05$ (Table 2). However, when using all FG SNPs that were genome-wide significant ($P < 5 \times 10^{-8}$) regardless of whether or not they were nominally associated with T2D, there was a strong causal relationship between FG and T2D, $P < 0.01$ across all methods. There was, however, a high degree of horizontal pleiotropy, $P_{Egger\ intercept} < 0.01$, which underscores the complex nature of T2D (Table 3). All observational pooled estimates remained virtually unchanged in the sensitivity analysis (Supplementary Figs. 1–3).

We further tested for pleiotropy and presence of outliers using the Mendelian Randomization Pleiotropy RESidual Sum and Outlier (MRPRESSO) method for outcomes where outliers were detected—coronary artery disease (CAD), AS and any ischemic stroke (AIS). This method detects horizontal pleiotropy, corrects for it, and also tests the distortion between the corrected and uncorrected causal estimates[6]. The outlier-corrected results did not differ with the inverse-variance weighted (IVW) results for these outcomes (Table 4). In addition, we conducted leave-one-out sensitivity analyses of the relationship between prediabetes and CAD, one using the original 28 SNPs and another using SNPs corrected for outliers using MRPRESSO, to assess whether this association was being driven by one or more influential SNPs. Our results show that the relationship between prediabetes and CAD is not driven by a single (or more) influential genetic variant (s) (Fig. 5). When we used 2-h glucose levels as an instrumental variable for prediabetes, only two SNPs remained after routine quality control (QC) and use of all genome-wide significant SNPs ($n = 7$ after QC) did not return significant results in association with CAD (Supplementary Note 2 and Supplementary Table 1). Further sensitivity assessments of the relationship between our

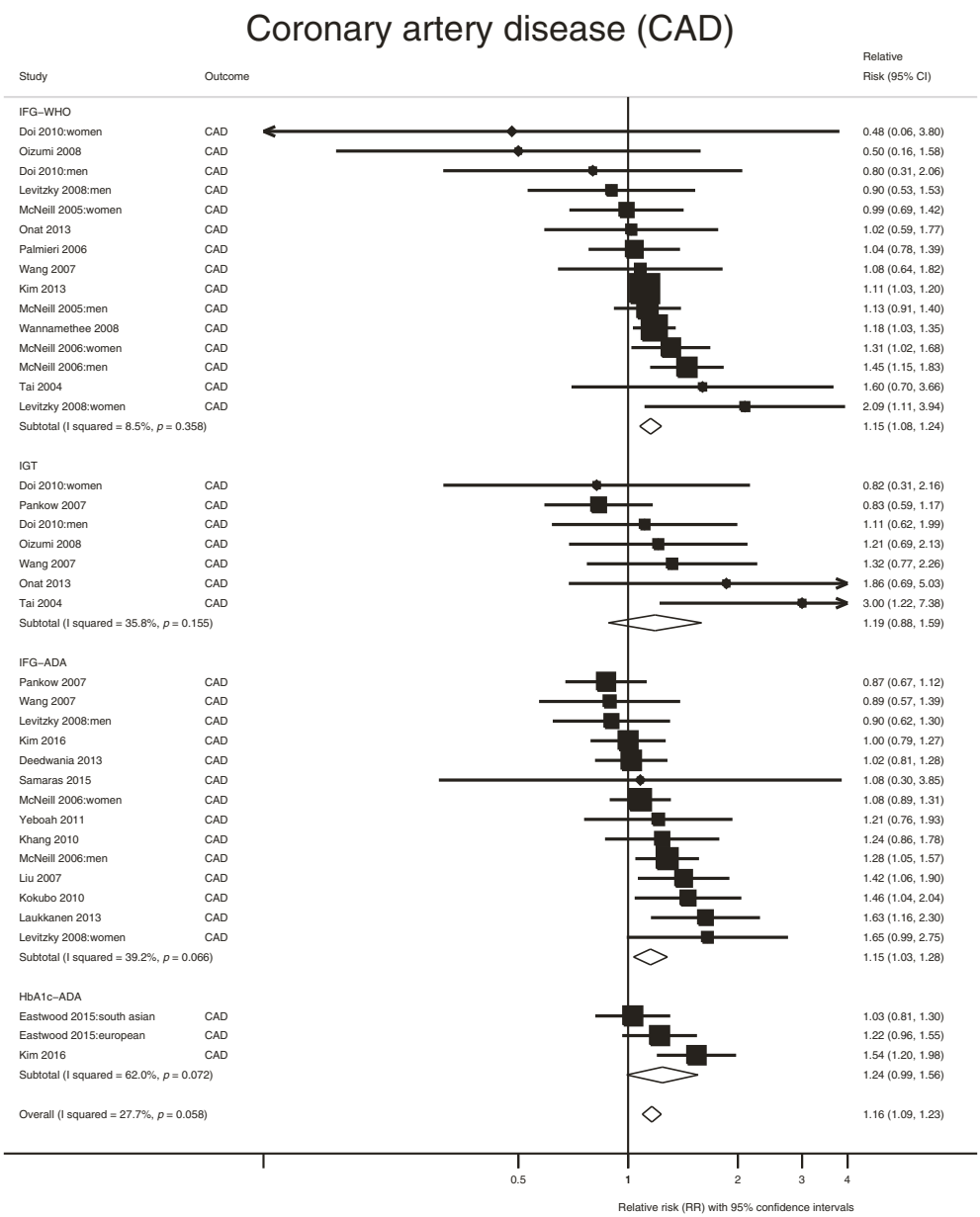


Fig. 1 Meta-analysis of the association between prediabetes and CAD. The square and diamond shapes represent effect size (relative risk estimates), while the horizontal bars represent the 95% confidence intervals. A total of 21 studies are included. All *P* values are two-sided. Source data are provided as Source Data file.

Table 1 Causal relationship between genetically determined prediabetes and vascular outcomes.

Trait associated with FG	IVW_{robust} (OR (95% CI))	MR-Egger (OR (95% CI))	Egger intercept *P* value	Weighted median (OR (95% CI))
CAD	1.26 (1.14, 1.38)	1.30 (1.09, 1.567)	0.76	1.29 (1.13, 1.47)
Any stroke	0.88 (0.68, 1.13	0.71 (0.47, 1.08)	0.34	0.82 (0.64, 1.07)
AIS	0.92 (0.73, 1.16)	0.70 (0.48, 1.02)	0.16	0.88 (0.67, 1.15)
LAS	0.83 (0.49, 1.40)	0.66 (0.33, 1.35)	0.48	0.79 (0.43, 1.46)
CES	1.10 (0.75, 1.63)	0.79 (0.39, 1.58)	0.21	1.04 (0.63, 1.73)
SVS	0.78 (0.46, 1.31)	0.49 (0.19, 1.22)	0.23	0.61 (0.33, 1.11)
CKD	1.04 (0.87, 1.25)	0.83 (0.56, 1.22)	0.32	0.93 (0.75, 1.16)
HbA1c-CAD[a]	1.03 (0.64, 1.64)	0.17 (0.04, 0.79)	0.01	0.83 (0.53, 1.31)

Data are presented as odds ratios and 95% CI for three methods of the Mendelian randomization analysis. Source data are provided as Source Data file.
IVW inverse-variance weighted, *CAD* coronary artery disease, *AIS* any ischemic stroke, *LAS* large artery stroke, *CES* cardioembolic stroke, *SVS* small vessel stroke, *CKD* chronic kidney disease.
[a]Two-sample MR results of the association between genetically determined HbA1c levels and CAD using robust IVW.

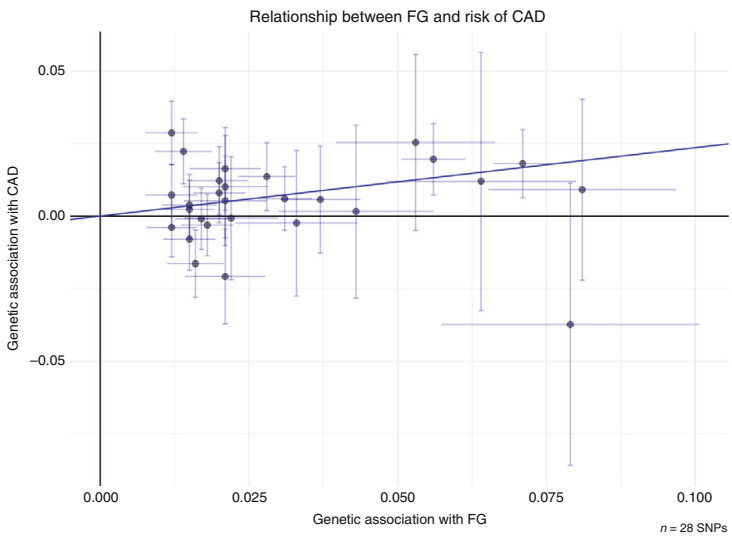


Fig. 2 Relationship between genetic effects of prediabetes only and CAD. Data are represented as log-odds and 95% confidence intervals for each trait. Slope of the line represents an estimate of the causal effect of fasting glucose on risk of CAD. The points represent effect sizes for each individual genetic variant (SNPs) for each of the traits on both axes. The horizontal and vertical bars at each point represent the 95% confidence intervals for genetic associations with FG and CAD, respectively. FG fasting glucose, CAD coronary artery disease. Source data are provided as Source Data file.

prediabetes instruments and other cardiovascular risk factors (Total, LDL, and HDL cholesterol levels; tryglyceride levels; and body mass index) did not show any significant association (Supplementary Note 2 and Supplemetary Tables 2–6).

Discussion

It is unclear if prediabetes is pathogenic or merely a prelude to the disease state of diabetes. We sought to address this important question using MR to estimate the causal effect of nondiabetic variations in FG on the major complications of diabetes. We compared these findings with those obtained through meta-analysis of published observational data from 1,326,915 participants. In the observational analysis, prediabetes was modestly associated with CAD and stroke, but not with CKD. In the MR analyses however, only prediabetic blood glucose was associated with CAD, with a 26% higher odds of CAD per mmol L^{-1} increase in fasting glucose. Elevation in genetically determined HbA1c did not confer a statistically significant increase in the odds of CAD or any other outcomes, though the number of instruments was less ($n = 8$) and the instruments were unclassifiable.

To date, there has been no medicinal products approved for the treatment of prediabetes in the EU or US. While lifestyle measures are clearly recommended as first-line intervention to improve glycaemia in people at high risk of developing diabetes, it is widely acknowledged that additional drug therapy may be beneficial in people with prediabetes, if their risk of diabetes is elevated for other reasons.

Current regulatory requirements for supportive evidence include showing that delay in disease progression is accompanied by other indicators of clinical benefit[7]. To provide this evidence, large, long-term clinical trials are needed, the high cost of which inhibits the development of prediabetic medicinal products. Moreover, there are reimbursement challenges of treating very

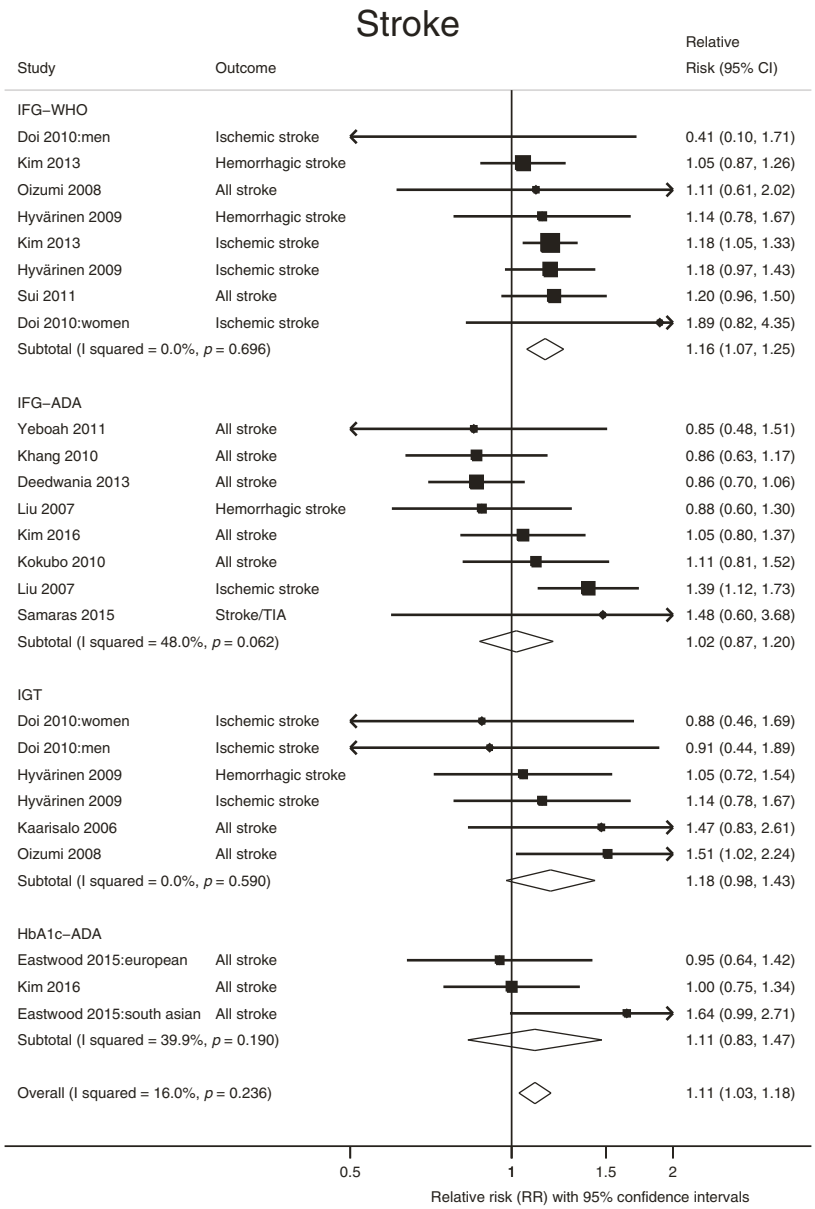


Fig. 3 Meta-analysis of the association between prediabetes and stroke. The square and diamond shapes represent effect size (relative risk estimates), while the horizontal bars represent the 95% confidence intervals. A total of 14 studies are included. All *P* values are two-sided. Source data are provided as Source Data file.

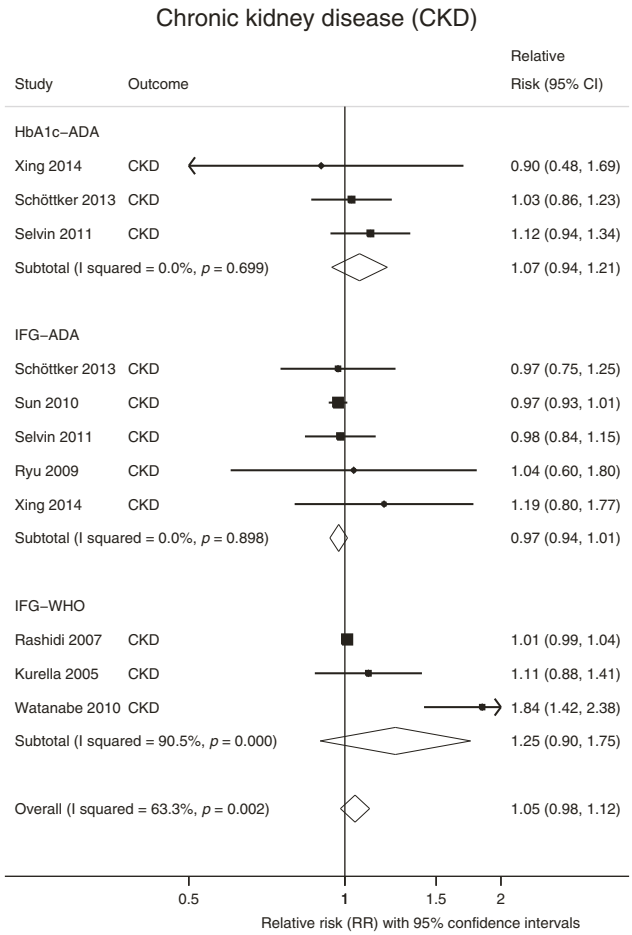


Fig. 4 Meta-analysis of the association between prediabetes and CKD. The square and diamond shapes represent effect size (relative risk estimates), while the horizontal bars represent the 95% confidence intervals. In total, eight studies are included. All *P* values are two-sided. Source data are provided as Source Data file.

Table 2 Causal association between prediabetes only and risk of T2D.

Method	OR	Lower 95% CI	Upper 95% CI	*P* value
Weighted median	0.98	0.82	1.14	0.79
IVW	1.02	0.90	1.16	0.76
Robust IVW	1.02	0.90	1.15	0.77
MR-Egger	0.91	0.73	1.14	0.42
$Intercept_{MR\text{-}Egger}$	1.00	1.00	1.01	0.23
Robust MR-Egger	0.91	0.77	1.07	0.25
$Intercept_{Robust\ MR\text{-}Egger}$	1.00	1.00	1.01	0.15

n = 28 SNPs. Results are from two-sample Mendelian randomization analyses and *P* values are two-sided. Results are unadjusted for multiple comparisons. Source data are provided as Source Data file.
IVW inverse-variance weighted, *OR* odds ratio.

large numbers of people with prediabetes. Determination of the health implications and risk assessment of prediabetes would, therefore, aid design of smaller, shorter, and potentially less expensive, clinical trials by providing alternative health benefits. It would also help address the value of treating large populations over longer periods, by showing cost effectiveness.

MR is often considered an analogue of RCTs. In the latter, treatment allocation is randomized to help ensure that any potential confounding factors that exist within the cohort prior to treatment assignment are distributed evenly between treatment arms, thus neutralizing their impact. In MR analyses, germline DNA variants are used as proxies (instrumental variables) for the exposure of interest (in this case, prediabetes). The random assortment of alleles during meiosis and the stability of DNA variants across the lifespan reduce to a bare minimum the possibility that the observed effect of the instrumental variable

Table 3 Causal association between fasting glucose (all GWA significant) and risk of T2D.

Method	OR	Lower 95% CI	Upper 95% CI	*P* value
Weighted median	1.55	1.23	1.94	1.67×10^{-4}
IVW	2.26	1.37	3.74	1.43×10^{-3}
Robust IVW	2.35	1.50	3.67	1.75×10^{-4}
MR-Egger	0.46	0.19	1.12	0.09
$Intercept_{MR\text{-}Egger}$	1.05	1.03	1.08	5.05×10^{-5}
Robust MR-Egger	0.96	0.45	2.03	0.91
$Intercept_{Robust\ MR\text{-}Egger}$	1.03	1.01	1.04	5.54×10^{-3}

$n = 74$. Results are from two-sample Mendelian randomization analyses and *P* values are two-sided. Results are unadjusted for multiple comparisons. Source data are provided as Source Data file.
IVW inverse-variance weighted, *OR* odds ratio.

Table 4 MRPRESSO analysis of relationship between prediabetes and outcomes with detected outliers.

Outcome	MR analysis	OR (95% CI)	*P* value
Coronary artery disease	Raw	1.27 (1.09, 1.47)	4.9×10^{-3}
	Outlier-corrected	1.24 (1.12, 1.38)	5.8×10^{-4}
Any stroke	Raw	0.92 (0.73, 1.17)	0.51
	Outlier-corrected	0.90 (0.72, 1.11)	0.32
Any ischemic stroke	Raw	0.95 (0.75, 1.22)	0.71
	Outlier-corrected	0.90 (0.74, 1.09)	0.28

All *P* values are two-sided. "Raw" refers to original FG SNPs ($n = 28$). Source data are provided as Source Data file.
OR odds ratio, *CI* confidence interval.

on the outcome is confounded or attributable to reverse causality[4].

Here, we specifically sought to isolate the causal effects of prediabetes from those of diabetes by selecting variants that are robustly associated with fasting glucose and HbA1c variation but not with diabetes. It is hard to envisage a clinical trial where this could be recapitulated, as participants would need to be exposed to prediabetes without progressing to diabetes long enough for complications to occur. Consider, too, that the method used to maintain the prediabetic state would need to function without directly affecting the trial's outcomes, excluding virtually all known blood glucose therapeutics. Thus, for this specific research question, MR is an especially powerful method for causal inference.

One of few naturally occurring examples where blood glucose can remain in the prediabetic state for long periods is a rare form of monogenic diabetes (MODY2), caused by mutations in the glucokinase gene (*GCK*). In MODY2, the blood glucose set-point is elevated, but is generally not linked with progressively deteriorating glycemic control. Moreover, most MODY2 patients do not develop macro- and micro-vascular complications[8]. As intriguing as this is, the physiological idiosyncrasies of the disease limit inferences about vascular risk in prediabetes. For example, unlike many people with prediabetes, MODY2 patients have normal post-prandial glycemic responses, virtually no insulin resistance and cardioprotective lipid profiles[9].

Although this is the first study to our knowledge to undertake a comprehensive systematic literature review coupled with a detailed MR analysis to specifically examine the causal effects of prediabetic blood glucose variation in micro- and macro-vascular disease, previous studies have examined the cardiogenic effects of diabetic and nondiabetic blood glucose variations. In general, the findings from these studies support the clinical consensus that T2D causes heart disease[10].

At least one previous MR study examined fasting glucose variation (inclusive of diabetes) in ischemic stroke and found no statistically robust evidence of effect[11]. However, a published MR analysis that, like our study, harnessed genetic variants associated with glucose but not diabetes[12], also reported evidence of causal associations with CAD. Another measure of glycemia, HbA1c, which reflects average glucose levels over the preceding 3 months, was shown in a recent study to be causally associated with cardiovascular complications[13]. However, as shown here, these results may not be independent of the effects of fasting glucose in CVD.

MR is not without limitations. Canalization is a widely described caveat of MR analyses; the phenomenon occurs when genetic perturbations are offset by coexisting and compensatory mechanisms, effectively short-circuiting the exposure-outcome relationships that MR analyses seek to assess[4]. There are no established methods to detect canalization in MR analyses. Canalization could invalidate MR findings by altering the effect of the genetic instrument on the outcome of interest without affecting the association between genotype and exposure of interest[4]. There are other established methodological limitations of MR, such as horizontal pleiotropy and population stratification, which were overcome in the current analysis using established statistical solutions. A further important consideration is that the exposures characterized in MR experiments should be viewed as having lifelong effects, whereas the timeframe for prediabetes exposure will be confined to a much shorter duration. Thus, the estimated effect of prediabetes in CAD derived from our MR analysis may be greater in magnitude than one would observe in the real world. However, the results from our observational meta-analysis are largely consistent with our MR estimates.

A major limitation of observational studies is the potential that participants progress to diabetes. Therefore, we went to great lengths to identify and stratify those studies which excluded individuals with diabetes in the analysis. Those which we deemed having the most likelihood of enrolling diabetics (i.e., those recruiting participants only with HbA1c or fasting glucose) were further stratified into a specific subgroup for re-analysis; results remained virtually unchanged (see Supplementary Material 2, Table 1, subgroup analysis). By no means do we claim that the observational evidence is definitive; on the contrary, this motivated us to contest these observational data and explore causality through the MR approach.

In conclusion, we report the synthesis of a very large body of epidemiological evidence linking prediabetes with the life-threatening complications caused by diabetes and validate these findings using MR. We found that prediabetes is likely to be causal in CAD, whereas it is not likely to cause kidney disease or stroke. The major implication of this finding is that interventions for the prevention of diabetes-related CAD may be more effective

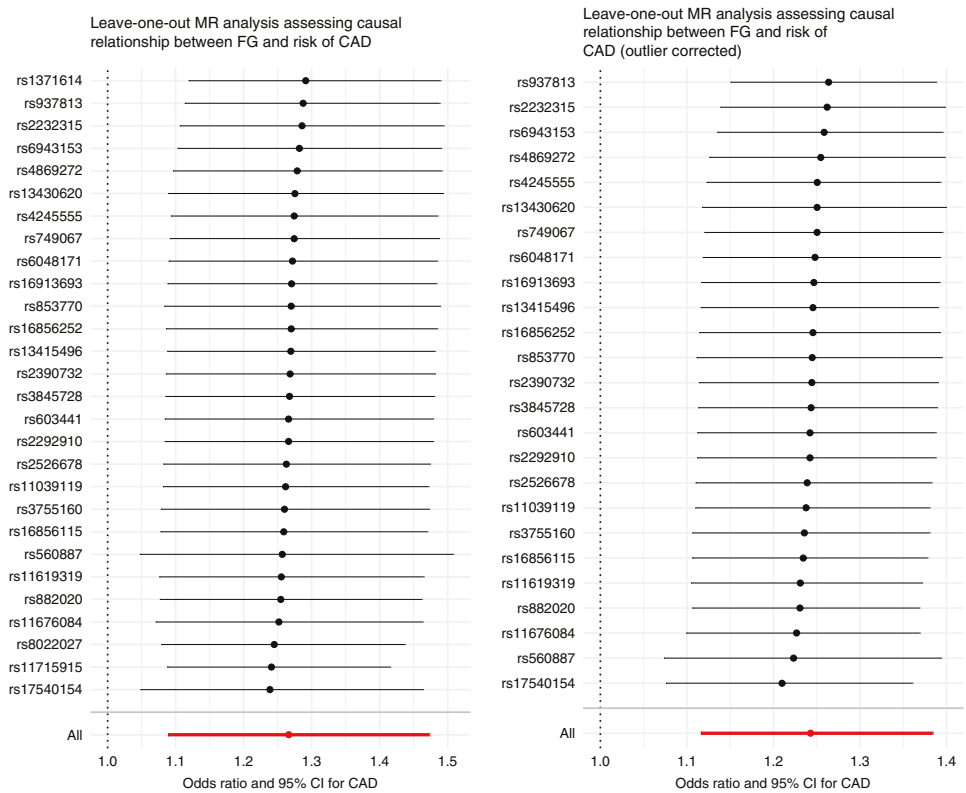


Fig. 5 Leave-one-out analysis plots of causal relationship between fasting glucose and CAD. Data are presented as odds (OR) ratio and 95% confidence interval (95% CI) of the exposure-outcome relationship for each SNP. Center points represent the causal effect estimate and the horizontal bars represent the respective 95% CI. Left panel represents data from all SNPs that passed QC ($n = 28$) while right panel represents SNPs retained after correcting for outliers using MRPRESSO, $n = 25$ SNPs. Source data are provided as Source Data file.

if initiated prior to diabetes onset. This may also help explain why CAD prevention in people with established diabetes has proven extremely challenging[14].

Methods

Observational data meta-analysis. We first performed a systematic literature review of published epidemiological studies focusing on "prediabetes and diabetic complications" and extracted summary statistics that we, thereafter, combined through meta-analysis. We then tested the hypothesis that these observational associations were of a causal nature using MR and compared effect estimates derived from the observational meta-analysis and the MR analyses.

A combined medical subject headings term and text search strategy was formulated restricted to "humans" and English language articles (Supplementary Data 1 shows the search strategy in detail). A search of the electronic database PubMed was carried out for all cohort studies published through November 30th, 2017, according to the following criteria: prediabetes defined by IGT, IFG per WHO[15] or ADA criteria, and glycated hemoglobin (HbA1c) per ADA criterion[16]. Studies were included if participants were drawn from the general population, glycaemia was measured at baseline, and the subsequent outcomes at follow-up were CAD, CKD, or stroke, and were compared with the group of normoglycaemic participants. Studies with individuals known to be diagnosed with diabetes or with diabetic values at baseline or follow-up were excluded from the analysis. Figure 6 shows the study selection procedure.

Data extraction: two authors (H.P-.M. and P.M.M.) independently identified, screened, and reviewed for eligibility the papers identified using the approach defined above. We systematically abstracted data relating to: author(s), year published, country or region, prediabetes definition, prevalence (%), sample size, gender ratio of the study population (%), participants' age, duration of follow-up, glycaemic status at baseline, outcome definition and ascertainment, covariates and approach used to control for confounding, risk estimates and 95% confidence intervals, in a standard form (Supplementary Data 2 shows the studies' characteristics). Discrepancies in study identification were adjudicated by a third researcher (G.N.G.). Quality of the studies and bias assessment was determined using the Newcastle–Ottawa scale[15] (Supplementary Data 2). Reported findings by subgroups (i.e., sex or ethnicity) were included separately by strata for statistical analysis. Effect estimates (relative risk, hazard ratio, and odds ratio, converted to RR) were logarithmically transformed and standard errors calculated[16]. A priori, we assumed there would be heterogeneity across the cohorts given the differences in population characteristics, follow-up duration, research methods, and outcome definitions. Therefore, the DerSimonian and Laird random-effects model for meta-analysis was used, which is considered more conservative than fixed-effect models[16]. Heterogeneity between and within studies was explored through subgroup analysis (Supplementary Data 2).

Publication bias was assessed using funnel plots and the Begg's and Egger's test. Sensitivity analysis was carried out by omitting one study at a time. All statistical meta-analyses were undertaken with the software Stata 13.0 (Stata Corp LP, College Station, TX).

MR analyses. MR is a method that employs instrumental variables to assess the causal association between a given exposure and an outcome[4]. For an instrument to be valid, it must mediate its effect on the outcome only through the exposure and not via other pathways. Further, it should only be associated with the exposure and not be associated with cofounders of the exposure-outcome association[17]. To

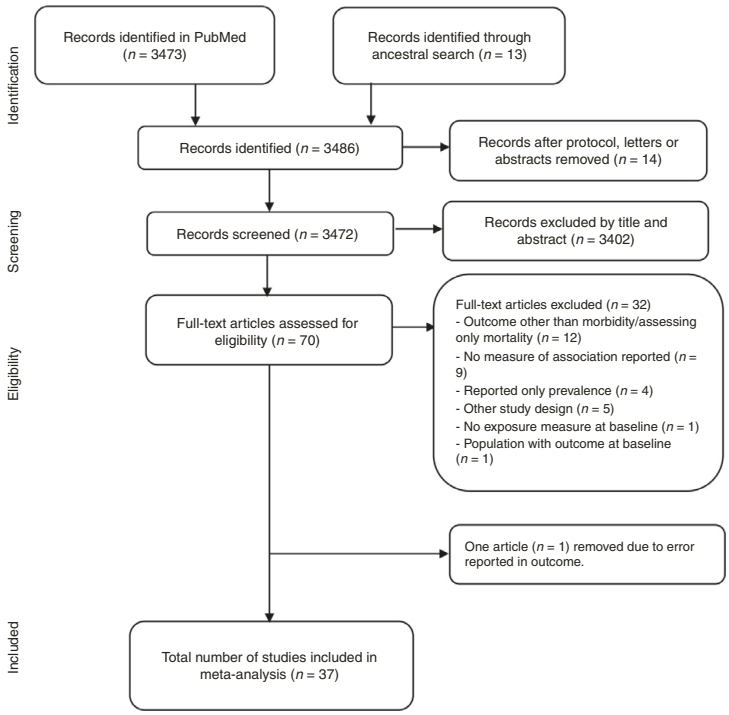


Fig. 6 Outline of study selection procedure. Source data are provided as Source Data file.

reduce potential bias due to population stratification, we restricted MR analyses to participants of European descent.

We defined two sets of instruments that specifically characterized variations in fasting glucose and HbA1c within the nondiabetic range. We achieved this by selecting SNPs that are associated with fasting glucose and HbA1c at a genome-wide level of statistical significance ($P < 5 \times 10^{-8}$) within the most recent MAGIC database[18,19], but which are not associated with type 1 or T2D ($P > 0.05$) in the most recent release of the Diabetes Genetics Replication and Meta-analysis database[20,21]. The sets of instruments derived from these variants were then examined within GWAS databases for any respective "diabetic" complications. Specifically, we used publicly available GWAS meta-analysis summary statistics from various consortia. Fasting glucose (exposure) data were obtained from the Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC, $n = 133{,}010$ for fasting glucose)[22]. The MAGIC GWAS meta-analysis includes 32 cohorts, which comprised participants of European descent adjusted for age and sex. Fasting glucose was expressed in mmol L^{-1} and was untransformed in the analyses[18].

HbA1c (exposure) data were also obtained from the latest MAGIC transethnic genome-wide association meta-analysis of genetic variants associated with HbA1c. This meta-analysis included 159,940 participants from 82 cohorts of different ancestries (European, South and East Asian, and African). Individuals of European ancestry were the majority, about 120,962 across 55 cohorts. All participants were diabetes free and studies reported HbA1c as percentage[19].

CAD GWAS summary statistics were obtained from the latest cardiomics meta-analysis data repository[23]. This data comprised of 34541 cases of CAD and 26,1984 controls from the UK Biobank and replication was done in 88,192 cases and 162,544 controls from Coronary Artery Disease (C4D) Genetics consortium (CARDIoGRAMplusC4D)[24,25].

Summary statistics for five phenotypes of stroke (AS, AIS, large artery stroke, cardioembolic stroke, and small vessel stroke) were obtained from the most recent MEGASTROKE consortium meta-analysis data repository[26] in which the analysis for European only ancestry consisted of 40,585 cases and 406,111 controls[27].

Data on renal disease were obtained from the CKDGen GWAS summary data repository[28]. GWAS meta-analysis for CKD (defined as eGRFcrea <60 ml per min per 1.73 m^2) was performed on a sample of 745,348 and replicated in a sample of 280,722 giving a combined sample size of more than one million[29].

Selection of glucose-associated SNPs from MAGIC[30], as outlined above, resulted in 47 SNPs for fasting glucose and 10 for HbA1c that we considered reflective of prediabetic glucose variation. To rule out linkage disequilibrium (LD) between SNPs, we performed LD-clumping restricted to $r^2 < 0.2$, a 1000 kb window and retained SNPs with the lowest P value resulting in final sets of 28 uncorrelated fasting glucose SNPs and 8 HbA1c SNPs. For each outcome, these genetic variants were further validated for use in the final analysis. Specifically, the exposure-outcome datasets were harmonized to ensure the same number of SNPs in exposure and outcome sets, similar strand orientation, correct direction of effect sizes, and correcting for palindromic SNPs[31].

Statistical analysis. All MR analyses were conducted with the R statistical software v3.6.1 using the MendelianRandomization[32] and TwoSampleMR packages[33].

We used the robust IVW method for the main analysis and the robust MR-egger and weighted median methods for sensitivity analyses. IVW is a widely-accepted approach for MR analyses, which involves regressing the effect sizes of the SNP-outcome association on the SNP-exposure association with the inverse of the variance used as weights. In robust regression, extreme values are penalized to minimize bias.

MR-Egger is used to test for directional horizontal pleiotropy, a violation of the instrumental variable assumption where the effect of the instrumental variable on the outcome is mediated via another pathway other than the exposure of interest. MR-Egger tests for violation of IV assumptions and bias in the inverse variance-weighted (IVW) methods and includes the intercept as part of the regression (unlike IVW, where the intercept is forced to zero)[34]. The resulting coefficient, therefore, provides an asymptotically consistent estimate of the causal effect, even if all variants are pleiotropic with the outcome[35]. This holds when the Instrument Strength Independent of Direct Effect assumption is true, i.e., the instrument strength is independent of its pleiotropic effect. When this criterion is met, MR-Egger provides an unbiased assessment of the association between the exposure and outcome, providing the intercept, which provides the average pleiotropic effect, does not significantly differ from the null. When the intercept is significantly different from the null, it represents an estimate of the directional horizontal pleiotropic effect of the genetic variants[35]. The median-weighted method provides

a reliable estimate of the causal association between exposure and outcome when at least half of the instrumental variables are valid[36].

Sensitivity analyses and instrument validation. To rule out false positive associations, we conducted sensitivity analyses to further test the veracity of our instrumental variables. First, we tested the association between the prediabetes instruments with T2D to demonstrate that our instruments represented pre-diabetes only and rule out any pleiotropic relationship with T2D. Second, we tested the association between all fasting glucose SNPs that reached GWA significance ($n = 74$ after QC) and the risk of T2D, to cement the above facts. Further, we tested if there was any causal relationship between fasting glucose and other cardiometabolic risk factors i.e., BMI, cholesterol levels (total, LDL, and HDL), and triglyceride levels. We also additionally used MRPRESSO to test for horizontal pleiotropy and outliers[6].

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

Data availability

The GWAS summary statistics data analyzed here are available in the following public repositories. CAD (Dataset: CAD_META.gz): https://data.mendeley.com/datasets/gbbsrpx6bs/1#file-67c31537-5906-40bb-9820-8764b1554666 (https://doi.org/10.17632/gbbsrpx6bs.1)[23]. CKD (Dataset: CKD overall European ancestry): http://ckdgen.imbi.uni-freiburg.de/[28]. T2D (Dataset: T2D GWAS meta-analysis—Unadjusted for BMI[20]): https://www.diagram-consortium.org/downloads.html[21]. Fasting glucose, 2-h glucose, and HbA1c: https://www.magicinvestigators.org/downloads/[22]. The fasting and 2-h glucose datasets are filed under Metabochip replication datasets, and the zipped file contains both datasets (ftp://ftp.sanger.ac.uk/pub/magic/MAGIC_Metabochip_Public_data_release_25Jan.zip). The HbA1c dataset can be retrieved at ftp://ftp.sanger.ac.uk/pub/magic/HbA1c_METAL_European.txt.gz. Stroke: https://megastroke.org/download.html[26]. The dataset (MEGASTROKE_data.zip) is accessible after agreeing to terms of use and submitting a brief project description. Lipids: http://csg.sph.umich.edu/willer/public/lipids2013/[37]. The datasets are filed under "RESULT FILES," subheading "JOINT ANALYSIS OF METABOCHIP AND GWAS DATA." The names of the files are LDL Cholesterol, HDL Cholesterol, Triglycerides, and Total Cholesterol. Body mass index: http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files[38]. The dataset is filed under "BMI and Height GIANT and UK BioBank Meta-analysis Summary Statistics." The name of the file is "Meta-analysis Wood et al. + UKBiobank 2018 GZIP". Source data are provided with this paper.

Received: 21 August 2019; Accepted: 31 July 2020;
Published online: 14 September 2020

References

1. International Diabetes Federation. *IDF Diabetes Atlas* 8th edn, 150 (International Diabetes Federation, Brussels, Belgium, 2017).
2. Tabák, A. G., Herder, C., Rathmann, W., Brunner, E. J. & Kivimäki, M. Prediabetes: a high-risk state for diabetes development. *Lancet* **379**, 2279–2290 (2012).
3. Haffner, S. M., Stern, M. P., Hazuda, H. P., Mitchell, B. D. & Patterson, J. K. Cardiovascular risk factors in confirmed prediabetic individuals: does the clock for coronary heart disease start ticking before the onset of clinical diabetes? *JAMA* **263**, 2893–2898 (1990).
4. Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).
5. VanderWeele, T. J., Tchetgen Tchetgen, E. J., Cornelis, M. & Kraft, P. Methodological challenges in mendelian randomization. *Epidemiology* **25**, 427–435 (2014).
6. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
7. Enzmann, H. et al. Guidelines on clinical investigation of medicinal products in the treatment or prevention of diabetes mellitus: Draft. No. CPMP/EWP/1080/00 Rev. 2. (EMA, London, UK, 2018).
8. Steele, A. M. et al. Prevalence of vascular complications among patients with glucokinase mutations and prolonged, mild hyperglycemia. *JAMA* **311**, 279–286 (2014).
9. Fendler, W. et al. Less but better: cardioprotective lipid profile of patients with GCK-MODY despite lower HDL cholesterol level. *Acta Diabetol.* **51**, 625–632 (2014).
10. Leon, B. M. & Maddox, T. M. Diabetes and cardiovascular disease: epidemiology, biological mechanisms, treatment recommendations and future research. *World J. Diabetes* **6**, 1246–1258 (2015).
11. Larsson, S. C. et al. Type 2 diabetes, glucose, insulin, BMI, and ischemic stroke subtypes: Mendelian randomization study. *Neurology* **89**, 454–460 (2017).
12. Merino, J. et al. Genetically driven hyperglycemia increases risk of coronary artery disease separately from type 2 diabetes. *Diabetes Care* **40**, 687–693 (2017).
13. Au Yeung, S. L., Luo, S. & Schooling, C. M. The impact of glycated hemoglobin (HbA1c) on cardiovascular disease risk: a Mendelian Randomization Study using UK Biobank. *Diabetes Care* **41**, 1991–1997 (2018).
14. The Look AHEAD Research Group. Cardiovascular effects of intensive lifestyle intervention in type 2. *Diabetes* **369**, 145–154 (2013).
15. Wells, G. A. et al. Newcale-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp (2014).
16. Higgins, J. P. T. & Green, S. (editors). *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, (2011).
17. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–R98 (2014).
18. Scott, R. A. et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).
19. Wheeler, E. et al. Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: a transethnic genome-wide meta-analysis. *PLoS Med.* **14**, e1002383 (2017).
20. Mahajan, A. et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* **50**, 559–571 (2018).
21. Consortium, D. *Diabetes Genetics Replication and Meta-analysis* https://www.diagram-consortium.org/downloads.html (2018).
22. Consortium, M. *The Meta-Analyses of Glucose and Insulin-related traits Consortium* https://www.magicinvestigators.org (2010).
23. Pim van der Harst. *CAD meta-analysis, Mendeley Data, v1* https://data.mendeley.com/datasets/gbbsrpx6bs/1 (2017).
24. Deloukas, P. et al. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.* **45**, 25–33 (2013).
25. Schunkert, H. et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43**, 333–338 (2011).
26. Malik, R. et al. *MEGASTROKE Consortium, The International Stroke Genetics Consortium*, https://megastroke.org/index.html (2018).
27. Malik, R. et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat. Genet.* **50**, 524–537 (2018).
28. Wuttke, M. et al. *The CKDGen Consortium* http://ckdgen.imbi.uni-freiburg.de/ (2019).
29. Wuttke, M. et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.* **51**, 957–972 (2019).
30. Morris, A. P. et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
31. Hartwig, F. P., Davies, N. M., Hemani, G. & Davey Smith, G. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int. J. Epidemiol.* **45**, 1717–1726 (2016).
32. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **46**, 1734–1739 (2017).
33. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, https://doi.org/10.7554/eLife.34408 (2018).
34. Burgess, S. & Thompson, S. G. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur. J. Epidemiol.* **32**, 377–389 (2017).
35. Burgess, S., Bowden, J., Fall, T., Ingelsson, E. & Thompson, S. G. Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. *Epidemiology* **28**, 30–42 (2017).
36. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).
37. Willer, C. J. et al. *Global Lipids Genetics Consoritum Results* http://csg.sph.umich.edu/willer/public/lipids2013/ (2013).
38. Yengo, L. et al. *Giant Consortium data files* http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files (2018).

Acknowledgements

We extend our gratitude to the many research groups that have made GWAS summary statistics data publicly available and accessible to the rest of the research community and all participants involved in the numerous studies. This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under Grant agreement No. 115881 (RHAPSODY). This Joint Undertaking receives support from the European Union's Horizon 2020 Research and Innovation Programme and EFPIA. RHAPSODY is also supported in part by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 16.0097. This study also received support from the Swedish Research Council, Strategic Research Area Exodiab, (Dnr 2009-1039), the Swedish Foundation for Strategic Research (IRC15-0067), the Swedish Research Council, Linnaeus Grant (Dnr 349-2006-237), and the European Research Council (CoG-2015_681742_NASCENT). The MEGASTROKE project received funding from sources specified at http://www.megastroke.org/acknowledgments.html. The opinions expressed and arguments employed herein do not necessarily reflect the official views of these funding bodies.

Author contributions

P.M.M.: literature search, data analysis, data interpretation, and writing of the manuscript; H.P-.M.: literature search, data analysis, data interpretation, and writing of the manuscript; N.A-.P.: data interpretation and writing of the manuscript; N.J.: revised the manuscript critically; R.A.: revised the manuscript critically; N.L.D.: revised the manuscript critically; J.F.T.: bioinformatic data retrieval; G.N.G.: data interpretation and writing of the manuscript; P.W.F.: conceived the study design, data interpretation, and writing of the manuscript.

Funding

Open Access funding provided by Lund University.

Competing interests

The authors declare no competing interests.

Ethics approval

This study was conducted using publicly available data and therefore did not require ethical approval.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-18386-9.

Correspondence and requests for materials should be addressed to P.W.F.

Peer review information *Nature Communications* thanks Matthew Budoff, Timothy Frayling and the other anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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