Genetic and phenotypic discordance in cardiometabolic diseases

Coral, Daniel

2024

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

Link to publication

Citation for published version (APA):
Coral, D. (2024). *Genetic and phenotypic discordance in cardiometabolic diseases*. [Doctoral Thesis (compilation), Department of Clinical Sciences, Malmö]. Lund University, Faculty of Medicine.

Total number of authors:
1

Creative Commons License:
CC BY

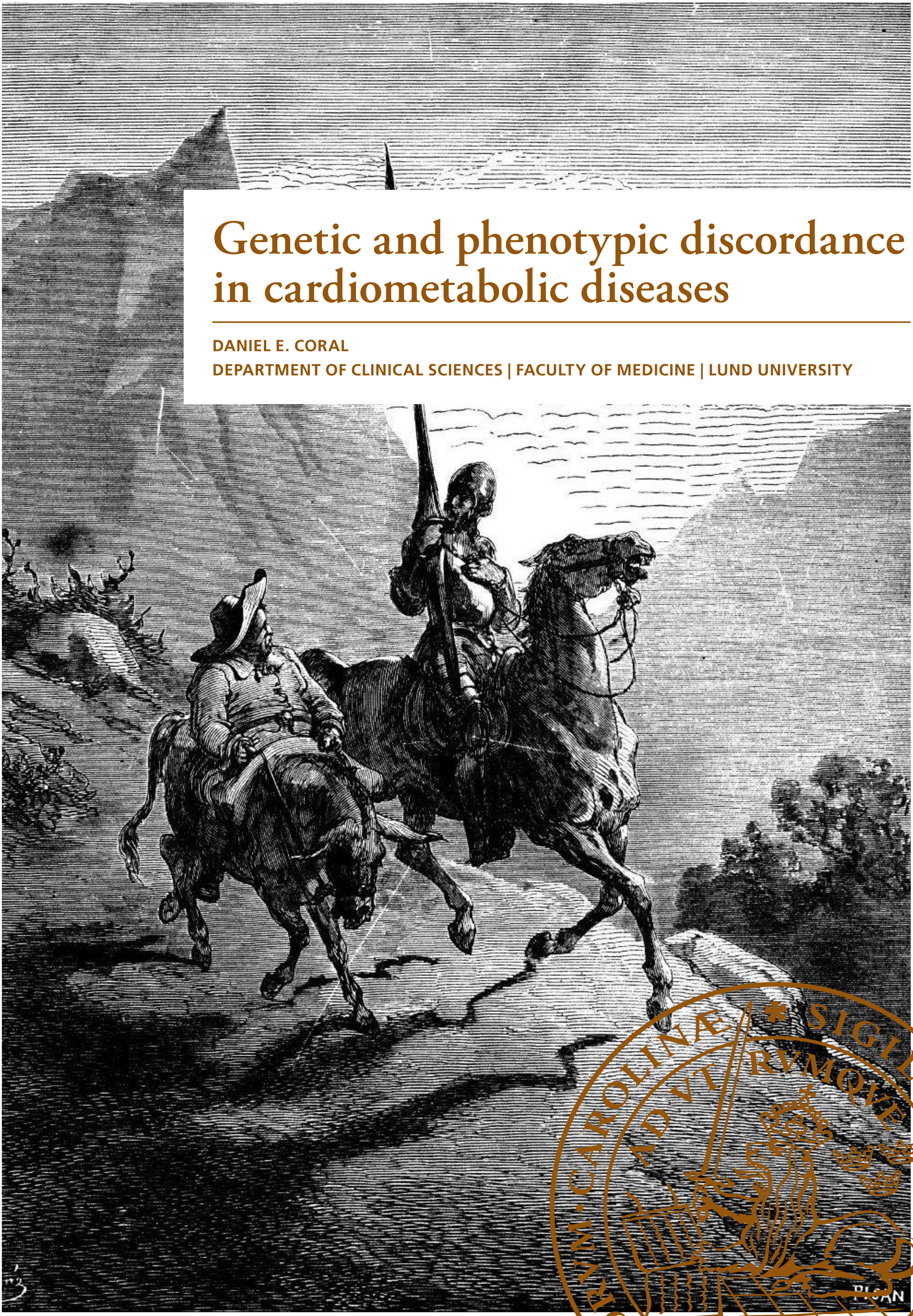
LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Genetic and phenotypic discordance in cardiometabolic diseases

DANIEL E. CORAL

DEPARTMENT OF CLINICAL SCIENCES | FACULTY OF MEDICINE | LUND UNIVERSITY



Printed by Media-Tryck, Lund 2024 NORDIC SWAN ECOLABEL 3041 0903

DANIEL E. CORAL earned his Medical Doctor degree from Universidad del Valle in Cali, Colombia, in 2013. He completed his Master's Program in Public Health at Lund University, Sweden, in 2019. Daniel has completed his PhD at the Genetic and Molecular Epidemiology Unit at the Lund University Diabetes Center in Malmö, Sweden. His doctoral research focuses on identifying genetic and phenotypic factors associated with cardiometabolic profiles at varying risk levels of complications.

LUND UNIVERSITY

FACULTY OF MEDICINE

Department of Clinical Sciences

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2024:41
ISBN 978-91-8021-534-3
ISSN 1652-8220

9 789180 215343

Genetic and phenotypic discordance in cardiometabolic diseases

Genetic and phenotypic discordance in cardiometabolic diseases

Daniel E. Coral

DOCTORAL DISSERTATION

Doctoral dissertation for the degree of Doctor of Philosophy (PhD) at the Faculty of Medicine at Lund University to be publicly defended on 22nd of March at 13.00 in Medelhavet, Department of Clinical Sciences, Jan Waldenströms gata 35, Malmö

Faculty opponent
Prof. Inês Barroso
Thesis advisors
Paul W. Franks, Ewan R. Pearson, Juan Fernandez-Tajes, Giuseppe N. Giordano

Organization: LUND UNIVERSITY

Document name: DOCTORAL DISSERTATION **Date of disputation:** March 22nd, 2024

Author(s): Daniel E. Coral **Sponsoring organization:** IMI-SOPHIA

Title and subtitle: Genetic and phenotypic discordance in cardiometabolic diseases

Abstract:

Cardiometabolic conditions such as obesity and type 2 diabetes (T2D) are among the first causes of death globally, and the number of people affected is rapidly increasing. Both conditions are intricately connected and are associated with many life-threatening cardiovascular disease (CVD). However, both obesity and T2D are highly heterogeneous, such that not all individuals with these conditions develop complications, while others are at disproportionatel higher risk. The purpose of this thesis is to improve our understanding of the heterogeneity in cardiometabolic conditions through the application of genetic analyses and machine learning techniques to identify profiles at disproportionately higher or lower risk of complications, providing insights into the mechanisms that give rise to these profiles and their potential clinical implications. In Paper I, I used cross-trait genetic analysis to derive genetically determined obesity profiles that are associated with higher body mass index (BMI) but are either concordantly associated with higher risk of T2D or discordantly associated with T2D protection. Through a comprehensive phenome-wide comparative analysis of these profiles, we highlighted adipose distribution, vascular function and extracelular matrix remodelling as key mechanisms uncoupling obesity from its diabetogenic risk, and prioritise 17 genes with potential discordant effects in obesity. In Paper II, I reformulated the cross-trait genetic approach from Paper I to construct genetically determined diabetic profiles that are either concordantly associated with higher risk of CVD or discordantly associated with CVD protection. The phenome-wide comparison yields VLDL metabolism as a key mechanism detaching T2D from CVD and 8 loci with cardioprotective effects in T2D that include known targets of current medications, such as statins and GLP-1 receptor agonists. Additionally, I showed through polygenic score analyses (PRS) that quantifying genetic discordance in the general population can aid in CVD risk prediction, improving predictions in 5% of cases and 3% of non-cases. In Paper III, I designed a probabilistic, graph-based clustering ensemble algorithm to identify subgroups of individuals whose levels of common biomarkers of CVD risk deviate from the expected given their body size. We deploy this approach on four large independent cohorts, finding that biomarker deviate substantially from the BMI expectation in ~20% of the general population, and tend to form 5 distinct profiles with specific BMI-biomarker discordance patterns. Considering these discordant profiles can improve CVD prediction with benefits comparable to lipid fraction quantification. In Paper IV, I contributed in an investigation of overall, sex-specific and nonlinear causal effects of BMI on multiple outcomes, including T2D and CVD. We found consistent positive effects of BMI on T2D across sexes, but sex-differential effects on CAD. Evidence of nonlinear effects of BMI were found for lipids and glycaemia. Understanding the mechanisms by which some individuals are at disproportionately higher or lower risk to develop the complications generally associated with obesity and type 2 diabetes can help design better and more precise interventions.

Key words:

Obesity, diabetes, cardiovascular disease, genome-wide association studies, heterogeneity

Supplementary bibliographical information:

Language: English

ISSN and key title: 1652-8220

ISBN: 978-91-8021-534-3

Lund University, Faculty of Medicine Doctoral Dissertation Series 2024:41

Recipient’s notes

Number of pages: 104

Price:

Security classification:

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature Date 2024-Feb-09

Genetic and phenotypic discordance in cardiometabolic diseases

Daniel E. Coral

LUND
UNIVERSITY

Coverphoto: Don Quijote de La Mancha and Sancho Panza by Gustav Doré, 1863. Public Domain

Faculty of Medicine

Department of Clinical Sciences

ISBN 978-91-8021-534-3

ISSN 1652-8220

Printed in Sweden by Media-Tryck, Lund University

Lund 2024

NORDIC SWAN ECOLABEL

Printed matter
3041 0903

Media-Tryck is a Nordic Swan Ecolabel
certified provider of printed material.
Read more about our environmental
work at www.mediatryck.lu.se

MADE IN SWEDEN

To my parents Carmen and Oswaldo,
my brothers Luis and José,
and my amazing wife Paula.

Table of Contents

List of Papers

Paper I

Coral D.E., Fernandez-Tajes J., Tsereteli N., Pomares-Millan H., Fitipaldi H., Mutie P.M., Atabaki-Pasdar N., Kalamajski S., Poveda A., Miller-Fleming T.W., Zhong X., Giordano G.N., Pearson E.R., Cox N.J., Franks P.W. **A phenome-wide comparative analysis of genetic discordance between obesity and type 2 diabetes.** *Nat Metab* 5, 237-247 (2023).

https://doi.org/10.1038/s42255-022-00731-5

Paper II

Coral D.E., Fernandez-Tajes J., Pigeyre M., Chong M., Atabaki-Pasdar N., Fitipaldi H., Kalamajski S., Gomez M.F., Paré G., Giordano G.N., Pearson E.R., Franks P.W. **A precision medicine approach to coronary artery disease risk prediction and mitigation in people with type 2 diabetes**. Preprint (Version 1) available at Research Square (2023).

https://doi.org/10.21203/rs.3.rs-3470871/v1

Paper III

Coral D.E., Smit F., Farzaneh A., Gieswinkel A., Fernandez-Tajes J., Sparsø T., Delfin C., Bauvain P., Wang K., Temprosa M., De Cock D., Blanch J., Fernández-Real J.M., Ramos R., Gomez M.F., Kavousi M., Panova-Noeva M., Wild P.S., Adriaens M., van Greevenbroek M., Arts I., Frayling T.M., Giordano G.N., Franks P.W., Le Roux C., Pearson E.R., Ahmadizar F. **Identification of phenotypic profiles with different cardiometabolic risks for given levels of body mass index: an IMI SOPHIA study**. Manuscript.

Paper IV

Mutie P.M., Pomares-Millan H., Atabaki-Pasdar N., **Coral D.E.**, Fitipaldi H., Tsereteli N., Fernandez-Tajes J., Franks P.W., Giordano G.N. **Investigating the causal relationships between excess adiposity and cardiometabolic health in men and women**. *Diabetologia* 66, 321–335 (2023).

https://doi.org/10.1007/s00125-022-05811-5

Author’s contribution to the papers

Paper I

Coral D.E. and Franks P.W. conceived and designed the analyses and wrote the manuscript. **Coral D.E.**, Fernandez-Tajes J., Miller-Fleming T.W. and Zhong X. performed the analyses. All other co-authors contributed materials/analysis tools and reviewed the manuscript.

Paper II

Coral D.E. and Franks P.W. conceived and designed the analyses, and wrote the manuscript. Coral D.E. Fernandez-Tajes J. and Pigeyre M. performed the analyses. All other co-authors contributed materials/analysis tools and reviewed the manuscript.

Paper III

Coral D.E., Franks P.W. and Pearson E.R. conceived and designed the analyses. **Coral D.E.**, Smit F., Farzaneh A. and Ahmadizar F. wrote the manuscript. **Coral D.E.,** Smit F., Farzaneh A. and Gieswinkel A. performed the analyses. All other co-authors contributed materials/analysis tools and reviewed the manuscript.

Paper IV

Mutie P.M. and Franks P.W. were responsible for the conception and design of the study, and data interpretation; Mutie P.M. analysed the data and drafted the manuscript. Fernandez-Tajes J., Tsereteli N. and **Coral D.E.** prepared the data and run sensitivity analyses. All co-authors critically revised the manuscript and approved the final version.

Papers not included in this thesis

Huang M., Claussnitzer M., Saadat A., **Coral D.E.**, Kalamajski S., Franks P.W. **Engineered allele substitution at *PPARGC1A* rs8192678 alters human white adipocyte differentiation, lipogenesis, and PGC-1α content and turnover**. *Diabetologia* 66, 1289–1305 (2023).

https://doi.org/10.1007/s00125-023-05915-6

Huang M., **Coral D.E.**, Ardalani H., Spegel P., Saadat A., Claussnitzer M., Mulder H., Franks P.W., Kalamajski S. **Identification of a weight loss-associated causal eQTL in MTIF3 and the effects of MTIF3 deficiency on human adipocyte function**. eLife 12:e84168 (2023). https://doi.org/10.7554/eLife.84168

Coral, D.E., Franks, P.W. **Proteogenomic mapping sets stage for precision medicine**. *Nat Metab* 5, 366–367 (2023).

https://doi.org/10.1038/s42255-023-00759-1

Pomares-Millan, H., Atabaki-Pasdar, N., **Coral, D.E.**, 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 14*, 1218 (2022). https://doi.org/10.3390/nu14061218

Atabaki-Pasdar N., Pomares-Millan H., Koivula R.W., Tura A., Brown A., Viñuela A., Agudelo L., **Coral D.E.**, van Oort S., Allin K., Chabanova E., Cederberg H., De Masi F., Elders P., Fernandez Tajes J., et al. **Inferring causal pathways between metabolic processes and liver fat accumulation: an IMI DIRECT study**. Preprint available at medRxiv (2021).

https://doi.org/10.21203/rs.3.rs-3470871/v1

Abbreviations

AIS	Acute Ischemic Stroke
ALT	Alanine Aminotransferase
BC	Baseline concordant profile
BMI	Body Mass Index
CAD	Coronary Artery Disease
CKD	Chronic Kidney Disease
CRP	C Reactive Protein
CVD	Cardiovascular Disease
DAL	Discordant adverse lipid profile
DBP	Diastolic Blood Pressure
DHG	Discordant hyperglycaemic profile
DHT	Discordant hypertensive profile
DIS	Discordant inflammatory state profile
DLT	Discordant liver transaminase profile
eQTL	Expression Quantitative Trait Loci
FDR	False Discovery Rate
GWAS	Genome-Wide Association Study
HDBSCAN	Hierarchical Density Based Spatial Clustering of Applications with Noise
HDL	High Density Lipoprotein
HEIDI	Heterogeneity in Dependent Instruments
HTN	Hypertension
ICD	International Classification of Diseases
IDL	Intermediate Density Lipoprotein
IVW	Inverse Variance Weighted
LACE	Local average causal effect
LD	Linkage Disequilibrium
LDL	Low Density Lipoprotein

LRT	Likelihood Ratio Test
MACE	Major Adverse Cardiovascular Events
MAF	Minor Allele Frequency
MCV	Mean Corpuscular Volume
MODY	Maturity Onset Diabetes of the Young
MR	Mendelian Randomization
NRI	Net Reclassification Improvement
OR	Odds Ratio
PAD	Peripheral Artery Disease
PheWAS	Phenome Wide Association Study
PRS	Polygenic Risk Score
PTH	Parathyroid Hormone
RA	Rheumatoid Arthritis
RAAS	Renin Angiotensin Aldosterone System
RNA	Ribonucleic Acid
RR	Relative Risk
SBP	Systolic Blood Pressure
SCr	Serum creatinine
SCORE2	Second Systematic COronary Risk Evaluation
SD	Standard Deviation
SMR	Summary data based Mendelian Randomization
SNP	Single Nucleotide Polymorphism
sQTL	Splicing Quantitative Trait Loci
TG	Triglycerides
TSLS	Two Stage Least Squares
UMAP	Uniform Manifold Approximation and Projection
WHR	Waist to hip ratio

Chapter I – Introduction

Our bodies, as all living organisms, are constantly exchanging matter and energy with their surrounding environments. They do so through an intricate symphony of chemical reactions that we know as metabolism (a term derived from the Greek word *metabole* meaning “change”). Metabolic activity is crucial to maintain the delicate chemical equilibrium that sustains life, preventing its degradation into entropy (1). This activity occurs in pathways – chains of sequential chemical reactions that lead to the conversion of substances into final products – that govern the utilization of nutrients, the production of energy, and the detoxification of waste products, all of which underpin the optimal functioning of our bodies. The elucidation of these pathways has been the focus of profound scientific inquiry, unveiling the impact that disruptions, whether subtle or pronounced, can have on metabolic balance and the development of diseases.

Woven into this complex fabric of chemical reactions lie the pathways dedicated to energy metabolism, responsible for harnessing the energy stored in food to power our daily activities. Highly efficient mechanisms carry out the conversion of macronutrients – carbohydrates, proteins, and fats – into adenosine triphosphate (ATP), the cellular energy endpoint. These mechanisms are highly adaptable to changes in energy intake and demand, orchestrating the appropriate shifts in energy utilisation and storage – mainly as fat – to ensure that our bodies have the power they need to function optimally at all times (2).

The capacity of our metabolism to adapt to varying scenarios of food availability and energy demands has likely evolved over millennia as a buffer against fluctuating nutritional circumstances (3,4). However, over the last centuries our metabolism faced a fast and profound transformation in dietary and physical activity patterns, from our past as hunter-gatherers to agricultural and industrial revolutions. We are now exposed to abundant highly processed foods which increase energy availability than those consumed by our ancestors. Fast-foods like french fries pack on average 3 kilocalories per gram, much higher than most fruits, which range from 0.15 to 1 kilocalories per gram (5). Additionally, while it is estimated that our hunter-gatherer ancestors usually engaged in over 2 hours of moderate to vigorous physical activity per day (6), our current average is less than 30 minutes per day (7). Beyond total calories consumed and spent, changes in the composition of our diets, such as the content of micronutrients, as well as components causing high peaks in blood sugar after meals, produce powerful and lasting hormonal responses (8). The interplay of these factors has favoured the emergence of two highly concerning outcomes: chronic excessive fat accumulation causing obesity and impaired glucose metabolism causing diabetes.

Obesity

For much of our history, obesity was a sign of wealth and privilege (Figure 1.1) (9). Nonetheless, even in ancient times it was known that obesity could lead to health problems. Hippocrates, for example, already 400 years B.C., stated (10):

> "It is very injurious to health to take in more food than the constitution will bear, when, at the same time one uses no exercise to carry off this excess (…) For as aliment fills, and exercise empties the body, the result of an exact equipoise between them must be to leave the body in the same state they found it, that is, in perfect health".

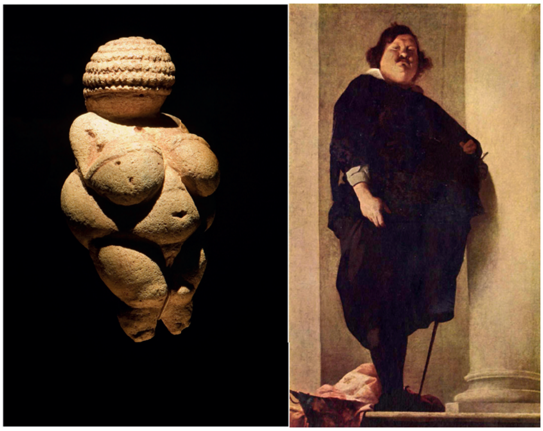


Figure 1.1. Obesity as a sign of wealth through history.
Left, the Venus of Willendorf, an 11cm figurine from 30,000 BC found in Austria. Right, The Tuscan General, by Alessandro del Boro, from about 1645. Source: Wikimedia Commons – Public domain.

This link between obesity and disease became progressively more evident with modernity and industrialisation. Of particular concern was the strong association between obesity and mortality observed in weight and height tables published by insurance companies at the beginning of the 20th century (11). This motivated the search for indices of relative weight that could be used as proxies of excess fat at the population level. In a seminal paper (12), Keys and colleagues in the 1970s revived an index devised more than a century before by Adolphe Quetelet, a Belgian polymath. In his quest to find the parameters defining the 'average man', he proposed measuring body size as the quotient of weight to squared height (13). Keys found that, aside from its simplicity, this index - named the body mass index (BMI) - had the lowest correlation with height, the highest correlation with body fat and the highest generalisability across populations compared to other measures. Due to these desirable properties, BMI quickly gained adoption within the medical community. By the end of the 20th century, international institutions such as the World Health Organization (WHO) embraced the BMI as the measure of choice for evaluating body corpulence within populations and its relation to disease within and between populations. Specific BMI thresholds were established to distinguish individuals with a 'normal' body size from individuals likely to be undernourished (BMI < 18.5 kg/m^2) and individuals likely to have excess fat with detrimental effects on health, categorised as having overweight (25 kg/m^2 $\leq$ BMI > 30 kg/m^2) or obesity (BMI $\geq$ 30 kg/m^2). These criteria were primarily informed by the population distribution of BMI and its correlation with mortality, which follows a J-shaped curve (14).

The widespread use of BMI enabled the global longitudinal surveillance of body size, revealing how changes in diet and lifestyle have impacted our bodies over the last decades. The global mean BMI has steadily increased from around 21 kg/m^2 in 1975 to close to 25 kg/m^2 in 2016 (15). While this has certainly led to a reduction the proportion of individuals with underweight, it has resulted in an increase in the proportion of individuals with obesity, now 3 times higher than the figures recorded in 1975 (16).

Longitudinal BMI surveillance also facilitated the quantification of the health burden arising from the striking rise in obesity rates. The share of total deaths attributable to obesity escalated from 4.7% in 1990 to 8.9% in 2019 (17). Likewise, the share of total years lived with disability globally due to obesity rose from 2.4% in 1990 to 4.75% in 2019 (17). Both higher mortality and disability impose tremendous economic pressures, not only increasing healthcare costs but also reducing work productivity. The costs attributed to obesity amounted to 2 trillion US dollars in 2020 (~2% of global gross domestic product, GDP) (18). If global BMI trends persist, it is projected that over 2 billion people around the world will be living with obesity by 2035, exposing a large portion of the population to the harmful effects of obesity, and representing global losses that would exceed 4 trillion US dollars, equivalent to ~3% percent of global GDP (18). It is this ominous trajectory that has compelled health authorities to declare obesity as a resounding global epidemic, calling for urgent action to curb its rise and mitigating its health impacts.

Type 2 diabetes

Like the trajectory of BMI, average blood sugar levels have been steadily increasing worldwide. Since 1980, blood glucose has risen around 0.08 mmol/L per decade (19). Although glucose is a fundamental source of energy in the human body, chronic hyperglycaemia can result in severe damage to multiple organs, with devastating consequences on health. Having glucose at levels where this damage is likely to happen is what we currently define as diabetes.

The term 'diabetes' was coined by Aretaeus of Cappadocia in the 2nd century A.D. to describe a rare syndrome of profuse urination, unquenchable thirst, and emaciation, all of which are manifestations of uncontrolled hyperglycaemia (20). Similar descriptions have been found in documents as old as Egyptian papyri from 1500 B.C. (21), demonstrating that diabetes is not a recent condition. The knowledge that these symptoms originate in derailments in sugar metabolism can also be dated back to writings from Indian physicians Sushruta and Chakara from 500 – 600 B.C., who associated these symptoms with sweet urine (22). Despite being widely known and studied over centuries, diabetes remained highly lethal and an effective treatment elusive until 1922, when a group of scientists from the University of Toronto, composed by Frederick Banting, Charles Best, John McLeod and James Collip, isolated insulin, the pancreatic hormone essential to promote glucose uptake from the bloodstream into cells (23). Nevertheless, around a decade later the British physician Harold Percival Himsworth observed that insulin administration worked best in younger individuals whose diabetes onset had been acute, but in older individuals who often had a more insidious onset and were more likely to have obesity, insulin was much less effective (24). It later became clear in that the former profile, termed 'type 1 diabetes' (T1D), was driven by autoimmune destruction of beta cells (located in the pancreas and the source of endogenous insulin). The second type, termed 'type 2 diabetes' (T2D), is mainly driven by resistance to insulin in peripheral tissues, and remains to this day the dominant type in the general population (23).

Consequently, along with the rising trends in blood glucose levels, diabetes prevalence has increased significantly over the past decades, from 3% in 1990 (~159 million individuals) to 6% in 2021 (~529 million individuals), and 96% of this increase has been driven by T2D (25). Moreover, despite advances in diabetes treatment, the share of total deaths attributed to diabetes has increased from 1.4% in 1990 to 2.8% in 2021 (25). Diabetes is also responsible for an increasing share of disability worldwide from 2.3% to 4.3% between 1990 and 2021 (25). The global cost of diabetes has risen from 200 million to 966 million US dollars during the same period (26). As the prevalence is expected to increase to ~10% by 2050 (over 1.3 billion individuals worldwide) (25), diabetes-related health expenditure is also projected to climb to >1 trillion US dollars (26).

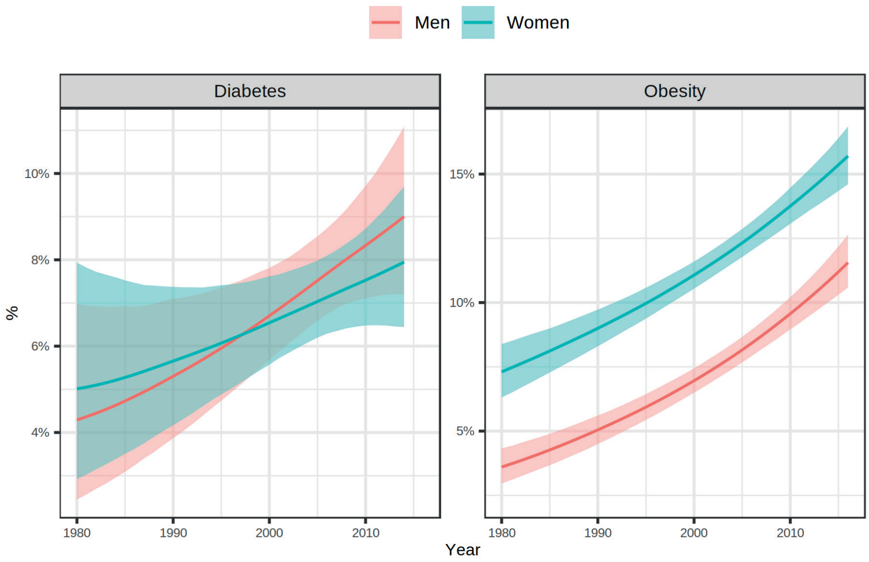


Figure 1.2. Global trends in prevalence of obesity and T2D.
Shaded areas are 95% confidence intervals. Source: NCD RisC Database.

Diabesity

People living with obesity have up to seven times higher risk of developing T2D compared to individuals without obesity (27) and around 60% of individuals with T2D have obesity (28). Due to the strength of the association between these two conditions, it has been proposed to refer to their co-occurrence as ‘diabesity’ (29), implying a causal pathophysiological link. The characteristic adipocyte hypertrophy in obesity causes stress, hypoxia, immune cell infiltration and fibrosis, shifting adipose tissue to a proinflammatory state. As a result, anti-inflammatory factors like adiponectin tend to decrease, while proinflammatory cytokines, free fatty acids and metabolically deleterious exosomes are released into the circulation, which interfere with insulin production and sensitivity (30). These processes favour ectopic lipid deposition, also adversely affecting insulin sensitivity, particularly in muscles and in intraabdominal organs like the liver, leading to an increase in gluconeogenesis (31). Ectopic fat in the pancreas can also in turn affect insulin secretion (32). Additionally, excess adiposity is linked to disruptions in signalling pathways of various hormones, such as leptin, a hormone produced by adipocytes with widespread neuroendocrine effects, including anorexigenic action on the hypothalamus, regulation of the pituitary, gonadal, thyroid and adrenal axes, as well as direct modulation of insulin sensitivity in peripheral tissues and insulin production by pancreatic beta cells (33).

In addition to the epidemiological and mechanistic links, interventions aimed to reduce weight in individuals with obesity are associated with a reduction in the risk of developing T2D. For example, the Diabetes Prevention Program showed that participants with overweight or obesity who underwent a diet and exercise intervention had a reduction in the risk of developing T2D of 58% in the short term (~3 years of follow-up) and 27% in the long term (~15 years of follow-up) compared to individuals who received placebo. Likewise, in the SOS study in Sweden, individuals with obesity without diabetes who underwent bariatric surgery had a 78% reduction in risk of incident T2D at 15 years of follow-up compared to individuals who did not opt for surgery (34).

Cardiovascular impacts of obesity and type 2 diabetes

The relationships between obesity, T2D and early mortality are primarily mediated through their association with coronary artery disease (CAD) and ischaemic stroke, the most prominent manifestations of cardiovascular disease (CVD). Compared to individuals with normal weight, individuals with obesity have on average a 30% higher risk of CVD mortality (35). Similarly, CVD is a common complication in individuals with T2D, who have around 4 times higher risk of dying from CVD than individuals without T2D (36). Over two thirds of deaths associated with obesity and half of those associated with T2D can be attributed to CVD (37,38).

The mechanisms linking obesity, T2D and CVD are multifaceted (Figure 1.3). The higher free fatty acid content in blood seen in obesity leads to an increase in triglyceride (TG)-rich lipoproteins, such as very low-density lipoproteins VLDL, with a subsequent reduction in high-density lipoprotein (HDL) particles, configuring a lipid profile that contributes to the formation of atherosclerotic plaques in vessel walls (39). Obesity is also associated with extracellular fluid volume expansion, increased blood flow and cardiac output, while impairing renal pressure natriuresis due to neuroendocrine activation of the renin-angiotensin-aldosterone system (RAAS) as well as physical compression of the kidneys by fat deposition, all of which increase blood pressure, adding more strain to blood vessel walls (40). This is further enhanced by the proinflammatory and prothrombotic state of obesity, which promotes endothelial dysfunction due lower nitric oxide bioavailability (41). Hyperglycaemia and hyperinsulinemia in T2D worsen endothelial inflammation and dysfunction and thrombotic risk (42), elevate blood pressure due to electrolytic imbalances and RAAS activation (43) and contribute to overproduction of VLDL in the liver (44). It is this multiplex of shared pathways that has been denominated the 'cardiometabolic' complex of diseases (45).

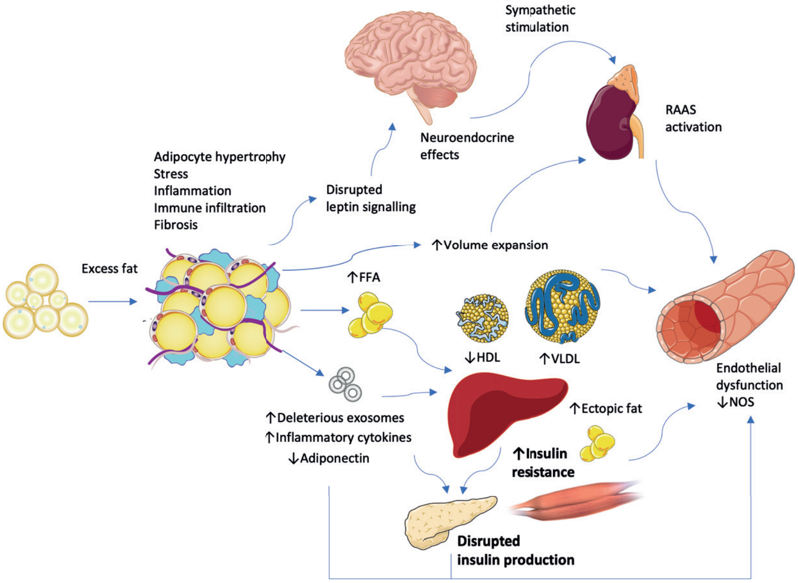


Figure 1.3. Mechanisms linking obesity, T2D and cardiovascular disease.
Excess fat leading to adipose hypertrophy causes an inflammatory cascade, release of free fatty acids and deleterious exosomes, and disruption in hormone signaling with widespread effects. Drawn with images from the public domain available at Wikimedia Commons and Servier Medical Art, and taken from Barilla et al (46).

Interventions aimed at reducing weight and glycaemia in individuals with obesity and T2D have demonstrated cardiovascular benefits. In the LookAHEAD trial, while an intensive lifestyle intervention did not show overall reduction in CVD after 9 years of follow-up (47), individuals who lost ≥10% of their body weight during the first year had a 20% reduction of CVD risk compared to those with stable weight after the intervention (48). The UKPDS study was the first to show that individuals with obesity and T2D who received metformin, an oral glucose-lowering agent with a modest weight loss effect, had a risk reduction of 32% of a composite outcome that included CVD and other diabetes-related microvascular complications, compared to the standard of care arm after 10 years of follow-up (49). Newer agents, particularly glucagon-like peptide 1 receptor agonists (GLP1RAs) significantly reduce weight, glycaemia and CVD risk in patients with obesity and T2D (50,51).

Heterogeneity and discordance in cardiometabolic diseases

The strong association of obesogenic diets and insufficient physical activity with cardiometabolic diseases has bolstered the notion that these conditions are the result of poor personal choices and lack of “willpower” (52). However, it has been demonstrated that these factors are not sufficient to produce the metabolic consequences observed in the population. In a landmark study, Sims and collaborators fed prisoners a diet of up to 10,000 kilocalories a day over a period of 40 weeks. While all participants gained weight, their metabolic rates also increased, causing almost all of them to return to their own initial weights by the end of the study (53). Similarly, losing weight through regimens of dietary energy restriction leads to a decrease in metabolic rate, again driving individuals to their initial weights (54).

Moreover, certain populations seem highly susceptible to weight gain and metabolic imbalances driven by lifestyle exposures, while others remain highly resilient. For example, the world’s highest prevalence of obesity occurs among inhabitants of islands in Micronesia and Polynesia, where it is 3 times the global average, reaching over 60%. In comparison, obesity prevalence is half in the neighbouring region of Melanesia (~30%), and further to the west in Indonesia, Malaysia and the Philippines, it is less than 10%, despite roughly similar changes in dietary patterns over time (15). Within more homogenous populations, such as cohorts in northern Sweden, distinct profiles of susceptibility and resilience have also been found, with the susceptible profile having significantly higher CVD risk compared to the resilient profile (55).

Moreover, while the relationship between BMI and metabolic dysfunction is generally positive, with an increase in BMI ‘concordantly’ corresponding to an increased risk of metabolic dysfunction, approximately 7% of individuals with obesity exhibit a ‘discordant’ profile, with no signs of metabolic dysfunction (56). Analogously, around 1 in 5 individuals within the normal range of BMI have a ‘discordant’ profile, characterised by the presence of multiple cardiometabolic risk factors despite their apparently healthy BMIs (57). Part of this discordance is because BMI is an imperfect proxy of the volume, distribution, and health of adipose tissue. For instance, individuals with similar BMIs can have vastly different adipose distributions: some individuals may have a higher proportion of visceral fat, which tends to promote inflammation and insulin resistance, and is strongly associated with CVD, while others may have a higher proportion of peripheral subcutaneous fat, which is generally more insulin sensitive, metabolically active and less prone to inflammation (58). This is also an important distinction between sexes, as men tend to accumulate more adipose tissue centrally, while females tend to accumulate adipose tissue around the hip and thighs (59).

It has also been also observed that in certain cases, especially following severe CVD, as well as other chronic diseases such as cancer, there is a paradoxical inverse association between BMI and mortality (60). Whether this phenomenon, named the "obesity paradox", reflects true biology or the result of selection bias in the populations where it has been found, is still a matter of debate (61).

There is also significant heterogeneity within T2D. The diagnosis of T2D is typically one of exclusion, leaving substantial inter-individual variability within the group that receive the diagnosis. This has motivated efforts to decompose T2D into subclasses that better reflect aetiology and risk of future complications. Data-driven analyses utilising various modelling strategies have consistently partitioned individuals into subgroups based on biomarker data at diagnosis, with significant differences in complications over time and differences in treatment response between the groups identified (62–64). For instance, these strategies have consistently identified T2D subgroups with phenotypic characteristics of severe insulin deficiency and resistance, which differ significantly in their disease evolution and treatment requirements after diagnosis.

As current prevention and treatment strategies for obesity and T2D are primarily based on average effects, the presence of significant heterogeneity can introduce errors in individualized care (65). It is therefore imperative to understand the origins of this heterogeneity and its clinical implications. This understanding may help refining existing strategies, while identifying mechanisms of disease susceptibility and resilience. Such insights have the potential to inform the development of new and more effective preventive and therapeutic targets to improve the management of cardiometabolic diseases (66,67).

Genetic factors in cardiometabolic diseases

An important explanation for why some individuals are at 'discordantly' higher or lower risk of cardiometabolic diseases is genetics. It has been known for a long time that both obesity and T2D run in families, in some rare cases with a strong Mendelian pattern of inheritance (68,69). Early evidence supporting the heritable basis of obesity in the general population came from the work of Stunkard et al in 1986 (70). They demonstrated that the BMI of Danish adopted adults approached the BMIs of their biological parents more than the BMIs of their adoptive parents. The same group of researchers also found, using a Swedish twin registry, that the BMIs of identical twins were highly correlated, irrespective of whether they were raised together or apart from each other (71). Around the same time, twin studies in the US reported that 58% of monozygotic twin pairs were concordant for T2D (both individuals had the diseases), and 65% of the twin co-pairs that were discordant (one of the pair did not have T2D) had hyperglycaemia (72).

Towards the end of the 20th century, advances in DNA sequencing led to the launch of the Human Genome Project, aiming to sequence the entire human genome. This endeavour propelled the search for the genetic determinants of diseases. Due to the high cost and difficulty of sequencing at the time, studies were restricted to few candidate genes with a priori evidence of an effect on obesity and T2D from experimental studies. Cases were also carefully selected, often with severe forms of disease, to increase the likelihood of finding the causal gene, but imposing limitations on the sample sizes. While these studies were frequently inconsistent across populations, they shed light on certain monogenic forms of disease, such as mutations in the leptin receptor *LEPR* and the melanocortin 4 receptor *MC4R* causing severe obesity (73,74), and mutations in the glucokinase gene *GCK* as one of the causes of maturity-onset diabetes of the young (MODY) (75).

The success of the Human Genome Project in delivering the first human genome sequence in 2006, together with progressive reductions in the cost and complexity of DNA sequencing as laboratory and computational techniques improved, helped unveil important elements of human genetic variation. The most common types of genetic mutations in the human sequence are single base-pair substitutions, known as single nucleotide polymorphisms (SNPs), the majority of which are biallelic, i.e., two alleles are segregated in the population. Hundreds of thousands of these SNPs covering many regions of the genome could be accommodated in microarrays enabling fast and accurate genotyping of many individuals.

Additionally, SNPs in physical proximity are often inherited together in what is known as a haplotype. This causes correlation of SNP alleles, a phenomenon called linkage disequilibrium (LD). This property was leveraged in imputation algorithms to accurately infer missing genotypes in regions of the genome adjacent to genotyped variants, effectively improving genotyping coverage.

Linking genotyped and imputed data to phenotypes gave rise to genome-wide association studies (GWAS). Merely one year after the publication of the human sequence by the Human Genome Project, in 2007, the first large-scale GWAS of BMI was published, pooling genetic data from more than 30,000 participants, and revealing that common variation in the *FTO* gene was associated with 70% higher odds of obesity (76). The first GWAS of T2D, also published that year, was conducted in a French cohort of over 5,000 individuals, finding four novel genetic loci robustly associated with T2D (77). Since these successful pioneering studies, the number of GWAS have increased exponentially, with increasing coverage of the genome and an exponential increase of sample sizes with many modern GWAS analyses incorporating data from millions of participants (Figure 1.4) (78). This trend has led to the discovery of hundreds of thousands of robust genetic associations, spanning many other clinical traits and phenotypes, including blood biomarkers, proteins, RNA expression in multiple tissues, and a plethora of other features (79).

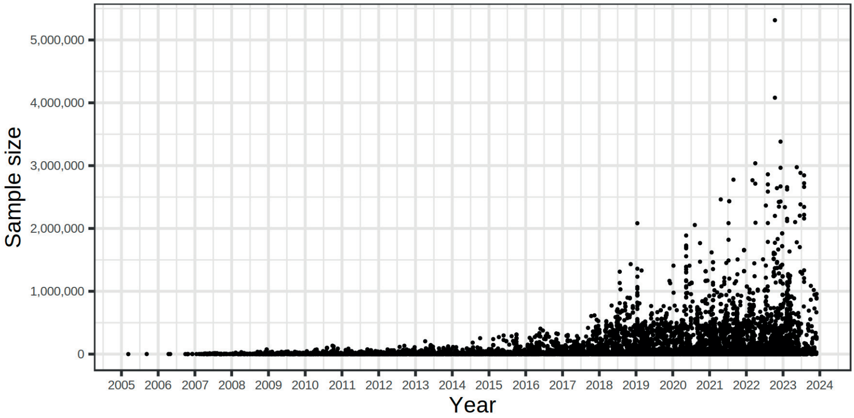


Figure 1.4. Trends in GWAS sample sizes over time.
Source: NHGRI-EBI GWAS Catalog.

GWAS have revealed that cardiometabolic diseases in the general population have an important heritable component, and this component is highly polygenic, with many genes contributing, each to a small degree, to the variance of the trait (79). Close to a thousand independent loci have been found to be associated with BMI (80), and over 300 associated with T2D (81). Together, these loci explain around 20% of the variance in BMI and T2D (80,82). These estimates are lower than heritability estimates from twin studies (the so-called 'missing heritability'), which could be explained by multiple factors, including additional common variants that are yet to be identified, non-additive (dominant or recessive) effects, the contribution of rare variants and non-genetic heritable factors and the inability of standard GWAS to account for gene-environment and gene-gene (epistatic) interactions (83).

Some of the loci identified in GWAS overlap with those associated with monogenic conditions, whose mechanisms are better understood, facilitating functional interpretation. An example of this is the *MC4R* gene, which has been associated with both monogenic and polygenic forms of obesity (84). However, most loci found in GWAS are located in non-coding areas of the genome (85). Bioinformatic strategies combining information from multiple sources, such as the association to other traits (a phenomenon called pleiotropy), and interaction with regulatory elements of nearby genes, has helped elucidate part of the biology linking these variants to disease (85). The mechanistic insights collected from these analyses have underscored the significant etiological heterogeneity of cardiometabolic diseases. For example, while some of the loci associated with BMI are likely to affect adipocyte biology, most of the associations link BMI with genes that predominantly act in the brain, particularly in the centres regulating appetite control (86). Additionally, while many genetic variants associated with BMI have deleterious consequences on glucose and lipid metabolism, following what is observed in observational data, some have rather been found to be associated with a favourable metabolic profile (87–90). Similarly, T2D-associated SNPs seem to cluster into groups acting through distinct pathways of insulin resistance or beta cell dysfunction and having different effects on complications (91,92), resembling the findings from phenotypic stratification previously exposed.

Purpose and aims

The global increase in cardiometabolic conditions, especially obesity and T2D, presents substantial health challenges. This is further complicated by the considerable heterogeneity among individuals with these conditions, with individuals at discordantly higher or lower risk, affecting the delivery of appropriate care. The purpose of this thesis is to improve our understanding of the heterogeneity in cardiometabolic conditions through the application of genetic analyses and machine learning techniques to identify these discordant profiles and provide robust insights into the mechanisms that give rise to this discordance, as well as its potential clinical implications. The aims of the papers included in this thesis are:

- In Paper I, I applied a genetic approach to identify two obesity profiles that are either ‘concordantly’ associated with higher T2D risk or ‘discordantly’ associated with protection against T2D. Then we conducted an agnostic phenome-wide comparison using various machine learning techniques to uncover the most prominent differences that distinctively characterize these profiles using data from various cohorts. The traits in which we found differences were taken forward to causal inference analyses to determine the causal relationships underlying discordant obesity.
- In Paper II, I reapplied the approaches used in Paper I to identify and characterise two genetically determined diabetogenic profiles that are either ‘concordantly’ associated with higher risk of CVD or ‘discordantly’ associated with protection against CVD. We additionally

assess the potential contribution of concordant and discordant profiles in CVD risk prediction.

- In Paper III, I designed a graph-based clustering approach to visualise and decompose the general population into profiles that represent phenotypic 'discordance' deviating from the 'concordant' linear relationship between clinical measures and BMI. We explored the characteristics of these profiles and their potential clinical implications in cardiometabolic risk in four large independent cohorts across Europe.
- In Paper IV, we explored the overall and sex-specific effects of BMI on T2D, CVD and multiple related biomarkers, and described the shape of these causal effects using linear and non-linear Mendelian randomisation techniques.

Chapter II – Data sources

GWAS databases

Human genetics has played a pioneering role in open science and data sharing mechanisms that improve reproducibility. The ability to share summary results of GWAS, without compromising participant privacy has been a cornerstone of this progress. This has enabled researchers to combine results from multiple cohorts to improve statistical power to discover genetic associations as well as shared genetic predisposition across various traits. Various databases have taken on the challenge of combining genetic data generated using different genotyping and processing methods, providing harmonised datasets that are publicly available and facilitate downstream analyses. Although the format of these datasets is not uniform, they typically consist of tables of millions of rows, each containing the effect estimate of a SNP on the trait being analysed, its accompanying standard error and p-value, as well as other useful information, such as allele frequency (93).

The GIANT consortium database

The Genetic Investigation of ANthropometric Traits (GIANT) consortium is the largest international collaboration focusing on genetic variation associated with human body size and shape. Established in 2007, the consortium aims to unravel the genetic basis of height, weight, and related traits through GWAS meta-analyses. Summary statistics generated from their publications are available online (https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium). We used genetic associations to BMI that were found in the latest and largest meta-analysis, that included over 700.000 individuals of European ancestry.

The DIAGRAM consortium database

The Diabetes Genetics Replication and Meta-analysis Consortium (DIAGRAM), formed in 2009, is the largest international collaborative effort focused on GWAS meta-analysis to uncover genetic factors contributing to T2D. Their summary results are also available online (https://www.diagram-consortium.org/) and was the source of genetic associations to T2D. We used the associations found in meta-analysis of cohorts of European ancestry, which had the largest sample size (around 900.000, of which 9% were cases), thus aligned to the ancestry of the GWAS of BMI and better powered to detect concordant and discordant associations.

The NHGRI-EBI GWAS Catalog

The NHGRI-EBI GWAS Catalog is a harmonised and curated collection GWAS results produced by a collaboration between the National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (EBI). It has provided data from published GWAS since 2008 and has been redesigned and relocated to EMBL-EBI in 2015. The new infrastructure includes a graphical user interface (www.ebi.ac.uk/gwas/), ontology-supported search functionality, and an improved curation interface. It is accessible via the website, an application programming interface (API), and R and Python packages to facilitate integration with third-party analytical tools. The GWAS Catalog contains publications, top associations, and full summary statistics. We used this database to identify the sources of genetic associations to CVD that were included in Paper II.

The Open GWAS database

The OpenGWAS database is a resource developed at the MRC Integrative Epidemiology Unit at the University of Bristol that provides a manually curated and harmonised collection of complete genome-wide association study (GWAS) summary datasets. The database currently contains over 125 billion genetic associations from more than 14,500 complete GWAS datasets, representing a range of different human traits and diseases. It is accessible via a website, an application programming interface (API), and R and Python packages, designed to allow programmatic access to GWAS summary data for thousands of traits simultaneously. This facilitates phenome-wide analyses for single or multiple SNPs, which we used in our comparative analyses of concordant and discordant profiles (94).

The MiBioGen consortium

The MiBioGen (Microbiome Genome) Consortium is an international collaborative initiative aimed at studying the influence of human genetics on the gut microbiota. It currently comprises 18 cohorts worldwide, with a total of 19,000 participants of predominant European origin (95). The consortium has standardised the analytical pipelines for both microbiota phenotypes and genotypes across the cohorts. This was the source of the concordant and discordant associations to gut microbiome in Paper I and II.

The GTEx database

The Genotype-Tissue Expression (GTEx) database is a comprehensive resource that provides insights into the relationship between genetic variation and gene expression across multiple human tissues. As most genetic associations found in GWAS lay in non-coding regions of the genome, the aim of developing this database is to bridge GWAS discoveries to genes, as a step to uncover the functional mechanisms of genetic variation. It incorporates data from 17,000 post-mortem samples across 54 tissue sites donated by over 900 individuals. Around 85% of participants were of European origin and 66% were male (96). We used this database to assess the effect concordant and discordant SNPs on the expression of nearby genes (±1 million base pairs [Mb] upstream and downstream in the genome). We used the 8th version of this database, which also includes genetic effects on splicing of nearby genes. Genetic variants with such effects are called expression and splicing quantitative trait loci, or eQTL and sQTL, respectively. Additionally, we used data from GTEx to provide evidence of target genes with potential discordant effects in cardiometabolic conditions.

The eQTLGen consortium

The eQTLGen is a database that integrates data from multiple large-scale studies of predominantly European ancestry to explore the genetic determinants of gene expression in whole blood (97). It incorporates a total of over 31,000 individuals. We used these data to complement our analysis of discordant cis-eQTL effects as done in GTEx.

Additional genetic analysis tools and databases

In addition to GWAS databases, I used existing tools and databases for genetic analyses and functional annotation of genetic variants and genes, as well as to aid in the construction of polygenic scores:

The 1000 Genomes database

The 1000 Genomes Project is a comprehensive international collaboration that aimed to create a catalog of human genetic variations by sequencing the genomes of a large number of individuals from different populations worldwide (98). Launched in 2008, its primary objective was to capture genetic diversity across various ethnic groups. The genotypes are publicly available, which I used mainly for LD calculations, as well as for quality control procedures of GWAS data, as explained in the methods section.

The Roadmap Epigenomics Project

The Roadmap Epigenomics Project is a collaborative initiative launched in 2008, with objective of systematic mapping of epigenetic marks, including DNA methylation, histone modifications, and chromatin accessibility, across a wide range of human cells (99). By providing an extensive database of epigenomic data, this project contributes to the understanding of regulatory mechanisms of the human genome. The integration of epigenomic information to GWAS results can therefore provide mechanistic insights of genetic associations. We integrated concordant and discordant SNPs to epigenetic data using this database to attempt to locate the most likely tissue of action, as is described later in the methods section.

DEPICT

DEPICT (Data-driven Expression Prioritized Integration for Complex Traits) is an analysis tool written in the Python programming language that helps pinpoint the genes, pathways, and tissues likely underlying the associations derived from a GWAS trait (100). It integrates data from multiple sources to map significant SNPs to one or multiple genes according to proximity, potential consequences, epigenetic interactions, effects on expression and protein interactions. The results are then used to identify functional pathways, tissues and cell types that are potentially implicated in the mechanisms linking genetic variation to the outcome under study. I used this tool in Papers I and II to characterise concordant and discordant genetic variation.

Drug-gene interaction databases

In Paper I we used two databases to identify drugs that could potentially interact with the genes identified in our analysis: the Drug-Gene Interaction database (DGIdb) (101), developed by researchers at the Washington University School of Medicine, and the PHAROS database, maintained by the US National Institutes of Health (NIH) (102). In both cases, lookups were performed using the web-based tool available online. DGIdb assigns an interaction score to the drug–gene interactions, which is the result of combining publication count, source count, relative drug specificity and relative gene specificity. This score allows researchers to gauge the confidence level associated with each potential interaction. The PHAROS database categorizes gene targets into four ‘Target Development Levels’ based on the available evidence of drug interactions:

- ‘Tdark’ encompasses understudied targets with limited information.
- ‘Tbio’ includes well-studied targets lacking known interactions with compounds.
- ‘Tchem’ includes targets known to bind to small molecules.
- ‘Tclin’ comprises targets with interactions involving approved drugs.

The PGS Catalog

The PGS Catalog is an open database of polygenic scores, which represent the genetic predisposition for a certain trait by aggregating multiple genetic variants associated with the trait found in GWAS (more details about polygenic scores in the Analytical methods section). The database stores the genetic variants contained in each polygenic score, which can be downloaded and used to compute scores in a population with genotype data. The database also provides metadata of the polygenic scores, such as trait information, sample description, performance metrics, and publication details, to ensure reproducibility and accurate application in biological and clinical research. It is accessible via the website, an application programming interface (API), and a Python package, which facilitate analytical integration (103).

Cohorts

UK Biobank

The UK Biobank is one of the largest ongoing prospective studies in the world, and it is the main cohort we used in our analyses, particularly for discovery. Recruitment run from 2006 to 2010, with postal invitations for participation sent to over 9 million individuals aged 40–69 years who were registered in the UK National Health Service and who resided at least 40 kilometres from one of the 22 participating assessment throughout England, Wales and Scotland. Approximately 5.5% (around 500,000 individuals) of the total invited attended the assessment centres and consented to participate. Participants are more likely to be female, older, healthier, and more affluent than nonparticipants. Data collected includes comprehensive genetic and phenotypic information, biochemical assays, and longitudinal health outcomes through health records, such as hospitalisation and mortality (104,105).

In our analyses we used anthropometric and biomarker data collected at recruitment, and we used hospitalisation and mortality records to derive hazard estimates for our outcomes of interest, including data from 10 years of follow-up. We also used genotyping data, which was generated using the UK Biobank Lung Exome Variant Evaluation and the Applied Biosystems UK Biobank Axiom Array. Genotype imputation was performed using the Haplotype Reference Consortium (HRC) panel.

Approximately 84% (N = 409,512) of the population in the UK Biobank who were genotyped were from European origin according to genetic principal component projections. We used this subset in our main analyses to ensure high statistical power while mitigating confounding due to population structure. In the context of our analysis of genetic discordance, this choice also ensured consistency with the European origin of the datasets used to identify concordant and discordant genetic variation. Additionally, only individuals who were part in the calculation of genetic principal components were included, which ensures minimal genetic kinship with other participants that could bias our results. We also excluded individuals with inconsistency between their reported and genetic sex, had sex chromosome aneuploidy or were outliers for heterozygosity or missingness. Population structure was further adjusted for in our regression analyses that included genetic data by adding the first ten genetic principal component as covariates.

BioVU

BioVU is an electronic health records (EHR) database established in 1990 in the Vanderbilt University Medical Center. It includes data on billing codes from the International Classification of Diseases, 9th and 10th editions (ICD-9 and ICD-10). Disease phenotypes ('phecodes') are derived from these billing codes and case, control and exclusion criteria are defined (106). Two codes on different visit days were required to instantiate a case for each phecode. It also includes various laboratory measurements taken during clinical care. EHR data is anonymised and linked to a biobank launched in 2007, which comprises excess blood samples that their donors had consented for use in biomedical research. This consent is obtained using an "opt-out" approach, with around 5% refusing to participate, and favouring fast data collection (500-1000 samples per week), reaching over 300,000 individuals in 2023 (107,108).

We used a data freeze that included over 48,000 individuals of European descent with genetic data linked to presence or absence of disease phenotypes, as well as 68,000 European and 14,000 African descent individuals with genetic data linked with laboratory measurements. These data were used to replicate the comparative analyses between BMI-T2D concordant and discordant genetic profiles. DNA samples were analysed using genome-wide genotyping platforms including Illumina multi-ethnic genotyping array. Genotype imputation was performed at the Michigan imputation server using the HRC reference panel. Populations of African American and European descent were identified by projecting individuals onto the major principal-component space derived from 1000 Genomes reference panel.

ORIGIN

The ORIGIN trial is an international multicentre randomized controlled trial that recruited participants from multiple ancestries with impaired fasting glucose, impaired glucose tolerance, newly detected diabetes, or established diabetes. They were also required to have documented manifestations of CVD, such as history of prior myocardial infarction or stroke, coronary or peripheral stenosis detected in angiography, left ventricular hypertrophy or albuminuria (109). Participants were then randomized to basal insulin titration with insulin glargine or placebo, and to polyunsaturated fatty acid supplementation or placebo.

The genotypes were assayed using the HumanCoreExome BeadChip 12 v1.0 and v1.1 from Illumina and imputed to the TOPMED reference panel. Population structure was assessed using principal-component analysis. We used data from European (and Latin American ancestry in paper II to measure the association of T2D-CVD concordant and discordant profiles to major adverse cardiovascular events (MACE) which was defined as fatal and non-fatal myocardial infarction or stroke. Participants were followed up for up to 7 years.

Cohorts in the SOPHIA consortium

Our analysis of phenotypic discordance to the expected for the BMI that is described in Paper III is the collaborative product of one of the working groups that compose the Stratification of Obesity PHenotypes to optimise future therapy (SOPHIA) consortium, an IMI project that aims to identify and characterise clinically meaningful subpopulations of patients with obesity (110). We proposed this analysis plan and performed the discovery analysis in the UK Biobank, and then we sought replication in three datasets that have been collected by partner institutions in SOPHIA:

- **The Maastricht Study** is an observational population-based cohort study based in the south of Limburg in the Netherlands, that aims to investigate the causes, consequences, and prevention of type 2 diabetes, cardiovascular disease, and other chronic conditions. It is enriched to contain 25% of participants with type 2 diabetes. It contains comprehensive demographic, biological, social, health, lifestyle, cognition, and mental health data. The study is a collaboration between Maastricht University, Maastricht University Medical Center, and the regional health authorities (111).
- **The Rotterdam Study** is a population-based cohort study conducted in the Ommoord district of Rotterdam, The Netherlands, with the primary objective of assessing common diseases among the elderly population. The study, which has been extensively documented (112) recruited 7983 individuals aged 55 years or older for the initial RS-I cohort in 1990. Subsequently, in 2000, the RS-II cohort was expanded by 3011 participants who either relocated to the study area or reached the age of 55. The cohort was further extended with 3932 participants aged 45 years

or older (RS-III). Baseline evaluations were conducted through home interviews and comprehensive physical examinations at the time of recruitment, followed by subsequent visits every 3-4 years for follow-up assessments. We included longitudinal outcome data up to 10 years after recruitment.

- **The Gutenberg Health Study** is a prospective and observational adult population-based cohort study in the Mainz-Bingen region of Rhine-Palatine in Germany. The study sample consisted of 15,010 participants aged 35-74 years who were enrolled at their baseline examination between 2007 and 2012. Each study participant underwent a comprehensive standardized clinical and laboratory examination at enrolment. We included follow-up outcome data up to 5-year after recruitment (113).

Chapter III – Analytical methods

Effect of variables on outcomes using regression

Regression is a flexible statistical framework that aims to understand how an outcome of interest changes as one or multiple variables vary. This framework can be adapted to different natures and distribution of the outcome of interest and can be used both for inference and prediction. It has been fundamental to understand the relationships across factors in cardiometabolic conditions, as well as understanding the effect of genetics, and it is therefore essential for this thesis. We describe here three modalities of regression we used in our analysis.

Linear regression

Linear regression is a statistical method used to model the relationship between a single or multiple variables and an outcome of continuous nature. The goal of linear regression is to find the best fitted line (or hyperplane) that describes the relationship between the variables and the outcome. This is defined by an equation of the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon \qquad (1)$$

Where y is the dependent variable, x_1 to x_m are the values of m independent variables, β_0 is the intercept (the value of y when the values of x_1 to x_m are 0) and β_1 to β_m are the coefficients, each representing the expected change in the dependent variable y for a unit change in the corresponding independent variable having the remaining fixed at their mean value. The residual variance term ε is expected to follow a normal distribution around 0, denoted $\mathcal{N}(0, \varepsilon)$.

The best set of coefficients is found using the least-squares method, which consists of minimizing the sum of the squared differences between the observed and predicted values of the dependent variable.

Linear regression can be used for predictive analysis, as it allows us to make predictions about the dependent variable based on the values of the independent variables. It is a popular statistical tool due to its easy implementation, interpretability, and scalability (114).

Logistic regression

Logistic regression extends linear regression to model the relationship between single or multiple variables and the occurrence of an event. To achieve this, the fitting procedure is performed on a transformed data space using the logit function, which is defined as the logarithm of the odds of an event y occurring vs not occurring:

$$\mathrm{logit}(y) = \log\left(\frac{y}{1-y}\right) \quad (2)$$

This transforms the outcome from a scale that is constrained between 0 (no event) and 1 (event) to an unconstrained scale (from minus to positive infinity). The independent variables are then assumed to be linearly associated to the outcome in this transformed space, similar to linear regression:

$$logit(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \quad (3)$$

As a consequence of Equations 2 and 3, the exponentiated value of the coefficients represent the relative increase or decrease in the odds of an event happening for every unit increase in the corresponding independent variable.

The expected probability of an event given the model coefficients can be obtained from this model using the inverse of the logit function:

$$E(y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_m x_m}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_m x_m}} \quad (4)$$

This function is used to find the values of the coefficients that best fit the data using maximum likelihood estimation. This consists of comparing the predicted probabilities $\mathrm{E(y)}_i$ to the observed event y_i for an individual i by calculating the log-likelihood, which is defined as:

$$\log(\mathrm{L}_i) = y_i \cdot \log(\mathrm{E(y)}_i) + (1 - y_i) \cdot \log(1 - \mathrm{E(y)}_i) \quad (5)$$

The sum of the log-likelihoods of all individuals is the model log-likelihood, which represents the accuracy of the model predictions. This value is maximised by updating the estimates using an iterative optimisation algorithm until convergence is reached, which means that there is no further gain by updating the coefficients. Several optimisation algorithms exist, although a detail explanation of them is beyond the scope of this thesis (115).

Cox regression

Cox regression is a statistical method used in the analysis of survival data, where the occurrence of an event of interest is recorded during a certain follow-up time. The aim is to investigate the effect of several variables on the time it takes for the event of interest to happen. This is commonly expressed as the hazard function:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t<T\leq t+\Delta t)}{\Delta t} \qquad (6)$$

Here, $P(t < T \leq t + \Delta t)$ is the probability of the event occurring at a particular time within a small interval between t and Δt, provided it has not happened before. It can be also interpreted as the instantaneous event rate at time T.

In a Cox model, the hazard function is linked with the predictor variables through the following equation:

$$h(t) = h_0(t) \cdot e^{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m} \qquad (7)$$

Which means that:

$$log\left(\frac{h(t)}{h_0(t)}\right) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m \qquad (8)$$

Thus, like in logistic regression, the exponentiated coefficients of this model, indicate the relative increase or decrease in the risk of an event at any point in time for a one-unit change in the corresponding covariate. This reflects a key assumption of the Cox model, known as the proportional hazards assumption, which implies that, irrespective of the baseline hazard function $h_0(t)$, the relative effects of covariates remain constant over time. The baseline hazard is therefore left unspecified, rendering the Cox model semi-parametric.

Coefficient estimation in Cox regression is achieved using the maximum likelihood optimisation approach in a similar manner to logistic regression. However, the likelihood function differs because for some individuals the event is not recorded, and hence it is not known if and when the event occurred for that individual, a feature of survival data known as censoring. To handle this, the likelihood in a Cox model is computed at every time t_i when an event occurs to an individual i. This likelihood is the ratio of the expected hazard for the individual who experienced the event $h(t_i)_i$, over the sum of the expected hazards of the set R_i composed by all individuals $j \in R_i$ who are still at risk at that particular time point (individuals who have not yet experienced the event):

$$L(t_i) = \frac{h(t_i)_i}{\sum_{j\in R_i} h(t_i)_j} = \frac{\mathrm{e}^{\beta_1 x_{1\mathrm{i}} + \beta_2 x_{2\mathrm{i}} + \ldots + \beta_m x_{mi}}}{\sum_{j\in R_i} \mathrm{e}^{\beta_1 x_{1\mathrm{j}} + \beta_2 x_{2\mathrm{j}} + \ldots + \beta_m x_{mj}}} \tag{9}$$

Additional adjustments are commonly done when two or more individuals experience an event at the same time, though without altering the general form of this function, known as the partial likelihood. The sum of the logarithm of the likelihoods over all times is the log-likelihood of the model, a measure of how close the model fits the data (116).

Pooling effects using meta-analysis

Meta-analysis is a statistical technique used to synthesise evidence from multiple independent studies that attempt to estimate a certain relationship between a variable and an outcome. It typically consists of combining effect estimates obtained from regressions run separately within each study into a single overall estimate. This has been fundamental to increase statistical power and precision in the estimated effects of risk factors in cardiometabolic conditions and has played a pivotal role in GWAS.

The most common meta-analytical method is the inverse-variance weighted (IVW) method (117). The core assumption in IVW meta-analysis is that every effect estimate $\hat{\theta}_k$ from study k is an estimate of the true effect estimate θ, and that better estimators have smaller standard errors $SE^2_{\hat{\theta}k}$, which reflect less sampling error. Thus, the best estimator of the true effect, $\hat{\theta}$, is a weighted average of $\hat{\theta}_k$, with weights proportional to the inverse of the standard errors:

$$\hat{\theta} = \frac{\sum_{k=1}^{K} \hat{\theta}_k \cdot w_k}{\sum_{k=1}^{K} w_k}, where\ w_k = \frac{1}{SE^2_{\hat{\theta}k}} \tag{10}$$

$$SE_{\hat{\theta}} = \frac{1}{\sqrt{\sum_{k=1}^{K} w_k}} \tag{11}$$

These equations implicitly assume that $\hat{\theta}$ is fixed, and that differences between $\hat{\theta}_k$ are only due to sampling error, known as the fixed effects model. This assumption is relaxed in a random effects model, where heterogeneity across studies is introduced by adding the between-study variance τ^2 to the weights in Equation 10:

$$w_k = \frac{1}{{s_k}^2 + \tau^2} \tag{12}$$

Several methods for estimating τ^2 exist, thoroughly reviewed in (118).

Cross-trait GWAS and genetic profiles

GWAS cross-referencing

In Papers I and II we were interested in identifying genetic variants strongly associated with two phenotypes (BMI and T2D in Paper I, T2D and CVD in Paper II) with coefficients that had either the same or opposite sign, representing the concordant and discordant genetic variation, respectively. To find these variants we cross-referenced GWAS summary data, for which we followed a quality control procedure:

- First, we ensured that GWAS sources used the same version of the human genome assembly to locate genetic variants and SNP identifiers – generally to the hg19 genome assembly.
- We only included biallelic SNPs as these are the most common type of mutations and are easier to cross-reference across studies and have generally fewer genotyping errors compared to other types, such as insertions, deletions and multiallelic SNPs.
- We only included SNPs with minor allele frequency (MAF) higher than 1% as rarer variants are generally more prone to genotyping errors, as well as higher errors in their effect estimates due to their low population frequency.
- We excluded palindromic SNPs with a MAF higher than 40%. This is because the two alleles of a palindromic SNP are complementary bases that can pair with each other in the double helix structure of DNA. This hinders the assessment of directional concordance or discordance, especially as MAF approaches 50%, because we would not know which of the two DNA strands was measured in the studies, which might differ across genotyping chips.

After this quality control procedure, we merged the GWAS summary statistics using their genomic location and SNP identifiers and checked that the two alleles matched. We removed SNPs in which the difference between MAF between studies was more than 20%. We also checked that these SNPs matched to SNPs in the 1000 Genomes reference panel. Finally, we aligned all effects to the BMI increasing allele in Paper I and to the T2D risk allele in paper II.

Cross-trait GWAS strategies and profile assembly

To identify SNPs strongly associated with both conditions, in Paper I we used a simple strategy: we selected variants with p-values for both conditions that reached the standard genome-wide significance threshold ($p < 5 \cdot 10^{-8}$). After recognising that this approach is too stringent, leading to loss of power to detect important discordant genetic variation, in Paper II we improved upon this strategy by applying a cross-trait GWAS approach. We used a method called C-GWAS, which optimises statistical power to detect SNPs with effects on the two traits by having into account their genetic correlation (119). This method discriminates true correlation of genetic effects Π from a background correlation Ψ that affect genetic effects and their correlation and can itself be affected by cryptic relatedness or population stratification. These estimated correlations are then used to combine the test statistics of each SNP from both GWAS into a single test statistic that reflect the deviation from the null hypothesis that the SNP is not associated with any trait. The process to obtain these estimates includes various optimisation algorithms achieved through multiple runs and systematic SNP sampling that can be extended to the analysis of multiple traits simultaneously. The following are the fundamental steps to combine two GWASs, which was the use case in this thesis:

- To accurately estimate Ψ, first each GWAS undergoes a genomic control step, where the observed test statistics are modelled in an iterative optimisation algorithm as a function of an inflation factor I, the proportion of SNPs with true effects p and an expected normal distribution of these true effects around 0 with variance Λ:

  $$\mathrm{E}(T^2) = 1 + I + p\Lambda \tag{13}$$

  The test statistics from each GWAS are then adjusted by the estimated inflation factor from Equation 13:

  $$T_{adj} = \frac{T}{\sqrt{1+I}} \tag{14}$$

- An initial estimate of Ψ between the two GWAS is calculated from the corresponding sets of adjusted test statistics $\mathbf{T}_1$and $\mathbf{T}_2$ from Equation 14:

  $$\Psi_0 = cor(\mathbf{T}_1, \mathbf{T}_2) \tag{15}$$

  A combined T-statistic for each SNP is computed using this correlation estimate:

  $$CT_{SNP} = (T_{1,SNP}, T_{2,SNP})\boldsymbol{\Psi_0}^{-1}(T_{1,SNP}, T_{2,SNP})^T \tag{16}$$

  Where $\boldsymbol{\Psi_0}$ is the 2x2 correlation matrix with 1 in the diagonal and Ψ_0 in the off diagonal. The set of SNPs with a CT_{SNP} equivalent to an F-statistic lower than a critical threshold (0.5) are considered unlikely to be associated to the traits under the joint distribution of GWAS:

  $$id = which(F_{\chi_2^2}\mathbf{CT} < \mathbf{0.5}) \tag{17}$$

This subset of SNP is then used to compute Ψ_f, an update from Ψ_0:

$$\Psi_f = cor(\mathbf{T}_{1,id}, \mathbf{T}_{2,id}) \tag{18}$$

This new estimate Ψ_f replaces Ψ_0 and Equations 15 through 18 are reiterated until there is no further gain in updating (i.e., $|\Psi_f - \Psi_0| \approx 0$), with the final Ψ_f being the best estimate of Ψ.

- An overall effect covariance matrix **H** is obtained by subtracting $\mathbf{\Psi}$ from **Z**, the observed covariance matrix of T-statistic. The effect correlation matrix $\mathbf{\Pi}$ is derived from **H**, with 1 in the diagonal and Π in the off diagonal.
- These matrices are used to calculate the single combined association statistic for each SNP using an approach called effect-based inverse-covariance weighted meta-analysis (EbICoW), an extension of the random effects IVW technique previously described. To illustrate this, in IVW the T-statistic of the overall estimate T_C, derived from Equations 10 to 12 and reformulated as a function of $\mathbf{w}$ and $\mathbf{t}$, the vectors containing the weights and T-statistics, is:

$$T_C = \frac{\sqrt{\mathbf{w}}^T \mathbf{t}}{\sqrt{\mathbf{w}^T \mathbf{1}}} \tag{19}$$

In EbICoW the T-statistic is additionally weighted by the matrices derived from the previous steps, thereby considering the covariance structure of the estimates, and improving statistical power:

$$T_C = \frac{(\mathrm{sign}(\Pi - \Psi)\boldsymbol{H} \circ \boldsymbol{w^T})\ \mathbf{\Psi}^{-1}}{\sqrt{(\mathrm{sign}(\Pi - \Psi)\boldsymbol{H} \circ \boldsymbol{w^T})\ \mathbf{\Psi}^{-1}(\mathrm{sign}(\Pi - \Psi)\boldsymbol{H} \circ \boldsymbol{w^T})^T}}\boldsymbol{t} \tag{20}$$

In line with other cross-trait GWAS methods, the estimated values of Ψ and Π for a given GWAS pair are assumed to be shared across all SNPs, and therefore remain unchanged in the computation of T_C for each SNP. However, this approach also accommodates the potential variability in the covariance structure across SNPs by allowing the matrix **H** to change during the computation for each SNP depending on its joint strength of association to the traits. This feature make this approach more flexible than other methods such as MTAG (120), and was the reason for its selection.

In Paper II, to consider a SNP to be associated with both traits, the univariate p-value for each trait had to be less than $2.23 \cdot 10^{-4}$, equivalent to the value of two joint tests reaching $5 \cdot 10^{-8}$, the standard genome-wide significance. Additionally, the p-value derived from the C-GWAS method had to be less than the genome-wide significance.

We then performed clumping of the signals identified, a procedure that identifies the most significant SNP (i.e., lowest p-value) in each block of the genome that is in LD. To clump we used a threshold for LD r^2 (a number from 0 to 1 indicating correlation of alleles between two SNPs) of 0.01 over a 500kb window. To assemble the concordant and discordant profiles, we classified these independent SNPs as concordant or discordant according to their positive or negative direction of association to T2D in Paper I and to CVD in Paper II.

Phenome-wide comparative analyses

Phenome-wide scan of concordant and discordant effects

After identifying SNPs within the concordant or discordant genetic profiles, we were interested in finding what are the main phenotypic characteristics that differ between these two profiles, which would give us clues on how discordance emerges. To accomplish this, we exploited the ability to combine genetic association from multiple sources using a Phenome-Wide Association (PheWAS) framework. A PheWAS is an orthogonal application of GWAS, in which the aim is to investigate the associations between genetic variants and a wide range of phenotypes. These pleiotropic effects can provide mechanistic information linking genetic variation to disease (106).

We performed a PheWAS of concordant and discordant SNPs using the OpenGWAS database through its R package. We followed a similar quality control procedure as the GWAS cross-reference to ensure SNP and allele matching. Additionally, in cases where the effect of a SNP on a trait was not found, we looked for the effect of the nearest proxy SNP up to an r^2 of 0.5 over a 500-kb window.

After aligning all SNP effects to the BMI increasing allele in Paper I and to the T2D risk allele in Paper II, we used the random effects IVW meta-analysis described before to combine the effects of each set of concordant and discordant SNPs separately on every trait, rendering a single concordant (β_C) and discordant (β_D) effect estimates.

Traits where concordant and discordant effects differ

We then proceeded to compare the concordant and discordant estimates on each trait. We calculated the absolute difference between the two estimates δ as (121):

$$\delta = |\beta_C - \beta_D| \qquad (21)$$

The corresponding standard error of this difference:

$$SE_\delta = \sqrt{SE_{\beta_C}^2 + SE_{\beta_D}^2} \qquad (21)$$

Given the large number of statistical tests we run in the comparative analyses, we adjusted the p-values of β_C, β_D and δ using a false discovery rate correction (FDR) of 5%. Traits in which any of the combined estimates and the difference reached statistical significance after this correction were taken forward to the second analytical stage, where we applied a Random Forest algorithm designed to ensure that the difference between β_C and β_D was consistent across the SNPs from each profile.

Using Random Forest to refine trait selection

The Random Forest algorithm is a technique that operates by constructing many decision trees during training. Each tree in the forest is trained on a random subset of the training data, typically chosen with replacement (bootstrapping). Moreover, at each split in the decision tree, a random subset of features is considered for splitting. These random selection processes introduce diversity among the trees in the forest and improve prediction accuracy (122). A key aspect of Random Forest is the variable importance measure, which is computed as the average drop in accuracy of decision trees when the variable is absent. This measure is an important tool for variable selection.

We apply the Random Forest algorithm to identify the most important features that separate concordant and discordant SNPs. To do this, we converted the effect estimates for each SNP and the selected traits during the first stage to z-scores. We then placed them in a SNP–trait matrix, with SNPs coded as ‘0’ if concordant and ‘1’ if discordant. Several Random Forest classifiers (1,000 iterations) were trained with this matrix, all attempting to classify SNPs in their correct category. To ascertain which traits were relevant to distinguish discordant from concordant SNPs, we used Boruta, an algorithm that creates randomly shuffled copies of all traits in the SNP–trait matrix, and then evaluates for each trait if its contribution to the accuracy of decision trees in the Random Forest is higher than its corresponding random set (123). This ensured that in the traits selected there was minimal heterogeneity of within-profile SNP effects while having maximal between-profile difference.

Polygenic risk scores (PRS)

In Papers I and II we combined concordant and discordant SNPs into two corresponding PRSs, which were calculated using individual genotype data as follows:

$$\mathrm{PRS}_{\mathrm{Pi}} = \sum_{j \in P}^{M_P} G_{ij} \tag{23}$$

Here, M_P is the set of SNPs belonging to the P profile (either concordant or discordant) and G_{ij} is the genotype for SNP j in individual i, coded as 0, 1 and 2 based on the number of risk alleles. This is a numerical representation of an individual's concordant or discordant genetic predisposition. We used these PRSs to evaluate the association of concordant and discordant profiles to multiple outcomes by adding them as covariates in regression models, which were adjusted for age, sex and the first 10 genetic principal components, to account for population stratification (124). These results extended our phenome-wide comparison of concordant and discordant profiles previously described.

Similarly, in paper III, we calculated PRSs for BMI and each of the biomarkers in which we found phenotypic discordance extracted from the PGS Catalog. The SNPs included were weighted by the corresponding GWAS coefficient, representing the strength of its association with the trait. Likewise, in paper IV, we calculated a PRS of BMI that was used in causal inference analyses as described in the next section. The SNPs were extracted from a GWAS meta-analysis from the GIANT consortium. A set of highly significant, independent SNP were identified through clumping using a p-value threshold of $5 \cdot 10^{-8}$ and a LD r^2 of 0.2 over a 250kb window. Each of these SNPs were also weighted in the PRS calculation by the corresponding BMI coefficient. These analyses were also adjusted for age, sex and the first 10 genetic principal components.

Causal inference analyses

To provide causal estimates we used the Mendelian randomisation (MR), an analytical framework that draws from meta-analysis as well as instrumental analysis from econometrics (125). MR leverages genetic variants as instruments of an exposure and assess the causal effect of the exposure on an outcome. Because genetic variation is randomly allocated at birth and remain invariant throughout life, allocation of the genetically determined exposure is akin to a randomized controlled trial. This random allocation of the exposure enables better estimations of causal effects.

The MR methodology relies on 3 strong assumptions: first, the genetic instruments selected must be reliably associated with the exposure. Second, the instruments should not be associated with confounding factors between the exposure and the outcome. Lastly, these variants should influence the outcome solely through their impact on the exposure (Figure 3.1).

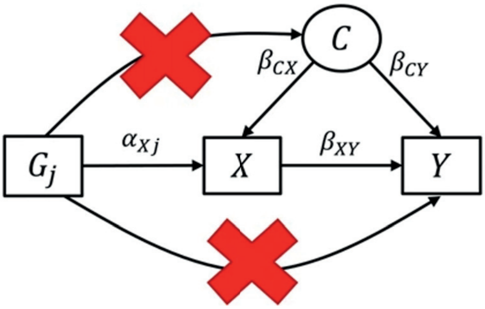


Figure 3.1. Causal diagram representing the core underlying assumptions in MR.
Here the genotype Gj is affecting the exposure X which in turn affects the outcome Y, an association confounded by C. Variables shown as rectangles or ovals, with ovals denoting potentially unobserved variables. Causal effects are indicated using one-sided arrows in the direction of the causal effect, with an accompanying effect size. Adapted from (126).

Potential causal factors of cardiometabolic discordance

We applied this framework to assess the potential causal impact of the traits that emerged from the phenome-wide comparative analyses on offsetting the diabetogenic effect of obesity in Paper I and the cardiovascular risk of T2D in Paper II. In Paper I, we selected genetic instruments for each of the traits identified in the phenome-wide comparison using SNPs that were also robustly associated with BMI (i.e., p-value for both BMI and the trait of interest $< 5 \times 10^{-8}$). In Paper II we paired each trait with T2D and run C-GWAS to identify SNPs associated with both traits. These SNP selection procedures were intended to identify instruments that reflect a dual exposure: trait of interest + BMI in Paper I, and trait of interest + T2D in Paper II.

Next, we then decomposed these instruments into two groups based on their direction of effect on the trait of interest, after alignment to the BMI-increasing allele in Paper I and T2D risk allele in Paper II, which reflected two distinct exposure groups. We then calculated the combined the effect of each of these two groups of SNPs on the outcome of interest (T2D risk on Paper I, CAD on Paper II) in a separate sample using the random effects meta-analytical approach previously described. We focused on the traits in which any of the two exposure groups conveyed a protective effect on these outcomes.

Genes and proteins with potential discordant effects

We exploited this framework also in Paper I and II to identify genes and proteins with potential discordant effects in cardiometabolic diseases. For this we used the Summary-based MR (SMR) method of colocalization, which consists of identifying for a protein or gene the strongest association signal, which is used as a genetic instrument to test for its pleiotropic (or colocalised) effect on an outcome (127).

After selection of genes and proteins whose expression is affected by discordant SNPs, we identified for each of these molecular exposures a SNP with the strongest association statistics to these genes and proteins, suitable to be used as a genetic instrument. This SNP was required to be located in the cis region (± 500 MB from the transcription start site), enhancing the likelihood that it is acting directly on the gene or protein of interest. We then computed for the SNP identified:

$$\beta_{xy} = \frac{\beta_{zy}}{\beta_{zx}} \tag{24}$$

Where β_{xy} represents extent that the genetic effect on the outcome of interest β_{zy} coincide with the genetic effect on expression β_{zx}. The statistical significance of this expression can be evaluated with the respective test statistics:

$$T_{SMR} = \frac{T^2{}_{zy}T^2{}_{zx}}{T^2{}_{zy}+T^2{}_{zx}} \sim \chi_1^2 \tag{25}$$

Here we also focused on genetic instruments that reflected a dual exposure to the molecular trait and the primary exposure in each Paper (BMI in Paper I, T2D in Paper II), and had a protective effect against the outcomes (T2D risk on Paper I, CAD on Paper II). Given the large number of genes and proteins we tested in our analysis, we applied 5% FDR correction. This is an advantage over other methods of colocalization (e.g. COLOC), where multiple test correction cannot be applied, making SMR better suited for genome-wide scans.

To ensure that these findings indicated true pleiotropy rather than mere linkage due to LD, SMR is combined with the HEterogeneity in Dependent Instruments (HEIDI) method (127). The main assumption in HEIDI is that in true pleiotropy the estimates β_{xy} calculated at SNPs in LD with the lead SNP used in SMR are homogeneous, varying only as a function of LD. Thus, using Equation 22, we can compute the difference between the estimates of each SNP i in LD against the lead SNP, which is expected to be normally distributed:

$$d_i = \beta_{xy,i} - \beta_{xy,lead} \sim \mathcal{N}(\mathbf{d}, \mathbf{V}) \qquad (26)$$

Where $\mathbf{V}$ is the covariance matrix of $\mathbf{d}$ and each element being:

$$cov(d_i, d_j) = cov(\beta_{xy,i}, \beta_{xy,j}) - cov(\beta_{xy,i}, \beta_{xy,lead}) - cov(\beta_{xy,j}, \beta_{xy,lead}) + var(\beta_{xy,lead}) \; (27)$$

And the covariance of two estimates depends on their LD correlation r:

$$cov(\beta_{xy,i}, \beta_{xy,j}) = r\frac{\sqrt{var(\beta_{zy,i})var(\beta_{zy,j})}}{\beta_{zx,i}\beta_{zx,j}} + \beta_{xy,i}\beta_{xy,j}\left(\frac{r}{z_{zx,i}z_{zx,j}} - \frac{1}{z_{zx,i}^2 z_{zx,j}^2}\right) \qquad (27)$$

To test whether the estimates of SNPs in LD are homogeneous, i.e., $\mathbf{d} = 0$, a vector of test statistics $\mathbf{z_d}$ can be derived from $\mathbf{d}$, expected to also be normally distributed around 0:

$$z_{d,i} = \frac{d_i}{\sqrt{var(d_i)}} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}) \qquad (28)$$

Where $\mathbf{R}$ is the correlation matrix of $\mathbf{z_d}$ and each element is:

$$r(z_{d,i}, z_{d,j}) = \frac{cov(d_i, d_j)}{\sqrt{var(d_i)var(d_i)}} \qquad (29)$$

The overall test statistic T_{HEIDI} is the sum of the squared vector $\mathbf{z_d}$ containing all M SNPs included in the calculation:

$$T_{\text{HEIDI}} = \sum_i^M z_{d,i}^2 \qquad (30)$$

The p-value derived from this test statistic p_{HEIDI} is approximated using the Saddlepoint method, a technique to approximate probability distributions. A higher p_{HEIDI} value means heterogeneity is less likely, which supports true pleiotropy across the gene/protein and outcome signal, while a lower p_{HEIDI} value means there is heterogeneity in the estimates, and the SMR signal is probably due to linkage. We consider an association to be potential true pleiotropy if $p_{\text{HEIDI}} > 0.01$.

We applied this technique on significant SMR signals that had at least 3 SNPs at sufficient LD with the lead SNP ($0.05 < r^2 < 0.9$) and with a sufficient strength of association to the gene or protein ($p < 1.53 \cdot 10^{-3}$, equivalent to a $\chi_1^2 > 10$) to be informative for HEIDI. We included in the calculations the top 20 SNPs ranked by their p-values of association to the gene or protein, as this is where the power of the

HEIDI test and its computational efficiency is maximal according to simulations (128).

Causal effects of BMI and its shape on cardiometabolic conditions

In Paper IV we used a BMI PRS as the instrumental variable and used the two-stage least squares (TSLS) method to assess the effect of genetically determined BMI on multiple outcomes. The first stage of TSLS involves regressing the exposure on the instrumental variable while adjusting for relevant covariates. Then, fitted values of the exposure are generated for use in the second stage model. Here, the outcome is regressed on these fitted values (used as the exposure) while adjusting for the same covariates as in the first stage. The regression coefficients of the fitted values represent estimates of the causal effect of the exposure on the outcome.

Additionally, in Paper IV we also applied an extension of MR that allows the exploration of non-linear relationships between the exposure and the outcome, known as the residual method (129). This method involves calculating local average causal effects (LACE) as ratios of coefficients within quantiles of the exposure. However, given that conditioning on a certain level of the exposure to find a causal estimate could lead to collider bias, the quantiles generated are instead based on the instrumental variable-free distribution of the exposure. This distribution is calculated as the residuals of a model in which the exposure is regressed on the IV. From these LACE values, the relationship between exposure and outcome can be fitted using non-linear regression, which is achieved by including fractional polynomial terms. Tests of nonlinearity are then applied to test the null hypothesis that the resultant non-linear model is no different from a linear model.

This method, however, relies on the strong assumption of a constant effect of the genetic instrument on the exposure across all quantiles, which can be violated when, for example, the effect of the instrument on the exposure is itself nonlinear (130). To overcome this issue a novel method to calculate the quantiles have been designed, called the doubly ranked method, where individuals are ranked first according to the level of the genetic instrument, and then according to their level of the exposure (131). This ranking procedure maximises the similarity in the distribution of the genetic instrument identical within each quantile, which controls the effect of heterogeneous instrument-exposure associations, while still obtaining strata where the average level of the exposure is increasing, as necessary to estimate the shape of the causal relationship. We used the doubly rank method to rerun the nonlinear analyses in Paper IV.

Phenotypic discordance with respect to the BMI

In Paper IV we estimated the degree of discordance between the observed value of 10 common clinical biomarkers against the expected for the BMI. We selected these biomarkers because of their clinical use in the assessment tools of different biological systems that are affected in obesity:

- Glycaemia: Fasting glucose.
- Lipid metabolism: HDL, LDL and TG.
- Systolic and diastolic blood pressure (SBP and DBP, respectively).
- Renal function: Creatinine (SCr).
- Liver function: Alanine transaminase (ALT).
- Fat distribution: Waist-to-hip ratio (WHR).
- Inflammation: C-reactive protein (CRP).

We estimated the age- and current smoking-adjusted change in all biomarkers for every unit increase in BMI using linear models. Then we calculated the difference between expected and observed values, which were centred and scaled to have zero mean and unit standard deviation. We then measure to what extent individuals deviated from the expected multivariate normal distribution of standardised residuals $\mathcal{N}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the observed covariance matrix of residuals across biomarkers. To this end we used the squared Mahalanobis distance, which is defined as (132):

$$\mathrm{D}_{M,i}^2 = (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \qquad (31)$$

Here, $\boldsymbol{X}$ is the vector containing biomarker discordance for individual i and $\boldsymbol{\mu}$ is the mean vector of the distribution, equal to a vector of zeros in this case. Since D_M^2 follows a χ^2 distribution with degrees of freedom equal to the number of biomarkers, we used this property to calculate for each individual their probability to belong to the expected distribution of residuals and quantified the proportion of individuals who are above the critical threshold of 0.05 (expected proportion 5%). We compared the observed proportion to the expected using a binomial test.

Probabilistic clustering of phenotypic discordance

To identify subgroups of individuals with similar patterns of biomarker deviations from the expected for their BMI, we applied Uniform Manifold Approximation and Projection (UMAP) to the residual data. UMAP is a dimension reduction technique that uses a network-based approach to represent the data, connecting individuals that are similar to each other, and then embedding this network in a lower dimensional space while preserving both the local and global structure of the network (133). We embedded the residual data in a two-dimensional space to visualise the distribution of BMI-discordance. Thereafter, we applied an ensemble of clustering algorithms to the underlying network to identify distinct discordant profiles, while recognising that individuals may display features of more than one profile, thus avoiding forcing individuals to have a single profile.

The ensemble algorithm consisted of several steps. We first applied two network-based algorithms to partition UMAP's network: first, we used the leading eigen vector algorithm to decompose the proximity matrix that represents the network (134). This provided stable initial seeds to subsequently run the Leiden algorithm, which partition the data iteratively in densely connected 'communities' of individuals with similar discordant profiles, while maximising the modularity score, a measure of the density of within-community relative to between-community connections, until no further improvements can be made (135). We iterated the Leiden algorithm over 500 times, resulting in hard partitions, where individuals are assigned to a single cluster.

To transform this hard partition into a probabilistic partition, we calculated for every individual the normalised eigen centrality scores for their respective clusters, which measures its importance within the cluster. These scores were used as weights to calculate the mean vector $\boldsymbol{\mu_k}$ and covariance matrix $\boldsymbol{\Sigma_k}$ of each cluster, thereby converting each hard cluster into a sub distribution, or profile, of discordance. We used these profiles to fit a Gaussian mixture model, where the overall probability distribution of discordance is modelled as a weighted sum of each profile (136):

$$\mathrm{P}(\boldsymbol{X}) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \qquad (32)$$

Where π_k is the weight of profile k, which reflects how frequent the k profile is in the population. To further improve the separation of highly discordant profiles deviating strongly from the expected normal distribution of residuals, we included in the Gaussian mixture a highly 'concordant' profile, where residuals follow a normal distribution around 0 with identity covariance matrix ($\mathcal{N}(0, \mathbf{I})$), reflecting no correlation of residuals beyond the explained by BMI. In the Gaussian mixture calculation, we kept $\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma_k}$ fixed and estimated π_k. The resulting partition included concordant and discordant profiles and every individual i was given a probability score $\pi_{k,i}$ for each profile, satisfying that for each individual:

$$\sum_{k=1}^{K} \pi_{k,i} = 1 \qquad (33)$$

We ran identical analyses in the discovery cohort (UK Biobank) and in the cohorts in SOPHIA and assessed whether individuals allocated to a profile in the discovery model with high certainty (i.e., a probability higher than 80%) also had a high median probability of being allocated to a profile found in any of the other three 'validation' cohorts (again higher than 80%). We considered a profile as being replicated if this condition was met in all three validation cohorts, which ensured that only robust clusters were included in the final model. We then readjusted the weights for each profile and focused all downstream analyses on these latter replicated clusters.

To assess cluster separation quality, we calculated the relative entropy of the final partition. This measure takes values from 0 to 1, indicating either identical probability distributions of all profiles (i.e., equal probabilities to all profiles for every subject) or perfect cluster separation (no overlap between clusters) (137). We also used the relative entropy to compare our final partition to partitions obtained from other clustering algorithms applied to the biomarker discordance data, all being based on different approaches for data partition: centroid-based (Gaussian mixture model directly applied to the data), boundary-based (archetypal model, where archetypes are located at the extremes of the data distribution) (138) and density-based (Hierarchical Density-Based Spatial Clustering of Applications with Noise, HDBSCAN, which is able to detect clusters of irregular shapes) (139). More details of these algorithms are found in Paper III.

Understanding probabilistic clustering as a composition

In Paper III, once we identified the concordant and discordant profiles and estimated the probability of profile allocation $\pi_{k,i}$ for every individual i to every profile k, we characterised each profile by calculating weighted averages and proportions using the corresponding $\pi_{k,i}$, thereby using data from all individuals in these calculations and improving statistical power (140).

Additionally, we assessed the risk to events of interest (MACE and diabetes) conveyed by profile allocations by fitting Cox proportional hazard models with π_k as predictors. Due to the sum-to-1 constrain in Equation 33, using π_k directly as predictors makes the model unidentifiable, i.e., it has more than one solution. We therefore turned to the log-contrast framework, commonly applied in the analysis of compositional data (141). In this framework, compositional predictors, which have the sum-to-1 constrain, such as π_k, can be accommodated in a regression model:

$$y = \sum_{k=1}^{K} \alpha_k \cdot log(\pi_k) \qquad (34)$$

With the constrain that:

$$\sum_{k=1}^{K} \alpha_k = 0 \tag{35}$$

To implement this constrain, the model can be reformulated to:

$$\mathrm{y} = \sum_{k=1}^{K \neq r} \beta_k \cdot log\left(\frac{\pi_k}{\pi_r}\right) \tag{36}$$

This transformation is known as the additive log-ratio transformation, where r is an arbitrarily selected reference component of the composition. We chose this reference component to be the concordant profile, thereby β_k being the effect of each discordant profile. This estimate represents the change in the outcome y expected from increasing the log-ratio of one discordant profile one unit while keeping the other log-ratios constant. Because of the sum-to-1 constrain, this is equal to increasing the probability of a particular discordant profile by a certain factor while decreasing all other profiles by the same factor, which would effectively keep the other log-ratios constant.

Given that we were interested in measuring the added value of BMI-discordance in prediction beyond the information provided by biomarkers and BMI alone, we fitted nested models with and without discordant log-ratios. In the models that included discordant log-ratios, a change in the probability of a certain discordant profile inevitably carries changes not only in the other profile probabilities but also in the biomarkers. In this context, discordant log-ratios essentially represent interactions terms that modify the relationship between biomarkers and risk, conditional on their pattern of discordance with BMI. This might be better illustrated in a simplified example where an outcome is modelled as a function of glycemia and BMI, as well as the discordance between observed and predicted glucose based on BMI:

$$\mathrm{y} = \beta_1 Glucose + \beta_2 BMI + \beta_3 (Glucose - E(Glucose|BMI))(37)$$

Here, β_3 would represent the effect of a certain discordant log-ratio in the high-dimensional case. Thus, the effect of increasing glucose on the outcome is modified by the degree of discordance between observed and predicted glucose. Therefore, to correctly estimate the effect of a shift in the probability distribution from the concordant to a specific discordant profile, while keeping the other discordant profiles fixed at their population value, we included all the changes – both in biomarker and discordant log-ratio terms – that would correspond to this shift.

Added value of a biomarker in risk prediction

Nested models

In Papers II and III we used Cox regression models to predict clinical outcomes and assessed whether adding information on genetic or phenotypic discordance improved the predictive ability of these models. The main clinical outcome we attempted to predict in these two Papers was MACE. The baseline model contained the predictors of SCORE2, the current risk stratification tool for primary care prevention of CVD recommended by the European Society of Cardiology (142). The competing models contained in addition information on T2D-CAD genetic discordance (Paper II) and biomarker-BMI discordance (Paper III). Relevant comorbidities, as well as antidiabetic, antihypertensive and lipid lowering medication were included as covariates in both models. In Paper III we run a similar analysis of competing models to predict diabetes diagnosis, where the baseline model included the biomarkers while the competing model had in addition information on biomarker-BMI discordance. These models were fitted separately for men and women. We ensured models were properly calibrated with calibration plots comparing the predicted versus the observed event rate, and then carried out comparison across the nested models using likelihood-based and ranked-based methods as explained below.

Likelihood-based measures

From Equation 9 it is shown that the partial log-likelihood in Cox regression represents how well the model separates individuals who experience the event of interest from those who are still at risk. Consequently, a powerful way to compare two models that are nested, one of which includes additional covariates, is by comparing their log-likelihoods. The ratio of the log-likelihoods of two nested models follows a χ^2 distribution with degrees of freedom equal to the number of additional covariates in the complete model. Thus, this test statistic, known as the likelihood ratio test (LRT) statistic, can yield a probability that the model comparison deviates from the null hypothesis of no improvement in prediction, while penalising for the added complexity of the model (140).

Moreover, once the LRT shows statistical significance, a quantity of added information can be derived from the LRT statistics from comparing each model to a null model with no coefficients (where essentially the likelihood is reduced to the cumulative risk at every event time). The ratio of the baseline to the complete model statistic (LRT_B/LRT_C), called the adequacy index, represents the proportion of explained variation in the complete model that is contained in the baseline model.

Hence, one minus this quantity is the fraction of additional variance F_{Add} explained by the additional covariates in the complete model (140):

$$F_{Add} = 1 - Adequacy\ Index = 1 - \frac{LR_B}{LR_C} \qquad (38)$$

Rank-based measures

To perform optimally, a Cox model should predict longer survival times for individuals who remain event-free for longer periods and shorter survival times for those who experience the event sooner. This means that if we rank individuals based on their predicted and actual survival times, these two ranks must be as highly correlated as possible. Two rank-based measures were used in our analyses: the c-statistic and the net reclassification improvement.

The c-statistic

The most common way to quantify the rank correlation is using the concordance statistic, or c-statistic. This is computed as a scoring system at each event time by pairing individuals who have experienced the event with those who have not yet done so. If the predicted probability for the individual who had the event is higher than the individual without the event, the pair is said to be concordant, adding 1 point to the score. Pairs with tied predicted probabilities add 0.5 to the score. The final statistic is the ratio of the score over the total number of pairs (143):

$$C = \frac{N_{Concordant} + 0.5 \cdot N_{Tied}}{N_{Pairs}} \qquad (39)$$

This statistic reflects the probability that if we take two individuals i and j at random, one with the event and one without, the individual with the event will have a higher predicted probability, i.e., will be ranked higher:

$$C = P(R_i > R_j | Y_i = 1, Y_j = 0) \qquad (40)$$

To calculate the variance of this statistic, an infinitesimal jack-knife approach is generally used, which consists of computing how much the statistic varies by leaving each observation out.

Competing models in the medical literature are often compared using their difference in C-statistic. The variance of this difference and its corresponding confidence interval can also be computed using the infinitesimal jack-knife method. However, the C-statistic difference is insensitive to clinically meaningful changes in predicted probabilities that do not change the ranks (115,116).

The net reclassification improvement (NRI)

Another approach commonly used to measure the added benefit of a biomarker is to evaluate if its inclusion in the prediction model help reclassify better cases and non-cases into higher and lower risk categories, respectively. This can be inspected in reclassification tables and plots, stratifying individuals based on their predicted probabilities by the baseline and the complete model and evaluating whether the predicted event rates from the complete model match better the observed event rates than the baseline model (144).

To quantify the overall proportions of individuals with and without the event who were correctly reassigned higher or lower risk, respectively, we used the Net Reclassification Improvement (NRI), adapted to the survival setting, and calculated separately for events (NRI_E) and non-events (NRI_N) (145):

$$NRI_E = \frac{P(T \leq t | R_C > R_B) \cdot P(R_C > R_B) - P(T \leq t | R_C < R_B) \cdot P(R_C < R_B)}{P(T \leq t)} \quad (41)$$

$$NRI_N = \frac{P(T > t | R_C < R_B) \cdot P(R_C < R_B) - P(T > t | R_C > R_B) \cdot P(R_C > R_B)}{P(T > t)} \quad (42)$$

In these equations, '$T \leq t$' and '$T > t$' denote the occurrence or absence of the event at a given time point t, and R_B and R_C are the predicted risks derived from the baseline and complete models, respectively. These measures have the advantage of being readily interpretable as proportions correctly reclassified within each group, while not being affected by threshold selections as the reclassification tables. They can, however, reflect changes in the predicted probabilities that might be too small to be clinically meaningful.

Decision curve analysis

While the previous measures are useful to define how well predictive models discriminate cases from non-cases, they do not provide direct guidance on how these models should be translated into clinical practice, where thresholds on predicted probabilities are used to make decisions of whether or not to intervene, and what intervention modalities to use.

Thresholds that guide interventions implicitly reflect how health practitioners and patients weigh true positives over false positives, as well as true negatives over false negatives (146). Lower thresholds give higher weight to detecting cases, while higher thresholds give higher weight to avoiding unnecessary interventions. For example, if a threshold to intervene is set to 10%, this means that a true positive is weighted 9 times higher than a false positive, or that it is acceptable to intervene 9 individuals who might not benefit from the intervention to treat 1 individual who will.

As a consequence of this, a measure of the net benefit (NB) of using a predictive model to guide an intervention can be derived by comparing the proportion of true positives TP against the appropriately weighted proportion of false positives FP, where the weight w is dictated by the threshold Pt that defines whether or not to intervene, and how TP are valued in terms of FP (146):

$$NB = TP - wFP \tag{43}$$

$$w = \frac{Pt}{(1 - Pt)} \tag{44}$$

In the context of survival data:

$$TP = P(T \leq t | R > Pt) \cdot P(R > Pt) \tag{45}$$

$$FP = (1 - P(T \leq t | R > Pt)) \cdot P(R > Pt) \tag{46}$$

Where $T \leq t$ denotes occurrence of an event at time t and R is the predicted risk. The NB indicates the number of true positives gained by using the model without increasing the number of unnecessary interventions. The NB can be calculated and plotted over a range of a range of plausible Pt, which enables the examination of the clinical utility of one or multiple models across different scenarios, with the model with the highest NB providing more clinical utility. Additionally, any predictive model has to be superior to the NB of default strategies of universal intervention (which has the greatest utility at lower Pt but drops below 0 when Pt matches the disease rate) or no intervention (which has always a $NB = 0$).

Likewise, the clinical utility can also be expressed in terms of net interventions avoided (NIA), which is calculated as a weighted difference in the NB of the model against that of the universal intervention strategy NB_{UI}, weighted by the inverse of w:

$$NIA = (NB - NB_{UI}) \cdot w^{-1} \tag{47}$$

Chapter IV – Results

Paper I

In this Paper we found 67 SNPs strongly associated with BMI and T2D, with 48 concordant and 19 discordant, which we used to construct the respective obesity profiles. In our phenome-wide exploration, we found that the concordant and discordant profiles differed strongly in three features: HDL, WHR and SBP (Figure 4.1). We also found differences in risk of CAD and stroke, which were lower in the discordant compared to the concordant profile. The levels of liver biomarkers such as ALT were lower in the discordant relative to the concordant profile. The sex-hormone binding globulin (SHBG), a protein also produced in the liver and associated with better metabolic function, was higher in the discordant as opposed to the concordant profile. There were also differences in red blood cell morphology, with discordant profile was associated with higher mean corpuscular volume. The concordant profile had higher levels of urate compared to the concordant profile. Interestingly, the odds of receiving treatment with alendronate was higher in the discordant than in the concordant profile, a drug indicated for osteoporosis.

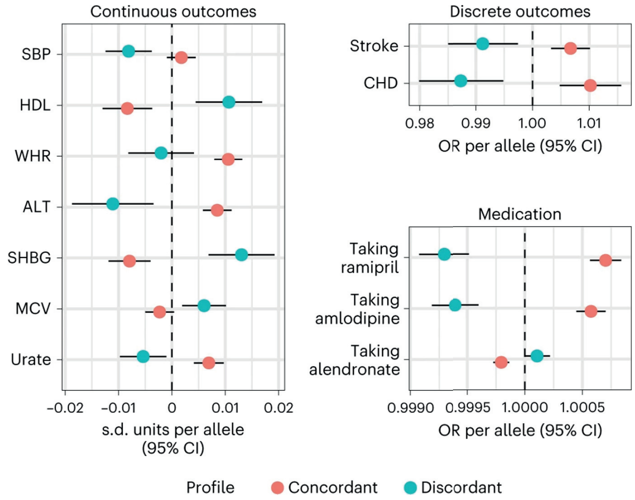


Figure 4.1. Phenome-wide comparison of concordant and discordant obesity profiles. Concordant and discordant effects on traits where we found significant differences between profiles using GWAS summary data.

In the PRS analyses in BioVU (Figure 4.2), we found that both profiles were associated with higher obesity risk, as well as the odds of receiving surgical interventions for obesity. In line with our expectations, the concordant profile was associated with higher risk of diabetes and higher levels of related biomarkers such as HbA1c, while the discordant profile associations tended to the opposite direction. This divergent association pattern was also reproduced in African American individuals. We replicated in BioVU our previous results of differences in lipids, blood pressure and red blood cell morphology, while finding additional differences, such as chronic kidney disease (CKD), which was higher in the concordant compared to the discordant profile, and osteoarthrosis, which was higher in the discordant compared to the concordant profile. We also found that leucocyte count, urea, creatinine, phosphate, C-reactive protein and PTH were all at higher in the concordant compared to the discordant profiles.

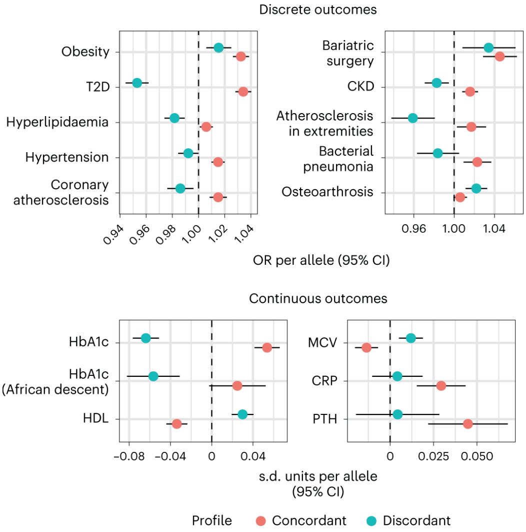


Figure 4.2. Comparison of concordant and discordant in BioVU.
The figure shows traits where we found significant differences after a 5% FDR correction.

The PRS analysis in the UK Biobank showed that the relationship between the two obesity profiles and early mortality was also divergent: the concordant profile was associated with higher mortality (HR per allele: 1.01, 95% CI: 1.01, 1.02), while the discordant profile was not (HR per allele: 0.99, 95% CI: 0.98, 1.01, pδ = 0.02).

In the analysis of the molecular differences between the two profiles, we found that the discordant profile was associated with higher cholesterol in lipoprotein particles of all densities, particularly HDL, while lower triglyceride content in lipoprotein particles of low densities, as opposed to concordant diabesity. The discordant profile also correlated with lower levels of branched chain (BCAA) and aromatic amino acids (Figure 4.3, left panel).

Although there were no differences between pooled concordant and discordant estimates for bacterial abundance in the gut that survived the FDR correction, we found nominal associations (p < 0.05) in taxa belonging the phyla Bacteroidetes – more abundant in the concordant profile – and Firmicutes – more abundant in the discordant profile (Figure 4.3, bottom right panel).

We found two proteins strongly influenced by discordant variants: heparan sulfate 6-O-sulfotransferase 2 (HS6ST2), which was higher in the discordant relative to the concordant profile, and metalloproteinase inhibitor 4 (TIMP4), which was under the influence of a discordant SNP near PPARG. In the expression data we found around 800 genes whose expression/splicing in a variety of tissues was significantly influenced by concordant and discordant SNPs. Enrichment analysis indicated that the discordant, but not the concordant profile, was functionally enriched in adipose tissue.

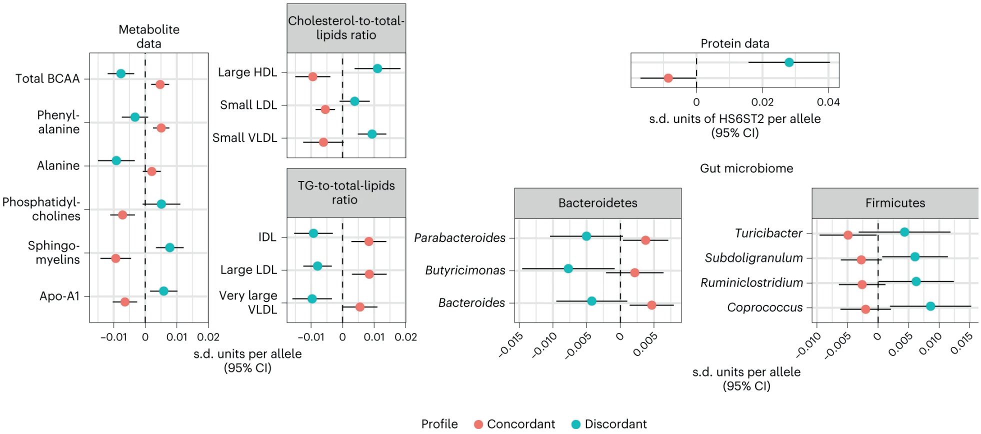


Figure 4.3. Molecular differences between concordant and discordant obesity profiles
Significant differences between estimates from each profile found in metabolite, protein and gut microbiome data. Metabolite and microbiome traits shown were significantly different between the two profiles. Gut microbiome data show bacterial taxa where we found nominal differences (p < 0.05).

In our causal inference analyses, we found that genetic profiles of higher BMI but WHR, lower SBP or higher free cholesterol content in HDL particles were associated with lower T2D risk. Additionally, we identified significant discordant effect of the TIMP4 protein levels and the expression of 17 genes in multiple tissues (Figures 4.4 and 4.5).

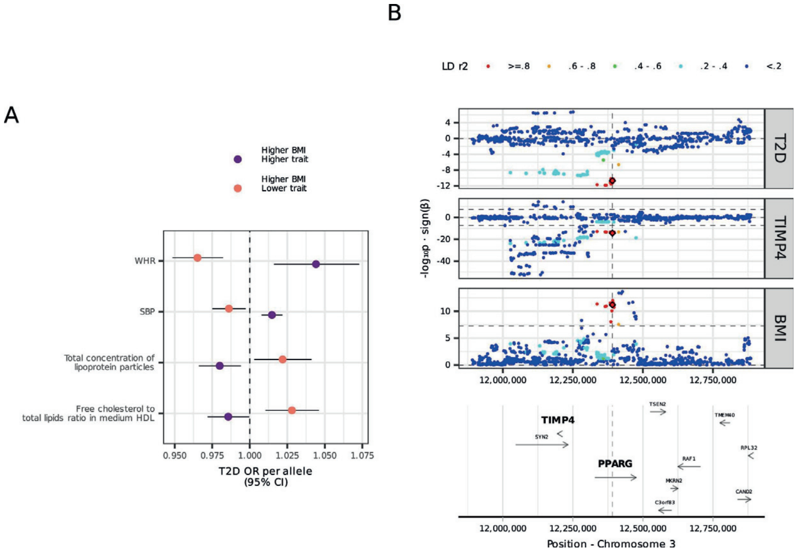


Figure 4.4. Estimates of traits with potential discordant protective effects.
Left, clinical and molecular traits where we found that a genetic profile associated with obesity and either higher or lower level of the trait was associated with T2D risk. Right, locus zoom plot showing regional association to BMI, TIMP4 levels and T2D risk.

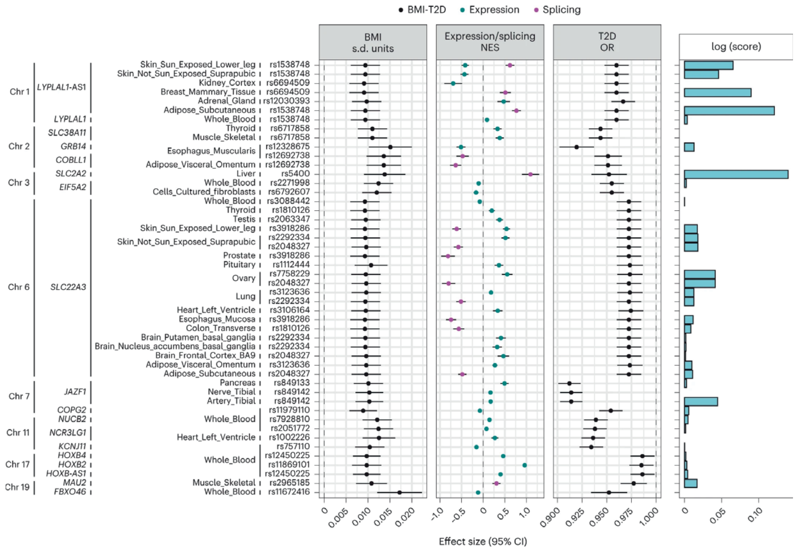


Figure 4.5. Genes with potential discordant protective effects in obesity.
Genes whose expression are pleiotropically associated with higher BMI but lower T2D risk, as evidenced by SMR and HEIDI methods. The three central panels show the association of the lead genetic instrument of each gene on BMI, expression and T2D risk. On the right, the logarithm of the score we used to identify the potential tissue of action of the instrument used to identify the gene.

Paper II

In this Paper we run cross-trait GWAS analyses between T2D and major cardiovascular complications of diabetes: CAD, acute ischaemic stroke (AIS) and CKD. We only found enough signals to run our comparative analysis in the T2D-CAD dyad: 83 SNPs, of which 70 were concordant and 13 discordant. We focused all downstream analyses to this dyad.

In the PRS analyses of concordant and discordant T2D-CAD profiles, we found that despite both PRSs showing associations with higher T2D, they exhibited opposing effects on primary incidence of MACE. When the two PRSs were added to SCORE2, we observed a statistically significant increase in the LRT in men, with a F_{Add} of around 1.5%. In this sex group, the difference in C-statistic between the two models was small but statistically significant. In the reclassification tables (Figure 4.6) we found that approximately 8.6% of the population were reclassified into different risk categories with considerably better calibrated predictions. The NRI calculations showed that 5.4% of male incident MACE cases were correctly reassigned as having higher predicted risk, while 2.93% of men who were subsequently event-free were reassigned as having lower predicted risk.

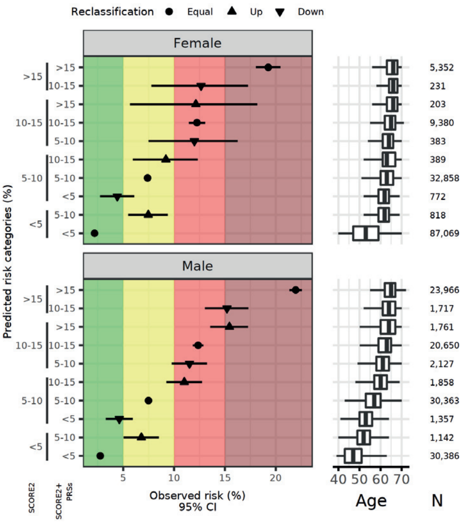


Figure 4.6. CVD risk reclassification using concordant and discordant profiles.
The y axis classifies individuals based on their predicted risk by both models. The x axis is the observed risk, showing Kaplan-Meier estimates within each category. To the right, distribution of age and size of each risk category.

We introduced interaction terms in regression models to assess the effect of the PRSs in high-risk groups. We found a statistically significant interaction between T2D and the discordant PRS in males, which translated into a lower MACE risk in individuals with T2D but a predominantly discordant genetic profile (Figure 4.7). We also rerun our analysis including individuals with and without MACE history and added interaction terms between each PRS and prior MACE. We found a significant interaction between the concordant PRS and prior MACE, such that the effect of the concordant PRS was significantly attenuated in individuals with prior MACE. We further explore this in the ORIGIN trial, where we found that the effects of the concordant and discordant PRSs on MACE prevalence were directionally consistent with our stratification in European and Latin American populations. In contrast to these cross-sectional associations, during follow-up the discordant PRS was significantly associated with higher MACE incidence, exceeding the estimated effect of the concordant PRS. In ancestry-specific analyses we found that in Native Latin population the concordant PRS was associated with lower incidence of MACE. We noted that this finding was driven by individuals with prior history of MACE. Nonetheless, the overall association of the discordant PRS with higher incidence was attenuated after adjusting for traditional risk factors (see Supplementary Note in Paper II).

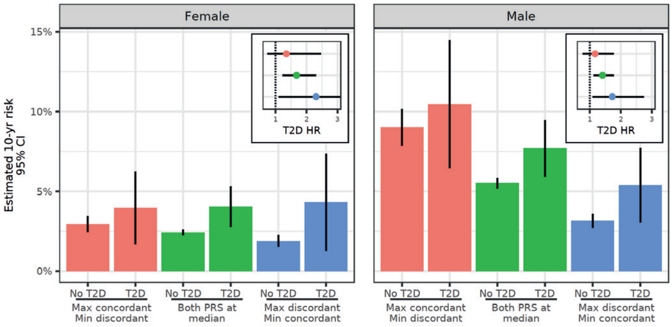


Figure 4.7. Differential effect of T2D by the level of concordant and discordant predisposition. This figure shows the predicted CVD risk in a 60-year old individual with or without T2D at varying levels of concordant and discordant PRS, obtained from a model with T2D*PRS interactions.

We found that concordant and discordant profiles differed in selected traditional CVD risk factors, especially in TG and SBP, both of which were lower in the discordant compared to the concordant profile (Figure 4.8, top panel). In the phenome-wide comparison, we found additional differences in other manifestations of atherosclerotic diseases (Figure 4.8, bottom panel). For example, the discordant profile was associated with lower odds of peripheral artery disease (PAD) and stroke relative to the concordant profile. The discordant profile also conveyed lower risks of diabetic hypoglycaemia compared to the concordant profile. Interestingly, we found that blood monocytes, the precursors of macrophages, were higher in the discordant compared to the concordant profile. Paternal lifespan was longer in the discordant compared to the concordant profile. Although there was a similar pattern of divergence in maternal lifespan, the difference between profiles was smaller and did not pass multiple test correction. The concordant profile had higher odds of dorsopathies, higher self-reported health satisfaction and lower odds of having been breastfed as a baby compared to the discordant profile.

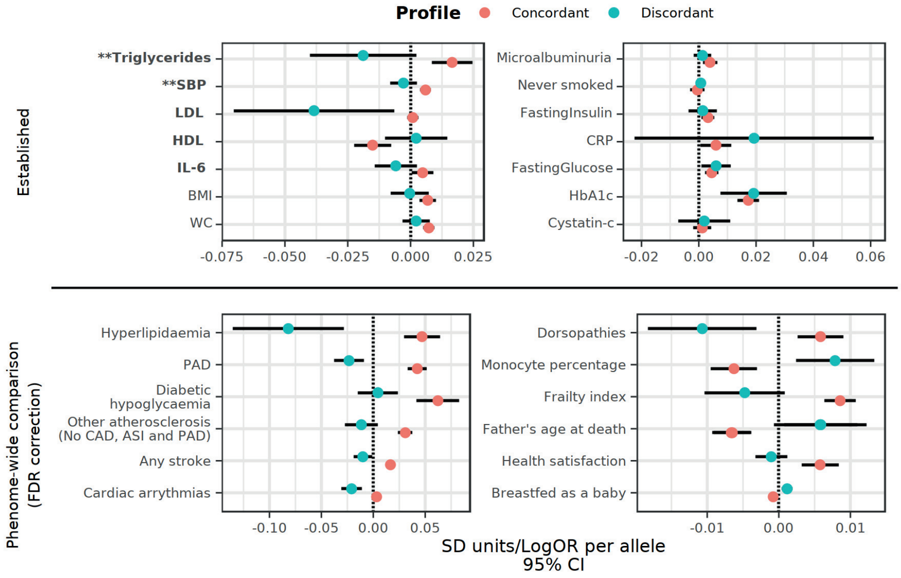


Figure 4.8. Phenome-wide comparison of T2D-CAD concordant and discordant profiles. The top panel shows differences within a group of selected established CVD risk factors. The bottom panel shows the differences obtained from running a phenome-wide comparison between the two profiles.

From the metabolite data we found that the concordant profile was associated with higher levels of small VLDL particles, as well as lower concentrations of large HDL particles with lower cholesterol and phospholipid content, compared to the discordant profile (Figure 4.9, top left panel). This is in line with the findings from the protein scan, where we identified 43 proteins whose levels in blood were significantly influenced by SNPs in the discordant profile, almost all driven by a missense mutation in APOE. These included lower levels of E2 and E3 isoforms of apolipoprotein E, as well as other proteins involved in lipid metabolism elsewhere in the genome, such as lower apolipoprotein B levels and lower levels of cardiotrophin-1, an inflammatory biomarker highly expressed in heart. We also found strong links between SNPs in the discordant profile and expression and splicing of over 100 genes in multiple tissues.

In the causal inference analysis, we found that the content of free cholesterol and the content of phospholipids in VLDL particles are likely within the causal pathways leading to protection against atherosclerosis in diabetes (Figure 4.9, top right panel). We prioritized 8 independent loci harbouring 33 genes whose expression in a variety of tissues was strongly associated with discordance between T2D and CAD (Figure 4.9 bottom panel). Among these genes was HMGCR, the target of statins, and KCNK5, a gene encoding the potassium channel K2P5 and in the vicinity of GLP1R, the target of glucagon-like peptide 1 receptor agonists.

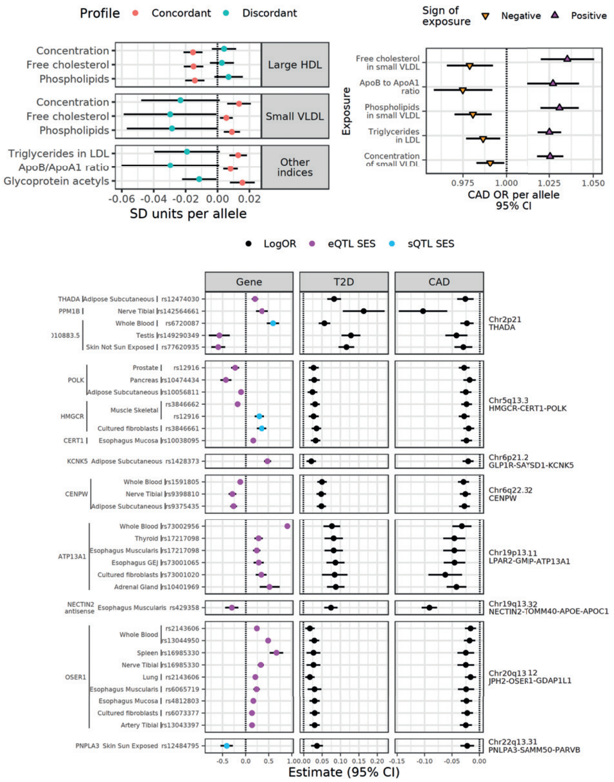


Figure 4.9. Molecular traits with potential discordant protective effects on T2D.
Top left panel: main differences between concordant and discordant T2D profiles in metabolite data. Top right panel: metabolites where we found that a genetic profile associated with T2D and either higher or lower level of the metabolite was associated with CAD risk. Bottom: genes whose expression are pleiotropically associated with higher T2D but lower CAD risk, as evidenced by SMR and HEIDI methods. The three panels show the association of the lead genetic instrument of each gene on BMI, expression and CAD risk.

Paper III

In this Paper we used UK Biobank data to quantify biomarker discordance to the expected given the BMI, finding that a significantly higher proportion of individuals displayed substantial biomarker discordance compared to the anticipated proportion under a normal distribution (expected proportion = 5%, observed proportion = 10.3%, $P_{\text{binomial}} < 0.001$). These individuals appeared clustered in subgroups in the UMAP projection (Figure 4.10), a pattern absent in projections generated under a normal distribution. Principal component projections were unable to detect these deviations. The UMAP projections in the SOPHIA cohorts had similar patterns.

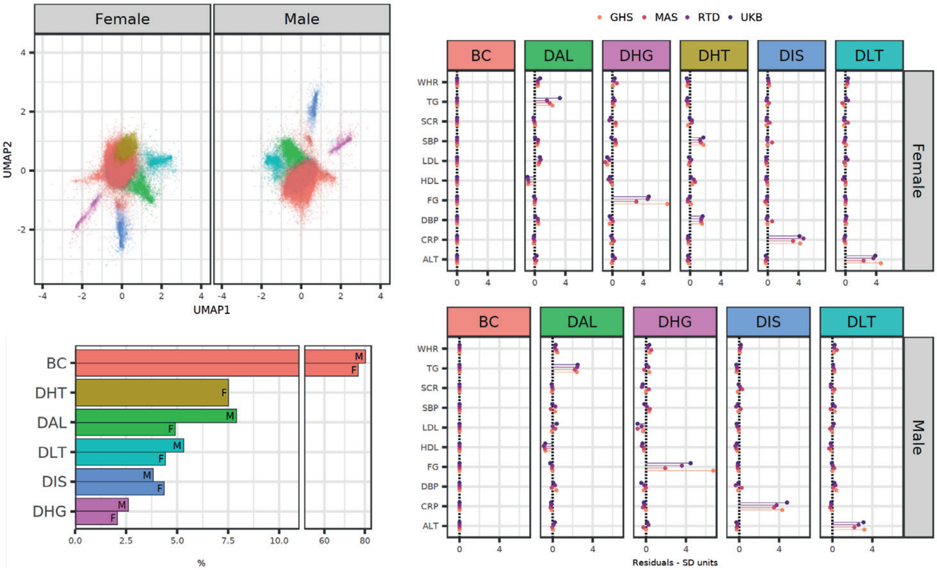


Figure 4.10. Concordant and discordant phenotypic profiles.
Profiles discovered in the UK Biobank and robustly replicated across 3 independent cohorts. Left upper panel: UMAP two-dimensional projection. Colours denote profile allocations. Left lower panel: Cluster weights. Right panels: Profile centres.

Our clustering algorithm run on each sex separately defined a concordant profile and five discordant profiles consistently replicated across all cohorts, with high relative entropy (>.8). Most individuals (~80%) had predominantly a concordant phenotypic profile (called the baseline concordant [BC] profile). Approximately 8% of females displayed a discordant hypertensive profile (DHT), with blood pressure values above the expected for their BMIs. This profile was not found in males. Around 5% of females and 7% of males showed a discordant adverse lipid profile (DAL), characterised by higher TG, lower HDL, and higher LDL than expected for their BMIs. Profiles of discordant liver transaminase (DLT) and discordant inflammatory state (DIS), characterised by higher ALT and CRP than expected for the BMI, respectively, were each observed in 4 to 5% of individuals in both sexes. About 2.5% had a discordant hyperglycaemic profile (DHG), with higher fasting glucose levels that correlated with lower LDL levels than the expected for the BMIs.

Among the differences at baseline between these profiles, the DHG profile in both sexes was associated with over 30-fold higher odds of T2D and around 3 times higher odds of CAD compared to the BC profile. In contrast, there were fewer cases of T2D and CAD in DAL compared to BC in both sexes. Individuals with this profile were also less likely to be taking lipid-lowering medication compared to BC.

Adding BMI-biomarker discordance information to fully adjusted models for MACE prediction led to a statistically significant increase in LRT and the c-statistic in men, with a value of F_{Add} that ranged between 1.4 to 5.4%. For a proper inference from these models, we derived the expected change in risk of MACE in a 60-year-old individual with a BMI of 30 kg/m2 whose probability to have a specific discordant profile increases 10%, at the expense of decreasing the probability to have a BC profile by the same amount. After multiple test correction, a 10% higher probability to have a DAL profile was associated with a higher risk of MACE compared to BC across sexes. In contrast, a 10% higher DHG profile probability was associated with lower risk of MACE compared to BC (Figure 4.11).

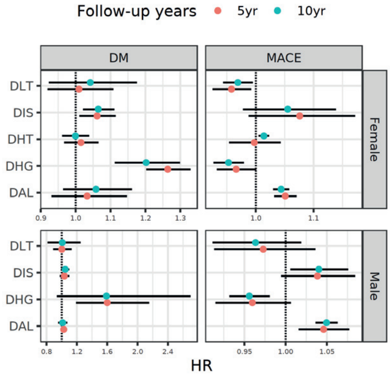


Figure 4.11. Risk of MACE and diabetes within each discordant profile relative to the concordant. Pooled estimates of the expected hazard ratio corresponding to a 10% higher probability to have a specific discordant profile while decreasing their probability of being concordant by the same amount. Estimates were derived using 5 years (which contained more individuals at baseline) and 10 years of follow-up.

In the case of predictive models for diabetes progression, we found greater gains in prediction from adding BMI-biomarker discordant information in individuals with higher median glucose values at baseline. For instance, in the Rotterdam Study, we found that Fadd was to 8 – 12%, much higher than in the UK Biobank (<1%). This was mainly driven by the DHG profile. A 10% increase in the probability of having a DHG profile and away from the BC cluster was associated with 20 – 60% increase in risk of progressing to diabetes compared to individuals in BC (see Table 1 in Paper III).

In decision curve analysis we found that BMI-biomarker discordance yielded a net benefit of 4 additional true positives and 37 additional true negatives per 10,000 men compared to the baseline model, when the threshold for intervention was set at 10% (Figure 4.12). To provide context for these values, we computed the additional net benefit of LDL, a recognized MACE risk factor, beyond models incorporating only age, which is 5 additional true positives and 42 additional true negatives per 10,000 men individuals tested. In diabetes progression models BMI-biomarker discordance provided 15 additional true positives and 135 additional true negatives per 10,000 women, and 4 additional true positives and 33 additional true negatives per 10,000 men.

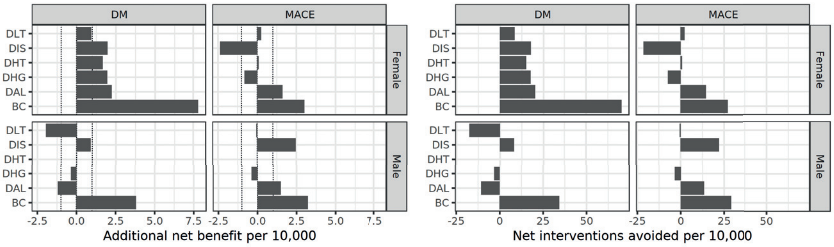


Figure 4.12. Additional net benefit and net interventions avoided by using discordant profiles. The results are discriminated by profile to determine the profiles with the highest and lowest gains in prediction. The dotted lines in the left panel correspond to the unity.

We evaluated how discordant profiles identified in European populations were distributed in African and South Asian populations in the UK Biobank. Both populations had higher odds of DHG profiles compared to the Europeans, although the risk was twice as high in South Asians, who were also more likely to have a DAL profile than Europeans. Improvement of models predicting MACE by incorporating discordant profile information was highest in South Asian men with a DAL probability, while in diabetes progression the highest gains were in African men with a DHG profile (Figure 4.13).

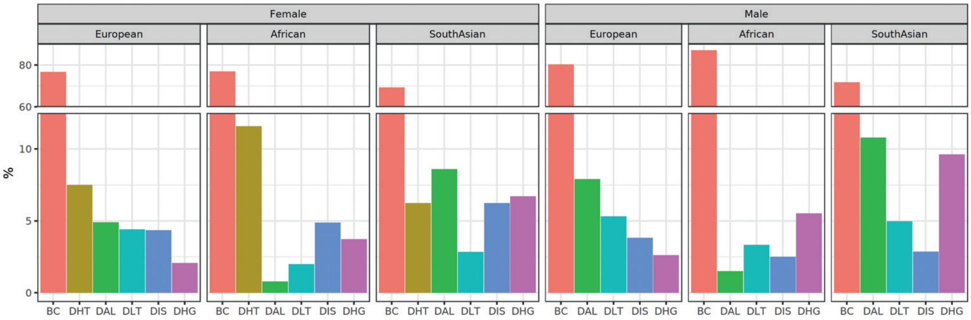


Figure 4.13. Distribution of concordant and discordant profiles by ancestry in the UK Biobank.
Each bar represents the weighted proportion of individuals mapped to the profiles identified in the European subset of the UK Biobank.

Paper IV

In this Paper we found linear associations of genetically predicted BMI with T2D, hypertension (HTN) and CAD, but not with CKD or stroke. Positive effects were found for glycaemia, TG and blood pressure, and inverse effects were found for total cholesterol, HDL and LDL. Sex-stratified analyses showed a stronger positive effect of BMI on CAD and a stronger negative effect on LDL in men compared to women (Figure 4.14).

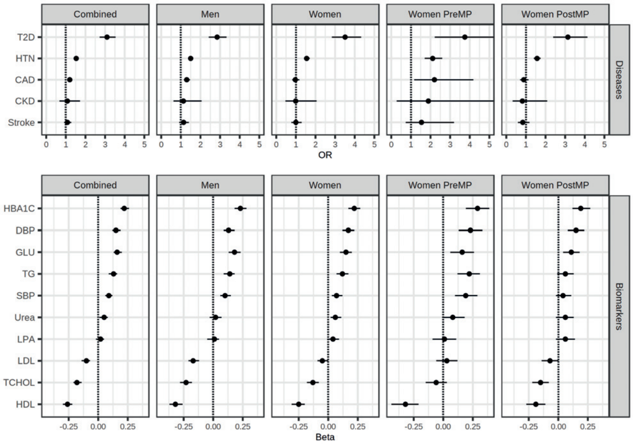


Figure 4.14. Causal estimates of BMI on selected cardiometabolic outcomes.
Odds ratio and standardised beta coefficients on each outcome per standard deviation unit of BMI. Lines represent 95% confidence intervals

Here we show the results of the doubly ranked nonlinear MR analyses. As found with the original methodology, we found consistent evidence to support a nonlinear causal effect of BMI on glycaemia, with at least three of the four tests of nonlinearity surpassing the significance threshold in the combined and the sex-specific analyses (Figures 4.15 – 4.17).

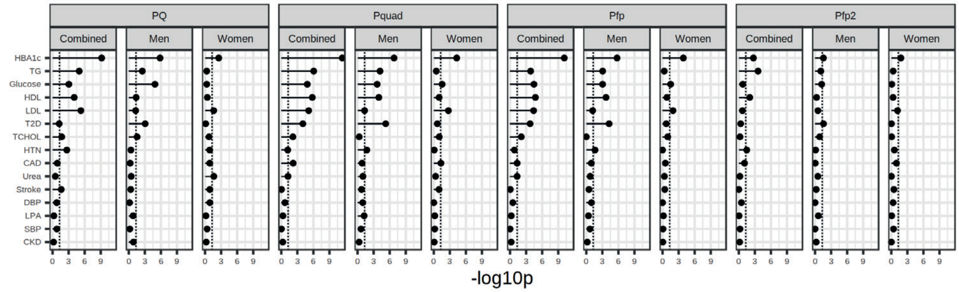


Figure 4.15. Tests for nonlinearity in the doubly ranked method.
PQ is the heterogeneity test across causal estimates from each quantile. Pquad is the significance of a quadratic model. Pfp is the significance of the fractional polynomial model. Pfp2 is the significance of a fractional polynomial model with two degrees compared to a simpler model with one degree. The dotted line represent -log10(0.05).

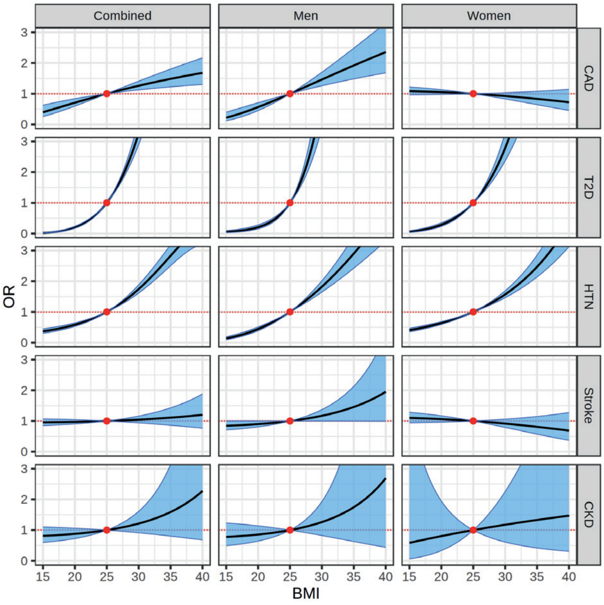


Figure 4.16. Shape of causal relationships of BMI on selected cardiometabolic outcomes.
Each line shows the odds ratio compared with the reference point at 25 kg/m^2. Shaded areas are 95% confidence intervals.

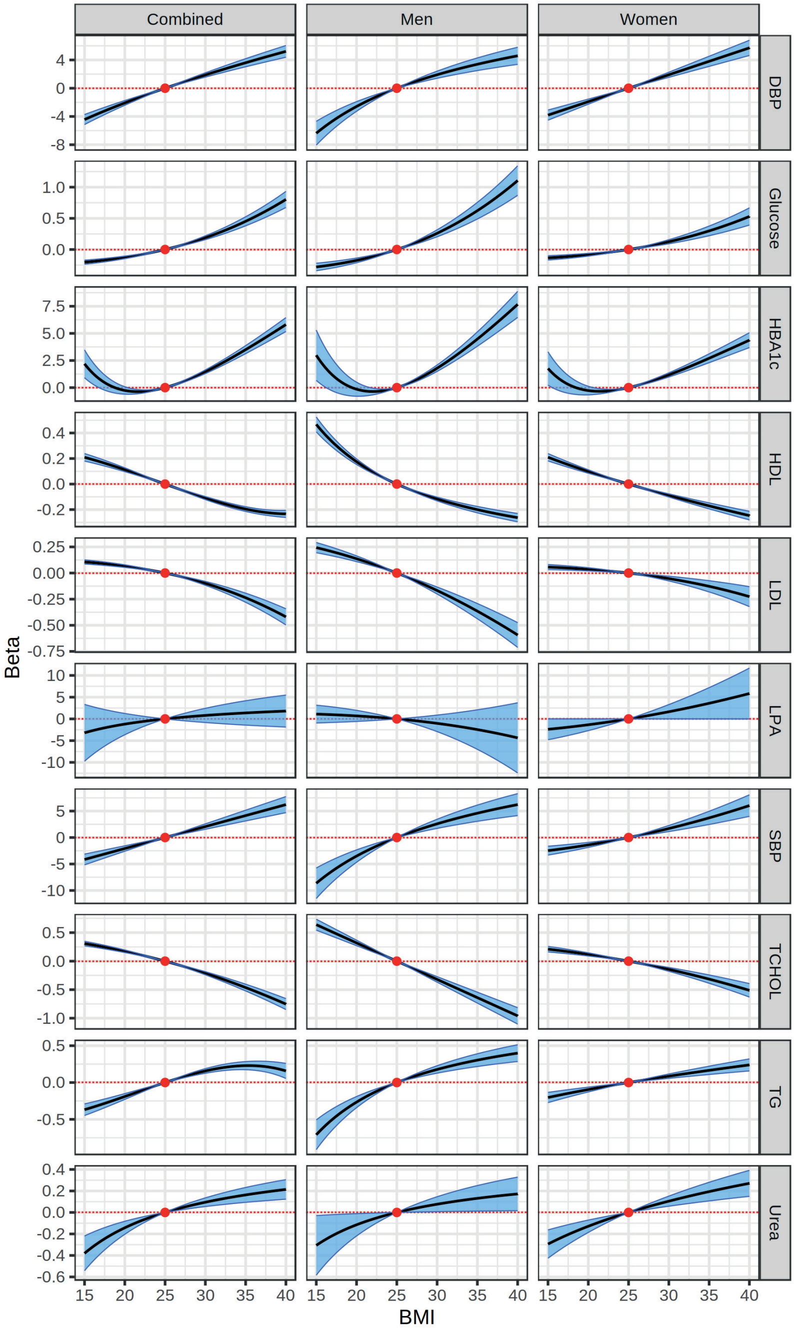


Figure 4.17. Shape of the causal relationships of BMI to continuous cardiometabolic traits. Each line shows the odds ratio compared with the reference point at 25 kg/m^2. Shaded areas are 95% confidence intervals.

Chapter V – Discussion

BMI-T2D discordance

Our genetic decomposition of the obesity phenotype into distinct diabetogenic or antidiabetogenic profiles facilitates the identification of crucial factors that either link or detach obesity from metabolic risks, which may help understand better the underlying mechanisms driving heterogeneity in obesity.

Our results highlight adipose distribution as a core feature of discordant obesity. We provide genetic evidence that an obesity profile that is less prone to accumulate adiposity in the visceral relative to the subcutaneous compartment in the hip, as indicated by a lower WHR, have lower risk of T2D compared to a contrasting obesity profile with higher WHR. Many factors have been proposed for this disparate effect of visceral versus subcutaneous adiposity on metabolic health. A positive energy balance that exceeds the capacity to store fat in the subcutaneous compartment drives visceral adipose accumulation, as seen in individuals with lipodystrophy (147). In addition, multiple lines of evidence suggest a biological difference in the adipose tissue from these two compartments. Animal studies have shown that transplanting subcutaneous, but not visceral fat, can improve the metabolic disturbances in obesity (148). Compared to subcutaneous adipose tissue, visceral adipose tissue responds less to insulin and preadipocyte differentiation stimulation but responds more to a more labile and prone to catecholamine-induced lipolysis. It is also more susceptible to immune cell infiltration and inflammation (149). Additionally, products from visceral adipose tissue may have a more direct effect on liver metabolism because of its drainage through the portal vein, which might underlie the difference we observe in biomarkers of liver failure between the two obesity profiles (150). Interestingly, due to its immune function, higher visceral fat can provide protection against intestinal infections in early life, but those who benefit from this protection might be more vulnerable to metabolic disturbances in adulthood (151).

Our results also highlight the importance of a sufficient vascular supply and a supportive extracellular matrix (EMC) remodelling in adipose tissue expansion without detrimental metabolic effects. We show that a genetically determined obesity profile coupled with better vascular function, reflected by lower SBP, have lower risk of T2D compared to its hypertensive counterpart. Additionally, we found that the expression of two proteins associated with extracellular matrix remodelling, HS6ST2 and TIMP4, are associated with a discordant phenotype in obesity. It is possible that both vascularisation capacity and EMC remodelling have roles in fat accumulation in subcutaneous and visceral compartments (152). Moreover, vascular dysfunction, although a consequence of T2D, can also has precede metabolic perturbations through its effects on nutrient and hormonal flux, not only in adipose tissue, but also in other tissues, such as pancreas and muscle (153).

Our analyses using gene expression that is determined by genetic variation provide potential targets that could be used to activate discordant metabolic processes to shift obesity into a more discordant phenotype. Some of the genes identified have already known interactions with currently used medications, such as metformin, thiazolidinediones and sulphonylureas. However, the mechanisms through which many of these targets contribute to discordance, and their medication interactions, are yet to be characterised.

In comparison to previous strategies to characterize the discordance between BMI and metabolic risk, our analysis is not constrained to a subset of traits selected a priori, but rather we determine the differential phenotypic structure of discordance in data-driven, agnostic manner across many phenotypic layers. Our strategy also enables profile comparisons using data generated by multiple datasets, rather than restricting to measurements in a single cohort. Both analytical choices enhance statistical power and minimise cohort-specific biases, which could be anticipated if the analyses were performed in a single cohort. Furthermore, using germline DNA variation helps mitigate reverse causality and other sources of confounding that hamper the interpretation of the differences between the obesity profiles.

Genetic discordance was defined using data from datasets of predominantly European ancestry, which potentially limits the transferability of our results to other populations. This choice that was dictated by the larger sample sizes available in this ancestral group, which facilitated the identification of concordant and discordant signals, and the availability of genetic associations with other multiple traits in our phenome-wide scans. This also mitigates the potential risk of spurious findings due to heterogeneity in allele frequencies across studies. Nonetheless, we found consistent results in African descent populations in BioVU, providing evidence of the trans-ethnic importance of BMI-T2D discordance.

T2D-CVD discordance

Similar to Paper I, in Paper II we decompose the diabetic phenotype into distinct genetic profiles with either susceptibility to or protection against CVD to identify mechanisms exacerbating or mitigating CVD risk in T2D, with additional analyses that show a potential role of quantifying genetic concordance and discordance for CVD risk stratification.

The use of univariate GWAS results has shown value for prediction of chronic diseases, including CVD (154). Further benefits might be achieved by partitioning the genetic predisposition of a disease represented by a single PRS into multiple complementary PRSs that represent distinct profiles and potential etiological bases of disease (155). This might provide a better understanding of the underlying processes leading to disease development in an individual, which can then help improve delivery of appropriate care. We found that quantifying highly concordant and highly discordant genetic predispositions improve CVD prediction in the general population, though likely not by modifying the risk ranks already given by traditional risk factors, but by giving better calibrated predictions. It is, however, important to note that our genetic stratification focuses on capturing variation in the predisposition to diabetic CVD, which is a subset (~30%) of total CVD. Our interaction analysis shows that the cardioprotective benefit of diabetes prevention is likely to be positively correlated with discordance.

The higher predictive value of concordant and discordant quantification in men than women is consistent with the higher heritability estimates of CVD reported previously in men than in women (156). This finding might be driven by the use of non-sex-specific estimates from GWAS to derive the PRSs, the enrichment of male cases in the cohorts used to discover genetic signals for CVD and differential participation of healthier women (157).

The comparative analyses across the phenome show that genetically determined T2D-CAD discordance is indicative of a broader T2D-systemic atherosclerosis discordance. In the subsequent causal inference analyses we found that the key mechanisms driving this discordance are the content of free cholesterol and phospholipids within small VLDL particles, as well as lower levels of ApoB (the main lipoprotein of VLDL) in relation to ApoA1 (the main lipoprotein of HDL). These findings align with previous investigations of genetic discordance between T2D and LDL as an intermediate phenotype of CVD (158), which underscore the central role of lipolytic remodelling of VLDL particles in peripheral tissues and their clearance from plasma in the discordant phenotype (159).

We prioritized eight loci with targets potentially capable of ameliorating cardiovascular risk in T2D. Two of these loci include targets of currently available drugs with established cardioprotective benefits, namely statins and GLP-1RAs. While statins have been shown to exert its cardioprotective effects primarily by lowering LDL cholesterol levels, they have been associated with higher glycemia, through mechanisms that are not well understood, and might include both interference with insulin secretion and insulin sensitivity (160). Nonetheless, it has been shown that they also decrease VLDL production in the liver and subsequent reductions in circulating VLDL levels (161). Similarly, GLP-1RAs decrease VLDL production in the liver while enhancing VLDL clearance from circulation, leading to improvements in lipid profiles and a reduction in cardiovascular risk (162).

Stratification approaches in diabetes, both using genetic and phenotypic data, often centre around intermediate traits. Instead, we focused on identifying subgroups based directly on the adverse clinical outcomes associated with diabetes. As described in Paper I, by leveraging genetic data, we minimize the risks of confounding and reverse causality. However, also like in Paper I, we based our genetic stratification on signals found in European populations. We show using data from the ORIGIN trial that at higher risk individuals of Latin American ancestry, there were still divergent associations with prevalence of MACE, which provides evidence of trans-ethnic importance of these profiles. However, as our analysis of the ORIGIN trial shows, more research is needed to identify genetic associations to secondary traits, while incorporating the potential of collider bias.

Observational BMI-biomarker discordance

In Paper III we focused on intermediate biomarkers of risk and defined 5 phenotypic profiles defined by specific patterns of biomarker discordance with BMI. These discordant profiles were robustly replicated across four independent large-scale population-based cohorts. The estimated weighted proportion of individuals having these profiles is around 20% of the general population. We found that these profiles convey distinct CVD and diabetes risks when compared to the more common concordant phenotypic profile, highlighting the considerable degree of heterogeneity in the relationship between BMI and cardiometabolic risk.

Enhancement of MACE prediction appears to be driven by the quantification of BMI discordance with lipid fractions, reflected by the DAL profile. This profile resembles the phenotype of familial combined hyperlipidaemia, characterised by disproportionate elevations of TG-rich lipoproteins for the same increases in adiposity, with accompanying higher levels of atherogenic small dense LDL particles and lower levels of HDL (163). Individuals with a DAL profile had a lower prevalence of MACE at baseline and were more commonly unmedicated, indicating that quantifying discordance might be useful for early identification and prevention of cardiovascular events in this subgroup of individuals.

The DHG profile, though enriched for individuals with prior history of multimorbidity and a higher risk of incident diabetes, was not associated with higher MACE incidence compared with the BC profile. Furthermore, within the DHG profile there is an inverse correlation between glycaemia and LDL. Thus, it is possible that this profile resembles the phenotypic signature of the T2D-CAD discordance described in Paper II.

Similarly, the DLT profile, which is associated with higher ALT, had no association with diabetes progression and had lower risk of MACE compared to the concordant profile. This is in line with findings in the large-scale NHANES population survey in the US, showing that ALT was positively associated with risk of diabetes-related CAD – most likely representing the concordant profile in our study – but was inversely associated with risk of CAD that was not attributable to diabetes (164).

Among the sex differences observed, the overall linear estimate of the increase in blood pressure per BMI unit was greater in women than men, as has been previously observed, even after adjusting for menopause (165). Moreover, a discordant hypertensive profile was identified in women but not in men, where the BMI-blood pressure association was enhanced. Nonetheless, the incidence of MACE in this profile was not significantly different than in the concordant profile.

Although we found similar estimates of MACE and diabetes risk associated with the discordant profiles in men and women, we found differences in the added predictive value of the profiles. In MACE, the added value was higher in men than women, possibly due to higher rates in men, but also because of the same limitations of healthy volunteer bias being stronger in women, as exposed in Paper II. Notably, we found that females in the DIS profile were classified less accurately when discordant profile information was incorporated in the predictive models, in contrast to the male counterpart where classification improved. Women tend to have higher CRP levels than men, a stronger relationship between adiposity and CRP levels, and a stronger relationship between CRP and fat distribution (166).

Conversely, in diabetes progression we found higher gains in predictive ability in women than men. Previous studies have shown that discrimination of models for diabetes progression generally perform better in women than men, especially when including anthropometric measures (167). Our study suggest that predictive ability can be further enhanced if discordance between anthropometric measures and other risk factors are considered.

Our strategy to define discordant subgroups applies nonlinear dimension reduction techniques to large-scale datasets, revealing the distribution of multivariate data without the constraints of linear assumptions. Similar techniques have been successfully utilized in the dissection of the clinical heterogeneity of patients with T2D at diagnosis (63). These techniques have the advantage of not assigning categorical labels to individuals. Categorical assignment has the problem of assuming that individuals within a category can be treated as a homogeneous group. It also ignores the possibility that some individuals at the boundary between two groups might share features of the two groups. Additionally, measurement error errors due to intra-individual variability is largely ignored in categorical allocations, while this is better handled in probabilistic allocations by incorporating allocation error. Thus, our approach offers a more nuanced representation of phenotypic discordance. Furthermore, this enables the evaluation of the effect of subtle discordances even within the concordant profile. According to our decision curve analyses, this is where the highest benefit of discordance quantification is observed.

Although our study included four large independent cohorts and the profiles identified were successfully replicated across all cohorts, the sample size of discordant profiles was small, which limits the statistical power of our analyses. This might also explain why we were not able to replicate certain subgroups. Notably, none of the profiles identified resembled a category of 'protective' discordance, where biomarker levels are at lower levels than the expected for the BMI. While this might imply that such a phenotypic pattern is likely to be part of the normal distribution around the concordant profile, rather than a discrete phenotype, better separation of this and other profiles can potentially be achieved through a more comprehensive biomarker assessment.

We used biomarker data from European populations, mainly because this was the predominant ancestry with data available across discovery and replication cohorts. We show nonetheless that taking the European discordance distribution as reference, we could assess the pattern of discordance in African and South Asian populations in UK Biobank. The samples sizes of these populations are much less than the European and are therefore not adequately powered to detect some of the discordant profiles.

Overall, sex-specific, and nonlinear BMI effects

In this study, we investigated the linear effect of BMI on multiple cardiometabolic outcomes, and the heterogeneity of causal estimates stratified by sex and age, and at different levels of BMI, deriving estimated shapes of these causal relationships. Overall, we found widespread effects of BMI on multiple biomarkers and disease phenotypes, with consistent positive effects on CAD, T2D, HTN and an unfavourable lipid profile characterised by lower HDL and higher TG, reflecting the predominant concordant obesity phenotype characterised in Paper I.

In the comparison of causal effects between sexes, we found that higher BMI is linked to CAD in men, but this connection is attenuated in women. Specifically, postmenopausal women show the most significant decrease in this association. Previous prospective studies have also noted that after menopause, central adiposity, rather than overall adiposity reflected by BMI, remains linked to CAD in women (168). Nonetheless, BMI was negatively associated with LDL, a major risk factor for CAD, and this association was more negative in men than women. It has been previously observed that individuals who have CAD despite having low levels of LDL are more likely to be male, and also have a worse overall profile, with higher glycaemia and triglycerides and lower HDL, compared to individuals with CAD and higher levels of LDL (169). This is likely a consequence of changes in lipid subfractions seen in obesity, with the predominance of small, dense LDL over larger LDL particles, a change that is more pronounced in men than women (170), and strongly associated with higher risk of CVD (171).

Our results highlight a nonlinear association between BMI and glycaemia. Similar findings were derived using orthogonal nonlinear MR techniques (172). The curve increases rapidly over a BMI of 25 kg/m^2 and is steeper in men than women, consistent with observational and overall MR estimates (173), indicating more detrimental effect on glycemia at higher BMIs, and also implying a beneficial effect of modest reductions in weight at higher BMIs (174). At the lower end of the BMI spectrum, estimates from the doubly ranked method show that the curve for glucose flattened while the curve for HbA1c was more variable and tended upwards. The interpretation of these effects is challenging because the segment of the population below the normal BMI range can include individuals with comorbidities associated with wasting that could affect glucose metabolism (175). However, it can also include lean cases of T2D, which occur more frequently in men, and is usually characterised by postprandial glucose peaks rather than sustained hyperglycaemia, which could be a cause of discordance between HbA1c and glucose (176,177).

It has been recently recognised that the original approach of nonlinear MR applied in Paper IV, referred to as the residual method, can produce biased estimates, especially when the assumption of a constant genetic effect on the exposure across strata in the population is violated (130). The doubly ranked method intends to solve through its ranking method which effectively controls the heterogeneity in the exposure association, while still obtaining strata where the average level of the exposure is increasing (131). The primary results from the residual method were found to be robust to changes in the approach, which enhances the validity of our findings. Both methods are, nonetheless, sensitive to selection bias in the population under study, which can give nonsensical results, such as causal associations of BMI with age and sex that varies across BMI strata (178). Given that this issue is less severe as samples are more representative of the source population (157,175), replication in unselected samples are needed to properly evaluate the external validity of our findings.

Summary and conclusions

Obesity profiles with either diabetogenic or antidiabetogenic proclivities reveal distinctive etiological subtypes, with key differences in fat distribution, blood pressure and cholesterol content in HDL particles. We identify proteins related to extracellular matrix remodelling as potential mediators of discordance in obesity and prioritise 17 genes potentially involved in the molecular mechanisms of discordance, involving pleiotropic effects across multiple tissues.

Similarly, we also identify two distinct T2D profiles with contrasting CVD risk, highlighting the key role of VLDL metabolism in separating T2D from CVD risk, and identifying eight discordant loci through bioinformatic evaluation, warranting subsequent validation in experimental investigations. We show that adding concordant and discordant predispositions can improve the predictive ability of current stratification models for CVD, especially in men, where 8% of people are reclassified into more appropriate risk categories, supporting the use of partitioned PRS in classifying individuals according to the etiological source of their disease.

We also identified five distinct phenotypic profiles exhibiting diverse relationships between BMI and risk biomarkers and varying degrees of CVD and diabetes risks, reflecting substantial heterogeneity in the link between BMI and risk. Conceptualising these as different subtypes of obesity requires further validation, but incorporating phenotypic discordance with BMI enhances the prediction of MACE and diabetes progression in the general population.

Finally, we use genetically driven causal inference analysis to estimate the overall and sex-specific effect of BMI on multiple outcomes, finding consistent positive effects on T2D, CAD, HTN and in dyslipidaemia. We found differences in the causal effect of BMI on CAD, with stronger effects in men than women, and a non-linear relationship BMI and glycaemia, with stronger positive effects at higher BMIs.

These studies highlight the diverse nature of cardiometabolic conditions by through genetic and phenotypic profilin, offering insights into the underlying mechanisms and clinical implications of this diversity. By recognizing individual variability, integrating these findings into clinical care could lead to more precise and tailored approaches in managing cardiometabolic diseases, potentially enhancing clinical care.

Future perspectives

Our work on defining discordance using genetics can be extended beyond cardiometabolic diseases, which could help gaining insights into the factors that give rise to more resilient or more susceptible phenotypes, while also informing pathways and therapeutic targets that could be used to uncouple diseases from its complications. Such an extension could include autoimmune diseases, neurological disorders, and cancer. Additionally, by investigating discordant phenotypes in diverse ancestral populations, we can assess the generalizability of the observed discordances and uncover population-specific factors influencing disease susceptibility and progression.

Comprehensive functional characterisation of discordant genetic variants holds promise in elucidating underlying biological mechanisms and molecular pathways contributing to phenotypic discordance. Two targets identified in our analyses (TIMP4 and DNAH10) are the subject of functional validation by other members of our research group. Integration of genomics and other omics data with functional experiments may reveal novel pathways underlying divergent disease outcomes, paving the way for targeted interventions and precision medicine approaches.

Investigating treatment response in individuals with discordant phenotypes offers insights into the effectiveness of interventions tailored to specific discordant profiles. Individuals with concordant and discordant profiles may receive the same treatment, such as both concordant and discordant obesity profiles having similar odds of undergoing bariatric surgery. Due to their underlying physiological differences, it is possible that their response varies. I am currently collaborating with a consortium investigating the genetic factors of bariatric surgery outcomes in thousands of individuals, which could give a unique opportunity to test these hypotheses.

The development and validation of partitioned polygenic scores based on discordant profiles can potentially enhance risk prediction while improving the understanding of the aetiology of disease. Integrating genetic data representing distinct phenotypic profiles may improve the accuracy and specificity of polygenic risk scores for personalized risk assessment, facilitating early detection and intervention strategies.

Utilizing nonlinear techniques allow for the exploration of complex relationships between genetic factors, biomarkers, and disease outcomes, which are unlikely to be uniformly linear. In particular, our nonlinear dimension reduction technique can be extended to incorporate additional biomarkers and diverse ancestries, improving the understanding of phenotypic heterogeneity, and informing targeted interventions and precision medicine approaches.

I am also involved in ongoing projects that explore the impact of longitudinal trends on heterogeneity in cardiometabolic diseases, characterising distinct trajectory patterns that could be used to better distinguish individuals with rapid changes leading to more severe disease phenotypes. In this matter, continuous monitoring technologies, and improved phenotyping methods is essential. Long-term follow-up studies and comprehensive phenotypic profiling can elucidate the dynamic nature of disease progression and identify novel prognostic markers and therapeutic targets, ultimately enhancing patient outcomes and public health.

I believe that understanding average trends and relationships is important to deliver effective interventions at the population level. However, as the papers in this thesis show, there are individuals who deviate from these expected trends. As we advance in data collection and analytical capabilities, my vision for the future is clinical care that effectively incorporates this information, potentially moving away from categorisation using hard clinical entities to a multidimensional and probabilistic approach to treatment.

References

1. DeBerardinis RJ, Thompson CB. Cellular Metabolism and Disease: What Do Metabolic Outliers Teach Us? Cell. 2012 Mar 16;148(6):1132–44.
2. Wang T, Hung CCY, Randall DJ. THE COMPARATIVE PHYSIOLOGY OF FOOD DEPRIVATION: From Feast to Famine. Annual Review of Physiology. 2006;68(1):223–51.
3. Cordain L, Eaton SB, Sebastian A, Mann N, Lindeberg S, Watkins BA, et al. Origins and evolution of the Western diet: health implications for the 21st century1,2. The American Journal of Clinical Nutrition. 2005 Feb 1;81(2):341–54.
4. Speakman JR, Elmquist JK. Obesity: an evolutionary context. life metab. 2022 Nov 3;1(1):10–24.
5. Brunstrom JM, Drake ACL, Forde CG, Rogers PJ. Undervalued and ignored: Are humans poorly adapted to energy-dense foods? Appetite. 2018 Jan 1;120:589–95.
6. Pontzer H, Wood BM, Raichlen DA. Hunter-gatherers as models in public health. Obesity Reviews. 2018;19(S1):24–35.
7. The Global Status Report on Physical Activity 2022 [Internet]. [cited 2023 Dec 14]. Available from: https://www.who.int/teams/health-promotion/physical-activity/global-status-report-on-physical-activity-2022
8. The carbohydrate-insulin model: a physiological perspective on the obesity pandemic. The American Journal of Clinical Nutrition. 2021 Dec 1;114(6):1873–85.
9. Sumińska M, Podgórski R, Bogusz-Górna K, Skowrońska B, Mazur A, Fichna M. Historical and cultural aspects of obesity: From a symbol of wealth and prosperity to the epidemic of the 21st century. Obesity Reviews. 2022;23(6):e13440.
10. Haslam D. Obesity: a medical history. Obesity Reviews. 2007;8(s1):31–6.
11. Medico-Actuarial Mortality Investigation. Obesity Research. 1995;3(1):100–6.
12. Keys A, Fidanza F, Karvonen MJ, Kimura N, Taylor HL. Indices of relative weight and obesity. Journal of Chronic Diseases. 1972 Jul 1;25(6):329–43.
13. Quetelet LAJ. A Treatise on Man and the Development of his Faculties [Internet]. Smibert T, editor. Cambridge: Cambridge University Press; 2013 [cited 2024 Jan 10]. (Cambridge Library Collection - Philosophy). Available from: https://www.cambridge.org/core/books/treatise-on-man-and-the-development-of-his-faculties/AB13A647A6C8727C06AE5399D7422887
14. Eknoyan G. A History of Obesity, or How What Was Good Became Ugly and Then Bad. Advances in Chronic Kidney Disease. 2006 Oct 1;13(4):421–7.

15. Abarca-Gómez L, Abdeen ZA, Hamid ZA, Abu-Rmeileh NM, Acosta-Cazares B, Acuin C, et al. Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults. The Lancet. 2017 Dec 16;390(10113):2627–42.

16. Obesity and overweight [Internet]. [cited 2024 Jan 10]. Available from: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

17. Institute for Health Metrics and Evaluation. GBD Compare Data Visualization [Internet]. Seattle, WA: IHME, University of Washington; 2020. Available from: http://vizhub.healthdata.org/gbd-compare

18. World Obesity Atlas 2023 [Internet]. World Obesity Federation; 2023. Available from: https://data.worldobesity.org/publications/?cat=19

19. Danaei G, Finucane MM, Lu Y, Singh GM, Cowan MJ, Paciorek CJ, et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2·7 million participants. The Lancet. 2011 Jul 2;378(9785):31–40.

20. Reed JA. Aretaeus, the Cappadocian History Enlightens the Present. Diabetes. 1954 Sep 1;3(5):419–21.

21. Polonsky KS. The Past 200 Years in Diabetes. New England Journal of Medicine. 2012 Oct 4;367(14):1332–40.

22. Barach JH. Historical Facts in Diabetes. Ann Med Hist. 1928 Dec;10(4):387–401.

23. Sims EK, Carr ALJ, Oram RA, DiMeglio LA, Evans-Molina C. 100 years of insulin: celebrating the past, present and future of diabetes therapy. Nat Med. 2021 Jul;27(7):1154–64.

24. Himsworth HP. DIABETES MELLITUS: ITS DIFFERENTIATION INTO INSULIN-SENSITIVE AND INSULIN-INSENSITIVE TYPES. The Lancet. 1936 Jan 18;227(5864):127–30.

25. Ong KL, Stafford LK, McLaughlin SA, Boyko EJ, Vollset SE, Smith AE, et al. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. The Lancet. 2023 Jul 15;402(10397):203–34.

26. International Diabetes Federation. IDF Diabetes Atlas, 10th edn [Internet]. Brussels, Belgium; [cited 2024 Jan 14]. Available from: https://diabetesatlas.org

27. Abdullah A, Peeters A, de Courten M, Stoelwinder J. The magnitude of association between overweight and obesity and the risk of diabetes: a meta-analysis of prospective cohort studies. Diabetes Res Clin Pract. 2010 Sep;89(3):309–19.

28. Centers for Disease Control and Prevention (CDC). Prevalence of overweight and obesity among adults with diagnosed diabetes--United States, 1988-1994 and 1999-2002. MMWR Morb Mortal Wkly Rep. 2004 Nov 19;53(45):1066–8.

29. Astrup A, Finer N. Redefining Type 2 diabetes: ‘Diabesity’ or ‘Obesity Dependent Diabetes Mellitus’? Obesity Reviews. 2000;1(2):57–9.

30. Klein S, Gastaldelli A, Yki-Järvinen H, Scherer PE. Why Does Obesity Cause Diabetes? Cell Metab. 2022 Jan 4;34(1):11–20.

31. Borén J, Taskinen MR, Olofsson SO, Levin M. Ectopic lipid storage and insulin resistance: a harmful relationship. Journal of Internal Medicine. 2013;274(1):25–40.

32. Gerst F, Wagner R, Kaiser G, Panse M, Heni M, Machann J, et al. Metabolic crosstalk between fatty pancreas and fatty liver: effects on local inflammation and insulin secretion. Diabetologia. 2017 Nov 1;60(11):2240–51.

33. Perakakis N, Farr OM, Mantzoros CS. Leptin in Leanness and Obesity: JACC State-of-the-Art Review. Journal of the American College of Cardiology. 2021 Feb 16;77(6):745–60.

34. Sjöström L. Review of the key results from the Swedish Obese Subjects (SOS) trial – a prospective controlled intervention study of bariatric surgery. Journal of Internal Medicine. 2013;273(3):219–34.

35. Bhaskaran K, dos-Santos-Silva I, Leon DA, Douglas IJ, Smeeth L. Association of BMI with overall and cause-specific mortality: a population-based cohort study of 3·6 million adults in the UK. The Lancet Diabetes & Endocrinology. 2018 Dec 1;6(12):944–53.

36. Raghavan S, Vassy JL, Ho Y, Song RJ, Gagnon DR, Cho K, et al. Diabetes Mellitus–Related All-Cause and Cardiovascular Mortality in a National Cohort of Adults. Journal of the American Heart Association. 2019 Feb 19;8(4):e011295.

37. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. New England Journal of Medicine. 2017 Jul 6;377(1):13–27.

38. Einarson TR, Acs A, Ludwig C, Panton UH. Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of scientific evidence from across the world in 2007–2017. Cardiovasc Diabetol. 2018 Dec;17(1):1–19.

39. Bays HE, Toth PP, Kris-Etherton PM, Abate N, Aronne LJ, Brown WV, et al. Obesity, adiposity, and dyslipidemia: A consensus statement from the National Lipid Association. Journal of Clinical Lipidology. 2013 Jul 1;7(4):304–83.

40. Hall JE, do Carmo JM, da Silva AA, Wang Z, Hall ME. Obesity-Induced Hypertension. Circulation Research. 2015 Mar 13;116(6):991–1006.

41. Engin A. Endothelial Dysfunction in Obesity. In: Engin AB, Engin A, editors. Obesity and Lipotoxicity [Internet]. Cham: Springer International Publishing; 2017 [cited 2024 Jan 15]. p. 345–79. (Advances in Experimental Medicine and Biology). Available from: https://doi.org/10.1007/978-3-319-48382-5_15

42. Low Wang CC, Hess CN, Hiatt WR, Goldfine AB. Clinical Update: Cardiovascular Disease in Diabetes Mellitus. Circulation. 2016 Jun 14;133(24):2459–502.

43. Ohishi M. Hypertension with diabetes mellitus: physiology and pathology. Hypertens Res. 2018 Jun;41(6):389–93.

44. Wu L, Parhofer KG. Diabetic dyslipidemia. Metabolism. 2014 Dec 1;63(12):1469–79.

45. Mechanick JI, Farkouh ME, Newman JD, Garvey WT. Cardiometabolic-Based Chronic Disease, Adiposity and Dysglycemia Drivers: JACC State-of-the-Art Review. Journal of the American College of Cardiology. 2020 Feb 11;75(5):525–38.

46. Barilla S, Treuter E, Venteclef N. Transcriptional and epigenetic control of adipocyte remodeling during obesity. Obesity. 2021;29(12):2013–25.

47. Cardiovascular Effects of Intensive Lifestyle Intervention in Type 2 Diabetes. New England Journal of Medicine. 2013 Jul 11;369(2):145–54.

48. Association of the magnitude of weight loss and changes in physical fitness with long-term cardiovascular disease outcomes in overweight or obese people with type 2 diabetes: a post-hoc analysis of the Look AHEAD randomised clinical trial. The Lancet Diabetes & Endocrinology. 2016 Nov 1;4(11):913–21.

49. Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (UKPDS 34). The Lancet. 1998 Sep 12;352(9131):854–65.

50. Davies M, Færch L, Jeppesen OK, Pakseresht A, Pedersen SD, Perreault L, et al. Semaglutide 2·4 mg once a week in adults with overweight or obesity, and type 2 diabetes (STEP 2): a randomised, double-blind, double-dummy, placebo-controlled, phase 3 trial. The Lancet. 2021 Mar 13;397(10278):971–84.

51. Sattar N, Lee MMY, Kristensen SL, Branch KRH, Del Prato S, Khurmi NS, et al. Cardiovascular, mortality, and kidney outcomes with GLP-1 receptor agonists in patients with type 2 diabetes: a systematic review and meta-analysis of randomised trials. The Lancet Diabetes & Endocrinology. 2021 Oct 1;9(10):653–62.

52. Jou C. The Biology and Genetics of Obesity — A Century of Inquiries. New England Journal of Medicine. 2014 May 15;370(20):1874–7.

53. Sims EAH, Danforth E, Horton ES, Bray GA, Glennon JA, Salans LB. Endocrine and Metabolic Effects of Experimental Obesity in Man11Supported in part by U.S. Public Health Service Grants 5 R01 AM 10254 (Dr. Sims), AM 13321 (Dr. Salans), 5 R01 AM 13307 (Dr. Horton), and FR 00109 (Clinical Research Center). In: Greep RO, editor. Proceedings of the 1972 Laurentian Hormone Conference [Internet]. Boston: Academic Press; 1973 [cited 2024 Jan 16]. p. 457–96. (Recent Progress in Hormone Research; vol. 29). Available from: https://www.sciencedirect.com/science/article/pii/B9780125711296500166

54. Leibel RL, Hirsch J. Diminished energy requirements in reduced-obese patients. Metabolism. 1984 Feb 1;33(2):164–70.

55. Pomares-Millan H, Poveda A, Atabaki-Pasdar N, Johansson I, Björk J, Ohlsson M, et al. Predicting Sensitivity to Adverse Lifestyle Risk Factors for Cardiometabolic Morbidity and Mortality. Nutrients. 2022 Jan;14(15):3171.

56. Smith GI, Mittendorfer B, Klein S. Metabolically healthy obesity: facts and fantasies. J Clin Invest. 2019 Oct 1;129(10):3978–89.

57. Wang B, Zhuang R, Luo X, Yin L, Pang C, Feng T, et al. Prevalence of Metabolically Healthy Obese and Metabolically Obese but Normal Weight in Adults Worldwide: A Meta-Analysis. Horm Metab Res. 2015 Oct;47(11):839–45.

58. Neeland IJ, Poirier P, Després JP. Cardiovascular and Metabolic Heterogeneity of Obesity. Circulation. 2018 Mar 27;137(13):1391–406.

59. Karastergiou K, Smith SR, Greenberg AS, Fried SK. Sex differences in human adipose tissues – the biology of pear shape. Biol Sex Differ. 2012 May 31;3:13.

60. Badrick E, Sperrin M, Buchan IE, Renehan AG. Obesity paradox and mortality in adults with and without incident type 2 diabetes: a matched population-level cohort study. BMJ Open Diabetes Research and Care. 2017 Mar 1;5(1):e000369.

61. Banack HR, Stokes A. The 'obesity paradox' may not be a paradox at all. Int J Obes. 2017 Aug;41(8):1162–3.

62. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, 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. 2018 May 1;6(5):361–9.

63. Nair ATN, Wesolowska-Andersen A, Brorsson C, Rajendrakumar AL, Hapca S, Gan S, et al. Heterogeneity in phenotype, disease progression and drug response in type 2 diabetes. Nat Med. 2022 May;28(5):982–8.

64. Wesolowska-Andersen A, Brorsson CA, Bizzotto R, Mari A, Tura A, Koivula R, et al. Four groups of type 2 diabetes contribute to the etiological and clinical heterogeneity in newly diagnosed individuals: An IMI DIRECT study. CR Med [Internet]. 2022 Jan 18 [cited 2023 Feb 9];3(1). Available from: https://www.cell.com/cell-reports-medicine/abstract/S2666-3791(21)00349-9

65. Young KG, McInnes EH, Massey RJ, Kahkoska AR, Pilla SJ, Raghavan S, et al. Treatment effect heterogeneity following type 2 diabetes treatment with GLP1-receptor agonists and SGLT2-inhibitors: a systematic review. Commun Med. 2023 Oct 5;3(1):1–20.

66. Leslie RD, Ma RCW, Franks PW, Nadeau KJ, Pearson ER, Redondo MJ. Understanding diabetes heterogeneity: key steps towards precision medicine in diabetes. The Lancet Diabetes & Endocrinology. 2023 Nov 1;11(11):848–60.

67. Chung WK, Erion K, Florez JC, Hattersley AT, Hivert MF, Lee CG, 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. 2020 Sep 1;63(9):1671–93.

68. Tattersall R, Pyke D, Nerup J. Genetic patterns in diabetes mellitus. Human Pathology. 1980 May 1;11(3):273–83.

69. Mayer J. Genetic Factors in Human Obesity. Annals of the New York Academy of Sciences. 1965;131(1):412–21.

70. Stunkard AJ, Sørensen TIA, Hanis C, Teasdale TW, Chakraborty R, Schull WJ, et al. An Adoption Study of Human Obesity. New England Journal of Medicine. 1986 Jan 23;314(4):193–8.

71. Stunkard AJ, Harris JR, Pedersen NL, McClearn GE. The Body-Mass Index of Twins Who Have Been Reared Apart. New England Journal of Medicine. 1990 May 24;322(21):1483–7.

72. Newman B, Selby JV, King MC, Slemenda C, Fabsitz R, Friedman GD. Concordance for Type 2 (non-insulin-dependent) diabetes mellitus in male twins. Diabetologia. 1987 Oct 1;30(10):763–8.

73. Montague CT, Farooqi IS, Whitehead JP, Soos MA, Rau H, Wareham NJ, et al. Congenital leptin deficiency is associated with severe early-onset obesity in humans. Nature. 1997 Jun;387(6636):903–8.

74. Yeo GSH, Farooqi IS, Aminian S, Halsall DJ, Stanhope RG, O'Rahilly S. A frameshift mutation in MC4R associated with dominantly inherited human obesity. Nat Genet. 1998 Oct;20(2):111–2.

75. Vionnet N, Stoffel M, Takeda J, Yasuda K, Bell GI, Zouali H, et al. Nonsense mutation in the glucokinase gene causes early-onset non-insulin-dependent diabetes mellitus. Nature. 1992 Apr;356(6371):721–2.

76. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, et al. A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. Science. 2007 May 11;316(5826):889–94.

77. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007 Feb;445(7130):881–5.

78. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A saturated map of common genetic variants associated with human height. Nature. 2022 Oct;610(7933):704–12.

79. Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 15 years of GWAS discovery: Realizing the promise. The American Journal of Human Genetics. 2023 Feb 2;110(2):179–94.

80. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. Human Molecular Genetics. 2018 Oct 15;27(20):3641–9.

81. Mahajan A, Spracklen CN, Zhang W, Ng MCY, Petty LE, Kitajima H, et al. Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. Nat Genet. 2022 May;54(5):560–72.

82. Vujkovic M, Keaton JM, Lynch JA, Miller DR, Zhou J, Tcheandjieu C, et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ethnic meta-analysis. Nat Genet. 2020 Jul;52(7):680–91.

83. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017 Jun 15;169(7):1177–86.

84. Farooqi IS, Keogh JM, Yeo GSH, Lank EJ, Cheetham T, O'Rahilly S. Clinical Spectrum of Obesity and Mutations in the Melanocortin 4 Receptor Gene. New England Journal of Medicine. 2003 Mar 20;348(12):1085–95.

85. Schipper M, Posthuma D. Demystifying non-coding GWAS variants: an overview of computational tools and methods. Human Molecular Genetics. 2022 Oct 15;31(R1):R73–83.

86. Loos RJF, Yeo GSH. The genetics of obesity: from discovery to biology. Nat Rev Genet. 2022 Feb;23(2):120–33.

87. Grant AJ, Gill D, Kirk PDW, Burgess S. Noise-augmented directional clustering of genetic association data identifies distinct mechanisms underlying obesity. PLOS Genetics. 2022 Jan 27;18(1):e1009975.

88. Huang LO, Rauch A, Mazzaferro E, Preuss M, Carobbio S, Bayrak CS, et al. Genome-wide discovery of genetic loci that uncouple excess adiposity from its comorbidities. Nat Metab. 2021 Feb;3(2):228–43.

89. Ji Y, Yiorkas AM, Frau F, Mook-Kanamori D, Staiger H, Thomas EL, et al. Genome-Wide and Abdominal MRI Data Provide Evidence That a Genetically Determined Favorable Adiposity Phenotype Is Characterized by Lower Ectopic Liver Fat and Lower Risk of Type 2 Diabetes, Heart Disease, and Hypertension. Diabetes. 2019 Jan 1;68(1):207–19.

90. Yaghootkar H, Lotta LA, Tyrrell J, Smit RAJ, Jones SE, Donnelly L, et al. Genetic Evidence for a Link Between Favorable Adiposity and Lower Risk of Type 2 Diabetes, Hypertension, and Heart Disease. Diabetes. 2016 Aug 1;65(8):2448–60.

91. Udler MS, Kim J, Grotthuss M von, Bonàs-Guarch S, Cole JB, Chiou J, et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. PLOS Medicine. 2018 Sep 21;15(9):e1002654.

92. Kim H, Westerman KE, Smith K, Chiou J, Cole JB, Majarian T, et al. High-throughput genetic clustering of type 2 diabetes loci reveals heterogeneous mechanistic pathways of metabolic disease. Diabetologia. 2023 Mar 1;66(3):495–507.

93. Reales G, Wallace C. Sharing GWAS summary statistics results in more citations. Commun Biol. 2023 Jan 28;6(1):1–6.

94. Elsworth B, Lyon M, Alexander T, Liu Y, Matthews P, Hallett J, et al. The MRC IEU OpenGWAS data infrastructure [Internet]. 2020 Aug [cited 2021 Dec 7] p. 2020.08.10.244293. Available from: https://www.biorxiv.org/content/10.1101/2020.08.10.244293v1

95. Kurilshikov A, Medina-Gomez C, Bacigalupe R, Radjabzadeh D, Wang J, Demirkan A, et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. Nat Genet. 2021 Feb;53(2):156–65.

96. THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020 Sep 11;369(6509):1318–30.

97. Võsa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021 Sep;53(9):1300–10.

98. Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Research. 2020 Jan 8;48(D1):D941–7.

99. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015 Feb;518(7539):317–30.

100. Pers TH, Karjalainen JM, Chan Y, Westra HJ, Wood AR, Yang J, et al. Biological interpretation of genome-wide association studies using predicted gene functions. Nat Commun. 2015 Jan 19;6(1):5890.

101. Freshour SL, Kiwala S, Cotto KC, Coffman AC, McMichael JF, Song JJ, et al. Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Research. 2021 Jan 8;49(D1):D1144–51.

102. Sheils TK, Mathias SL, Kelleher KJ, Siramshetty VB, Nguyen DT, Bologa CG, et al. TCRD and Pharos 2021: mining the human proteome for disease biology. Nucleic Acids Research. 2021 Jan 8;49(D1):D1334–46.

103. Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. Nat Genet. 2021 Apr;53(4):420–5.

104. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203–9.

105. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. American Journal of Epidemiology. 2017 Nov 1;186(9):1026–34.

106. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics. 2010 May 1;26(9):1205–10.

107. Roden D, Pulley J, Basford M, Bernard G, Clayton E, Balser J, et al. Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. Clinical Pharmacology & Therapeutics. 2008;84(3):362–9.

108. What is BioVU? – VICTR – Vanderbilt Institute for Clinical and Translational Research [Internet]. [cited 2024 Jan 23]. Available from: https://victr.vumc.org/what-is-biovu/

109. The ORIGIN Trial Investigators. Rationale, design, and baseline characteristics for a large international trial of cardiovascular disease prevention in people with dysglycemia: The ORIGIN Trial (Outcome Reduction with an Initial Glargine Intervention). American Heart Journal. 2008 Jan 1;155(1):26.e1-26.e13.

110. IMI Innovative Medicines Initiative [Internet]. 2020 [cited 2024 Feb 7]. IMI Innovative Medicines Initiative | SOPHIA | Stratification of obese phenotypes to optimize future obesity therapy. Available from: http://www.imi.europa.eu/projects-results/project-factsheets/sophia

111. Schram MT, Sep SJS, van der Kallen CJ, Dagnelie PC, Koster A, Schaper N, et al. The Maastricht Study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities. Eur J Epidemiol. 2014 Jun 1;29(6):439–51.

112. Ikram MA, Brusselle G, Ghanbari M, Goedegebure A, Ikram MK, Kavousi M, et al. Objectives, design and main findings until 2020 from the Rotterdam Study. Eur J Epidemiol. 2020 May 1;35(5):483–517.

113. Wild PS, Zeller T, Beutel M, Blettner M, Dugi KA, Lackner KJ, et al. Die Gutenberg Gesundheitsstudie. Bundesgesundheitsbl. 2012 Jun 1;55(6):824–30.

114. Harrell FE. General Aspects of Fitting Regression Models. In: Harrell Jr Frank E, editor. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis [Internet]. Cham: Springer International Publishing; 2015 [cited 2024 Feb 7]. p. 13–44. (Springer Series in Statistics). Available from: https://doi.org/10.1007/978-3-319-19425-7_2

115. Harrell FE. Binary Logistic Regression. In: Harrell Jr Frank E, editor. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis [Internet]. Cham: Springer International Publishing; 2015 [cited 2024 Feb 7]. p. 219–74. (Springer Series in Statistics). Available from: https://doi.org/10.1007/978-3-319-19425-7_10

116. Harrell FE. Cox Proportional Hazards Regression Model. In: Harrell Jr Frank E, editor. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis [Internet]. Cham: Springer International Publishing; 2015 [cited 2024 Feb 7]. p. 475–519. (Springer Series in Statistics). Available from: https://doi.org/10.1007/978-3-319-19425-7_20

117. Ebert MH Pim Cuijpers, Toshi Furukawa, David. Doing Meta-Analysis with R: A Hands-On Guide. New York: Chapman and Hall/CRC; 2021. 500 p.

118. Langan D, Higgins JPT, Simmonds M. Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. Research Synthesis Methods. 2017;8(2):181–98.

119. Xiong Z, Gao X, Chen Y, Feng Z, Pan S, Lu H, et al. Combining genome-wide association studies highlight novel loci involved in human facial variation. Nat Commun. 2022 Dec 20;13(1):7832.

120. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet. 2018 Feb;50(2):229–37.

121. Multiple Outcomes or Time-Points within a Study. In: Introduction to Meta-Analysis [Internet]. John Wiley & Sons, Ltd; 2009 [cited 2021 Nov 29]. p. 225–38. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470743386.ch24

122. Genuer R, Poggi JM. Random Forests. In: Genuer R, Poggi JM, editors. Random Forests with R [Internet]. Cham: Springer International Publishing; 2020 [cited 2024 Feb 7]. p. 33–55. (Use R!). Available from: https://doi.org/10.1007/978-3-030-56485-8_3

123. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. Journal of Statistical Software. 2010 Sep 16;36:1–13.

124. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020 Sep;15(9):2759–72.

125. Taliun SAG, Evans DM. Ten simple rules for conducting a mendelian randomization study. PLOS Computational Biology. 2021 Aug 12;17(8):e1009238.

126. de Leeuw C, Savage J, Bucur IG, Heskes T, Posthuma D. Understanding the assumptions underlying Mendelian randomization. Eur J Hum Genet. 2022 Jun;30(6):653–60.

127. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016 May;48(5):481–7.

128. Wu Y, Zeng J, Zhang F, Zhu Z, Qi T, Zheng Z, et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nat Commun. 2018 Mar 2;9(1):918.

129. Staley JR, Burgess S. Semiparametric methods for estimation of a nonlinear exposure-outcome relationship using instrumental variables with application to Mendelian randomization. Genetic Epidemiology. 2017;41(4):341–52.

130. Burgess S. Violation of the Constant Genetic Effect Assumption Can Result in Biased Estimates for Non-Linear Mendelian Randomization. Human Heredity. 2023 Aug 31;88(1):79–90.

131. Tian H, Mason AM, Liu C, Burgess S. Relaxing parametric assumptions for non-linear Mendelian randomization using a doubly-ranked stratification method. PLOS Genetics. 2023 Jun 30;19(6):e1010823.

132. Mahalanobis Distance. In: The Concise Encyclopedia of Statistics [Internet]. New York, NY: Springer; 2008 [cited 2024 Feb 7]. p. 325–6. Available from: https://doi.org/10.1007/978-0-387-32833-1_240

133. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software. 2018 Sep 2;3(29):861.

134. Newman MEJ. Finding community structure in networks using the eigenvectors of matrices. Phys Rev E. 2006 Sep 11;74(3):036104.

135. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019 Mar 26;9(1):5233.

136. Raftery LS Chris Fraley, T Brendan Murphy, Adrian E. Model-Based Clustering, Classification, and Density Estimation Using mclust in R. New York: Chapman and Hall/CRC; 2023. 268 p.

137. Jedidi K, Ramaswamy V, Desarbo WS. A maximum likelihood method for latent class regression involving a censored dependent variable. Psychometrika. 1993 Sep 1;58(3):375–94.

138. Eugster MJA, Leisch F. From Spider-Man to Hero — Archetypal Analysis in R. Journal of Statistical Software. 2009 Apr 29;30:1–23.

139. McInnes L, Healy J, Astels S. hdbscan: Hierarchical density based clustering. Journal of Open Source Software. 2017 Mar 21;2(11):205.

140. Harrell FE. Overview of Maximum Likelihood Estimation. In: Harrell Jr Frank E, editor. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis [Internet]. Cham: Springer International Publishing; 2015 [cited 2023 Sep 29]. p. 181–217. (Springer Series in Statistics). Available from: https://doi.org/10.1007/978-3-319-19425-7_9

141. Coenders G, Pawlowsky-Glahn V. On interpretations of tests and effect sizes in regression models with a compositional predictor. SORT-Statistics and Operations Research Transactions. 2020 Jun 26;44(1):201–20.

142. SCORE2 working group and ESC Cardiovascular risk collaboration. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. European Heart Journal. 2021 Jul 1;42(25):2439–54.

143. Therneau TM, Atkinson E. The concordance statistic. A package for survival analysis in R, vignettes. 2023 Mar 11;

144. Cook NR, Paynter NP. Performance of Reclassification Statistics in Comparing Risk Prediction Models. Biom J. 2011 Mar;53(2):237–58.

145. Pencina MJ, Steyerberg EW, D'Agostino RB. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med. 2011 Jan 15;30(1):11–21.

146. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagnostic and Prognostic Research. 2019 Oct 4;3(1):18.

147. Virtue S, Vidal-Puig A. It's Not How Fat You Are, It's What You Do with It That Counts. PLOS Biology. 2008 Sep 23;6(9):e237.

148. Tran TT, Yamamoto Y, Gesta S, Kahn CR. Beneficial Effects of Subcutaneous Fat Transplantation on Metabolism. Cell Metab. 2008 May;7(5):410–20.

149. Arner P, Rydén M. Human white adipose tissue: A highly dynamic metabolic organ. Journal of Internal Medicine. 2022;291(5):611–21.

150. Ibrahim MM. Subcutaneous and visceral adipose tissue: structural and functional differences. Obesity Reviews. 2010;11(1):11–8.

151. West-Eberhard MJ. Nutrition, the visceral immune system, and the evolutionary origins of pathogenic obesity. Proceedings of the National Academy of Sciences. 2019 Jan 15;116(3):723–31.

152. Spencer M, Unal R, Zhu B, Rasouli N, McGehee RE Jr, Peterson CA, et al. Adipose Tissue Extracellular Matrix and Vascular Abnormalities in Obesity and Insulin Resistance. The Journal of Clinical Endocrinology & Metabolism. 2011 Dec 1;96(12):E1990–8.

153. Karaca Ü, Schram MT, Houben AJHM, Muris DMJ, Stehouwer CDA. Microvascular dysfunction as a link between obesity, insulin resistance and hypertension. Diabetes Research and Clinical Practice. 2014 Mar 1;103(3):382–7.

154. Manikpurage HD, Eslami A, Perrot N, Li Z, Couture C, Mathieu P, et al. Polygenic Risk Score for Coronary Artery Disease Improves the Prediction of Early-Onset Myocardial Infarction and Mortality in Men. Circulation: Genomic and Precision Medicine. 2021 Dec;14(6):e003452.

155. Udler MS, McCarthy MI, Florez JC, Mahajan A. Genetic Risk Scores for Diabetes Diagnosis and Precision Medicine. Endocrine Reviews. 2019 Dec 1;40(6):1500–20.

156. Huang Y, Hui Q, Gwinn M, Hu YJ, Quyyumi AA, Vaccarino V, et al. Sexual Differences in Genetic Predisposition of Coronary Artery Disease. Circ Genom Precis Med. 2021 Feb;14(1):e003147.

157. Pirastu N, Cordioli M, Nandakumar P, Mignogna G, Abdellaoui A, Hollis B, et al. Genetic analyses identify widespread sex-differential participation bias. Nat Genet. 2021 May;53(5):663–71.

158. Klimentidis YC, Arora A, Newell M, Zhou J, Ordovas JM, Renquist BJ, et al. Phenotypic and Genetic Characterization of Lower LDL Cholesterol and Increased Type 2 Diabetes Risk in the UK Biobank. Diabetes. 2020 Jun 3;69(10):2194–205.

159. Tiwari S, Siddiqi SA. Intracellular Trafficking and Secretion of VLDL. Arteriosclerosis, Thrombosis, and Vascular Biology. 2012 May;32(5):1079–86.

160. Abbasi F, Lamendola C, Harris CS, Harris V, Tsai MS, Tripathi P, et al. Statins Are Associated With Increased Insulin Resistance and Secretion. Arteriosclerosis, Thrombosis, and Vascular Biology. 2021 Nov;41(11):2786–97.

161. Ooi EMM, Watts GF, Chan DC, Chen MM, Nestel PJ, Sviridov D, et al. Dose-Dependent Effect of Rosuvastatin on VLDL–Apolipoprotein C-III Kinetics in the Metabolic Syndrome. Diabetes Care. 2008 Aug;31(8):1656–61.

162. Dahl K, Brooks A, Almazedi F, Hoff ST, Boschini C, Bækdal TA. Oral semaglutide improves postprandial glucose and lipid metabolism, and delays gastric emptying, in subjects with type 2 diabetes. Diabetes, Obesity and Metabolism. 2021;23(7):1594–603.

163. Veerkamp MJ, de Graaf J, Bredie SJH, Hendriks JCM, Demacker PNM, Stalenhoef AFH. Diagnosis of Familial Combined Hyperlipidemia Based on Lipid Phenotype Expression in 32 Families. Arteriosclerosis, Thrombosis, and Vascular Biology. 2002 Feb;22(2):274–82.

164. Schooling CM, Kelvin EA, Jones HE. Alanine transaminase has opposite associations with death from diabetes and ischemic heart disease in NHANES III. Annals of Epidemiology. 2012 Nov 1;22(11):789–98.

165. Faulkner JL, Belin de Chantemèle EJ. Sex Differences in Mechanisms of Hypertension Associated With Obesity. Hypertension. 2018 Jan;71(1):15–21.

166. Khera A, Vega GL, Das SR, Ayers C, McGuire DK, Grundy SM, et al. Sex Differences in the Relationship between C-Reactive Protein and Body Fat. The Journal of Clinical Endocrinology & Metabolism. 2009 Sep 1;94(9):3251–8.

167. Kengne AP, Beulens JW, Peelen LM, Moons KG, van der Schouw YT, Schulze MB, et al. Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. The Lancet Diabetes & Endocrinology. 2014 Jan 1;2(1):19–29.

168. Chen GC, Arthur R, Iyengar NM, Kamensky V, Xue X, Wassertheil-Smoller S, et al. Association between regional body fat and cardiovascular disease risk among postmenopausal women with normal body mass index. European Heart Journal. 2019 Sep 7;40(34):2849–55.

169. Sacks FM, Tonkin AM, Craven T, Pfeffer MA, Shepherd J, Keech A, et al. Coronary Heart Disease in Patients With Low LDL-Cholesterol. Circulation. 2002 Mar 26;105(12):1424–8.

170. Hoogeveen RC, Gaubatz JW, Sun W, Dodge RC, Crosby JR, Jiang J, et al. Small Dense Low-Density Lipoprotein-Cholesterol Concentrations Predict Risk for Coronary Heart Disease. Arteriosclerosis, Thrombosis, and Vascular Biology. 2014 May;34(5):1069–77.

171. Liou L, Kaptoge S. Association of small, dense LDL-cholesterol concentration and lipoprotein particle characteristics with coronary heart disease: A systematic review and meta-analysis. PLOS ONE. 2020 Nov 9;15(11):e0241993.

172. Sulc J, Sjaarda J, Kutalik Z. Polynomial Mendelian randomization reveals non-linear causal effects for obesity-related traits. Human Genetics and Genomics Advances. 2022 Jul 14;3(3):100124.

173. Wainberg M, Mahajan A, Kundaje A, McCarthy MI, Ingelsson E, Sinnott-Armstrong N, et al. Homogeneity in the association of body mass index with type 2 diabetes across the UK Biobank: A Mendelian randomization study. PLoS Med. 2019 Dec 10;16(12):e1002982.

174. Wing RR, Lang W, Wadden TA, Safford M, Knowler WC, Bertoni AG, et al. Benefits of Modest Weight Loss in Improving Cardiovascular Risk Factors in Overweight and Obese Individuals With Type 2 Diabetes. Diabetes Care. 2011 Jun 17;34(7):1481–6.

175. Sun YQ, Burgess S, Staley JR, Wood AM, Bell S, Kaptoge SK, et al. Body mass index and all cause mortality in HUNT and UK Biobank studies: linear and non-linear mendelian randomisation analyses. BMJ. 2019 Mar 26;364:l1042.

176. Taylor R, Holman RR. Normal weight individuals who develop Type 2 diabetes: the personal fat threshold. Clinical Science. 2014 Dec 9;128(7):405–10.

177. Lontchi-Yimagou E, Dasgupta R, Anoop S, Kehlenbrink S, Koppaka S, Goyal A, et al. An Atypical Form of Diabetes Among Individuals With Low BMI. Diabetes Care. 2022 Jun;45(6):1428–37.

178. Wade KH, Hamilton FW, Carslake D, Sattar N, Davey Smith G, Timpson NJ. Challenges in undertaking nonlinear Mendelian randomization. Obesity. 2023;31(12):2887–90.

Popular science summary

Cardiometabolic diseases like obesity and type 2 diabetes represent significant public health concerns worldwide, contributing substantially to mortality rates, primarily due to cardiovascular disease. However, the manifestation of these conditions varies widely among individuals, with some experiencing severe complications while others remain relatively unaffected. To understand this variability better, in this thesis I used genetic and biomarker data from large studies, exploring the factors associated with a disproportionately higher or lower risk of complications commonly associated with obesity and type 2 diabetes.

In Paper I, I examined what are the main differences between two genetically determined obesity profiles, one of which is associated with a 'concordant' higher risk of type 2 diabetes, while the other profile, which we called 'discordant', is associated with protection against type 2 diabetes. We found that key distinctions between these two profiles are how adipose tissue is distributed in the body, the vascular function, reflected by blood pressure, and the microenvironment in which adipose tissue cells live. We identify 17 genes that, when more or less abundant, have a discordant effect, being protective against type 2 diabetes despite being associated with obesity. These genes could potentially be used, for example, as medication targets to improve sugar levels in individuals with obesity.

In Paper II, I extended this analysis to focus on genetically determined type 2 diabetes profiles, which are either concordant or discordant in their association with cardiovascular disease. By comparing these profiles, I uncovered insights into how certain metabolic pathways, particularly those involving very low-density lipoprotein metabolism, play a pivotal role in determining cardiovascular risk in individuals with type 2 diabetes. This research identified eight genetic loci associated with cardioprotective effects in type 2 diabetes, some of which contain target genes of current medications like statins and GLP-1 receptor agonists. Additionally, we found that adding polygenic scores representing these profiles to common cardiovascular risk markers can improve cardiovascular disease prediction, especially in men, updating the probabilities of 5% of cases and 3% of non-cases in the right direction.

In Paper III, I introduced an approach composed of various machine learning algorithms to identify distinct subgroups within the population with unexpected variations in common biomarkers of cardiovascular risk relative to their body mass index, that is, being either concordant or discordant for their body size. This method revealed that approximately 20% of individuals show significant deviations from expected biomarker levels based on their body mass index. We found five distinct patterns of discordance between biomarkers and body mass index, with significant differences in men and women. These discordant profiles differ from the concordant profile in their association with the prevalence and incidence of cardiovascular disease and can also enhance cardiovascular risk prediction. We found that adding discordant profile information to common cardiovascular risk markers led to improvements in prediction that are comparable to adding low-density lipoprotein levels, a known important cardiovascular risk factor.

Lastly, in Paper IV, I contributed to an investigation into the causal effects of BMI on various health outcomes, including type 2 diabetes and cardiovascular disease, using a method called Mendelian randomisation, which leverages the inherent randomness of genetics as a natural random experiment of nature, similar to the random assignment into different treatment groups in a drug trial. We found consistent positive effects of BMI on type 2 diabetes risk across sexes while observing sex-differential effects on coronary artery disease. Additionally, we found evidence of nonlinear effects of BMI particularly in lipid blood sugar levels.

In conclusion, through comprehensive genetic and phenotypic analyses of divergent manifestations of cardiometabolic diseases like obesity and type 2 diabetes, we underscore key mechanisms for why some individuals are more susceptible to complications than others. We show potential clinical applications of these analyses, by informing potential molecular targets for intervention and through improvements in cardiovascular risk prediction. These findings offer insights for the development of more targeted interventions and personalized treatment strategies in cardiometabolic diseases.

Populärvetenskaplig sammanfattning

Kardiometabola sjukdomar som fetma och typ 2-diabetes utgör betydande folkhälsoproblem över hela världen och bidrar väsentligt till dödligheten, främst till följd av hjärt-kärlsjukdomar. Dock varierar manifestationen av dessa tillstånd kraftigt bland individer, där vissa upplever allvarliga komplikationer medan andra förblir relativt opåverkade. För att bättre förstå denna variation använde jag genetiska och biomarkörsdata från stora studier för att utforska faktorer som är associerade med en oproportionellt högre eller lägre risk för komplikationer som vanligtvis är förknippade med fetma och typ 2-diabetes.

I första studien undersökte jag de främsta skillnaderna mellan två genetiskt bestämda fetmaprofiler, varav den ena är associerad med en 'konkordant' högre risk för typ 2-diabetes, medan den andra profilen, som vi kallade 'diskordant', är förknippad med skydd mot typ 2-diabetes. Vi fann att nyckelskillnaderna mellan dessa två profiler är hur fettvävnaden fördelas i kroppen, den vaskulära funktionen, som återspeglas av blodtrycket, och den närmaste miljön (s k mikromiljön) där fettvävnadsceller lever. Vi identifierade 17 gener som, när de är mer eller mindre uttryckta i celler, har en skyddande effekt mot typ 2-diabetes trots att de är förknippade med fetma. Dessa gener kan potentiellt användas för att utveckla nya behandlingar för att förbättra sockernivåerna hos personer med fetma.

I artikel II utvidgade jag denna analys för att fokusera på genetiskt bestämda typ 2-diabetesprofiler, som antingen är konkordanta eller diskordanta i sin association med hjärt-kärlsjukdom. Genom att jämföra dessa profiler avslöjade jag insikter om hur vissa metabola vägar, särskilt de som involverar metabolismen av mycket lågdensitetslipoprotein, spelar en avgörande roll för att bestämma hjärt-kärlrisken hos personer med typ 2-diabetes. Vi identifierade åtta genetiska markörer kopplade till skydd mot kardiovaskulär sjukdom vid typ 2-diabetes. Några av dessa är mål för nuvarande läkemedel så som statiner och GLP-1-receptorblockare. Vi kunde även visa att om man adderar genetisk risk poäng bestående av många genvarianter (s k polygen risk score) som representerar dessa profiler till vanliga markörer för hjärt-kärlrisk kan förbättra prognosen för hjärt-kärlsjukdom, särskilt hos män.

I artikel III introducerade jag en metod bestående av olika maskininlärningsalgoritmer för att identifiera specifika subgrupper inom populationen med oväntade variationer i vanliga biomarkörer för hjärt-kärlrisk i förhållande till deras BMI (ett mått på fetma), det vill säga att vara antingen konkordanta eller diskordanta för sin kroppsstorlek. Denna metod avslöjade att cirka 20% av individerna visar betydande avvikelser från förväntade biomarkörsnivåer baserat på deras BMI. Vi fann fem distinkta mönster av diskordans mellan biomarkörer och BMI, med betydande skillnader mellan män och kvinnor. Dessa diskordanta profiler skiljer sig från den konkordanta profilen i deras association med förekomsten och insjukningsfrekvens av hjärt-kärlsjukdom och kan också förbättra prognosen för hjärt-kärlrisk. Vi fann att genom att lägga till information om diskordanta profiler till vanliga markörer för hjärt-kärlrisk ledde till förbättringar i prognos som är jämförbara med att lägga till lågdensitetslipoproteinnivåer, en känd viktig riskfaktor för hjärt-kärlsjukdom.

Slutligen bidrog jag i artikel IV till en undersökning av de kausala effekterna av BMI på olika hälsorelaterade resultat, inklusive typ 2-diabetes och hjärt-kärlsjukdom, med hjälp av en metod som kallas Mendelsk randomisering. Metoden utnyttjar det faktum att gener fördelas slumpmässigt vid befruktning som ett naturligt experiment, liknande den slumpmässiga tilldelningen till olika behandlingsgrupper i en läkemedelsstudie. Vi fann positiva kopplingar mellan BMI och risken för typ 2-diabetes i både män och kvinnor medan vi observerade könsspecifika effekter på hjärtkärlssjukdom. Dessutom fann vi bevis för icke-linjära effekter av BMI särskilt på lipid- och blodsockernivåer.

Sammanfattningsvis, genom omfattande genetiska och fenotypiska analyser av olika manifestationer av kardiometabola sjukdomar så som fetma och typ 2-diabetes understryker vi viktiga mekanismer för varför vissa individer är mer mottagliga för komplikationer än andra. Vi visar potentiella kliniska tillämpningar av dessa analyser genom att visa möjliga molekylära mål för intervention och genom förbättringar av prognoser för hjärt-kärlrisk. Dessa resultat kan leda till utvecklingen av mer riktade interventioner och personliga behandlingsstrategier.

Acknowledgements

I am profoundly grateful for the support and guidance provided by my exceptional team of supervisors throughout these past four years, a period of considerable challenges that included a global pandemic. I really think they were the perfect combination; their complementary knowledge and wisdom helped me navigate the complexities of this bumpy journey and enriched my learning experience. Paul, thank you for opening the doors of the GAME unit and making me feel at home during a really critical moment in my life. For teaching me the principles of transparent and rigorous research, while fostering an environment where the personal dimension is not cast aside, but rather takes the importance that it deserves. Ewan, I really admire your enthusiasm and how you seamlessly integrate novel and quite complex methodological approaches with your clinical practice (how is that even done!?). Your expertise doing this has been fundamental to enhance the clinical relevance of my analyses. Juan, thank you for diving together with me into endless lines of code, patiently explaining to me in simple terms what algorithms are doing. This has been essential for my coding skills and to understand how to tackle analytical challenges more effectively. The time me and my wife spent at your place with your family in beautiful Galicia was among the best moments during these past years. Nick, I think your transition into leadership of the GAME unit has been exceptional, because it ensured a sense of security and confidence within the team, allowing us to focus on our work without worrying for anything else, which was truly remarkable.

I was also very lucky to receive guidance from Maria Gomez, whose thoughtful input not only greatly enriched the projects included in this thesis but also provided incredible support to a really interesting parallel collaboration in the ANDiS dataset, which I hope to continue developing. This project started thanks to the confidence that Valeriya Lyssenko placed on my analytical skills, for which I am very grateful. This collaboration led me to be part of BEAT-DKD consortium, a space where I gained invaluable knowledge and experiences from many other researchers.

The GAME unit has been a cornerstone for growing as a researcher and as a person. I want to thank Pascal Mutie for making me part of his paper which was included in this thesis. I really appreciate our shared late-night office sessions, during which we steered through the challenges of our respective projects together. Sebastian, thanks for your invaluable insights and explanations of the technical

aspects of genotyping and other omics technologies, the mechanism of selection and genetic drift, and the potential connections of the variants we identify to physiological mechanisms. It is talking to you that effect sizes and p-values actually start making biological sense. Hugo Fitipaldi, I consider myself very lucky to have had the opportunity to follow in your footsteps on this journey. I am grateful for your generosity not only at work, but also your insights into broader aspects of life. Hugo Pomares, it was your initiative that set up the crucial first meeting with Paul, leading to my involvement in the GAME unit. Having a friend who shared the same challenges and spoke my mother tongue was very comforting. Your kindness and willingness to assist anyone in the team added a layer of warmth to our work environment, making it truly enjoyable. Naeimeh, I am delighted to continue our collaboration on future projects. Your infectious happiness and positive energy have a remarkable impact on the entire team, making our work even more enjoyable. Your talent and expertise are truly inspiring, and I am eager to continue learning from you as we work together. Neli, as I embarked on my journey in data science, your help was invaluable to me. Everyone in the team benefited from your talent, which really showcases your selflessness, dedication, and versatility in swiftly transitioning between projects. Huang Mi is one of the most dedicated and modest scientists I've had the privilege to meet. I admire your ability to grasp cutting-edge genomic editing technologies while remaining approachable and patient in explaining fundamental techniques to us. Alaitz, your help was crucial in my initial steps into genetic analysis, being an essential contributor to my first paper. I learned a lot from your expertise and rigour in exploring alternative explanations and interpretations beyond surface-level solutions. Tibor, thanks for your advice on aspects of work that I found were often overlooked, which showed your true concern for the well-being of others in the team. Marketa, thank you for always providing helpful answers to my questions and requests, whether they were about science or university operations, which definitely made my life easier. Lastly, but certainly not least, Pernilla, whose impressive talent in navigating the complex administrative tasks of the university made everything ran smoothly behind the scenes, which did not go unnoticed, and for which I am truly thankful.

I would also like to extend my heartfelt gratitude to my friends and colleagues that I have met at LUDC, both current and former, whose dedication and achievements serve as a constant source of motivation, and whom with I have shared not only my work but also my joys and frustrations. I am truly grateful for their support and camaraderie, with special appreciation to Alice Giontella and Esther González-Padilla. Thanks also to all the members of the SOPHIA consortium, where I gained invaluable experience in conducting collaborative research, and I learned the importance of patient inclusion. Special thanks to Femke and Ali, whose contributions were instrumental in the completion of my third paper.

No journey can ever be completed without friends guiding your soul, even without them realising their contribution. Jag vill tacka Malin Claesson, som har varit en ovärderlig och villkorslös vän redan innan jag började min doktorsutbildning. Du har verkligen fått mig att känna att Sverige är mitt hem, och detta har varit mycket viktigt för mig för att kunna avsluta den här fasen av min karriär. Jag vill utöka detta tack till din familj och dina vänner som nu också är våra vänner, inklusive Julia, Cajsa och Ebba. I also would like to express my deepest gratitude to my friends: Chaitanya, Kanika, Tomas, Philipa, Angel, Christine, Yassine, Santiago, Melba, and Karl, who always had open doors and open arms for me and my wife, always offering invaluable support, generosity, fun and endless inspiration. To fully capture everything I have received from each of them I would have to write another thesis. Quiero hacer una mención especial a nuestros grandes amigos Sergio y Juliana que han compartido con nosotros la alegría y la calidez de su hogar, esenciales para terminar este proyecto, y que a pesar de haberse mudado lejos, nos siguen haciendo la vida agradable. También gracias en Alemania a Juan Sebastián Muñoz. A pesar de mi fracaso para mantener el contacto de manera constante, has estado presente, compartiendo el desafío a veces abrumador de adaptarnos a un entorno nuevo y extraño.

Como dice el gran Piero, yo vengo del tinto amargo y de la piel que tengo de ustedes, mi queridísima familia. Le debo todo a la tenacidad y sabiduría de mi padre Oswaldo, de quien he aprendido que la adversidad e incluso la muerte se enfrentan con el más profundo amor. A mi madre Carmen que no nos ha dejado desfallecer nunca con su energía, alegría e incondicional apoyo a nuestros proyectos. Gracias a mis hermanos: a Luis, por decidir venir a pasar estos dos años conmigo y con Paula, trasnochando conmigo trabajando, dándome consejos y escuchando los míos; y a José, no creo decirte suficiente que cada vez que tengo la oportunidad de verte y hablar contigo, siento que mis energías vuelven a recargarse y puedo continuar. Quiero además extender mi agradecimiento a mi abuelita Luz María y en general, a toda mi familia extendida en Colombia, la cual tengo la fortuna de que sea inmensa y amorosa, y que me han permitido con su apoyo estar aquí.

Paula, he reservado estas últimas líneas para dedicártelas a ti, pero seré breve porque jamás podría hacer justicia a todo lo que he recibido de tu amor con palabras, y más bien prefiero demostrártelo cada día. Aquí solo diré, desde lo más profundo de mi alma, que no creo haber tomado una mejor decisión que compartir mi vida contigo, y que estos años sólo me lo han confirmado. Te amo.