Elucidating causal relationships between energy homeostasis and cardiometabolic outcomes

Mutie, Pascal

2022

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

Link to publication

Citation for published version (APA):
Mutie, P. (2022). *Elucidating causal relationships between energy homeostasis and cardiometabolic outcomes*. Lund University, Faculty of Medicine.

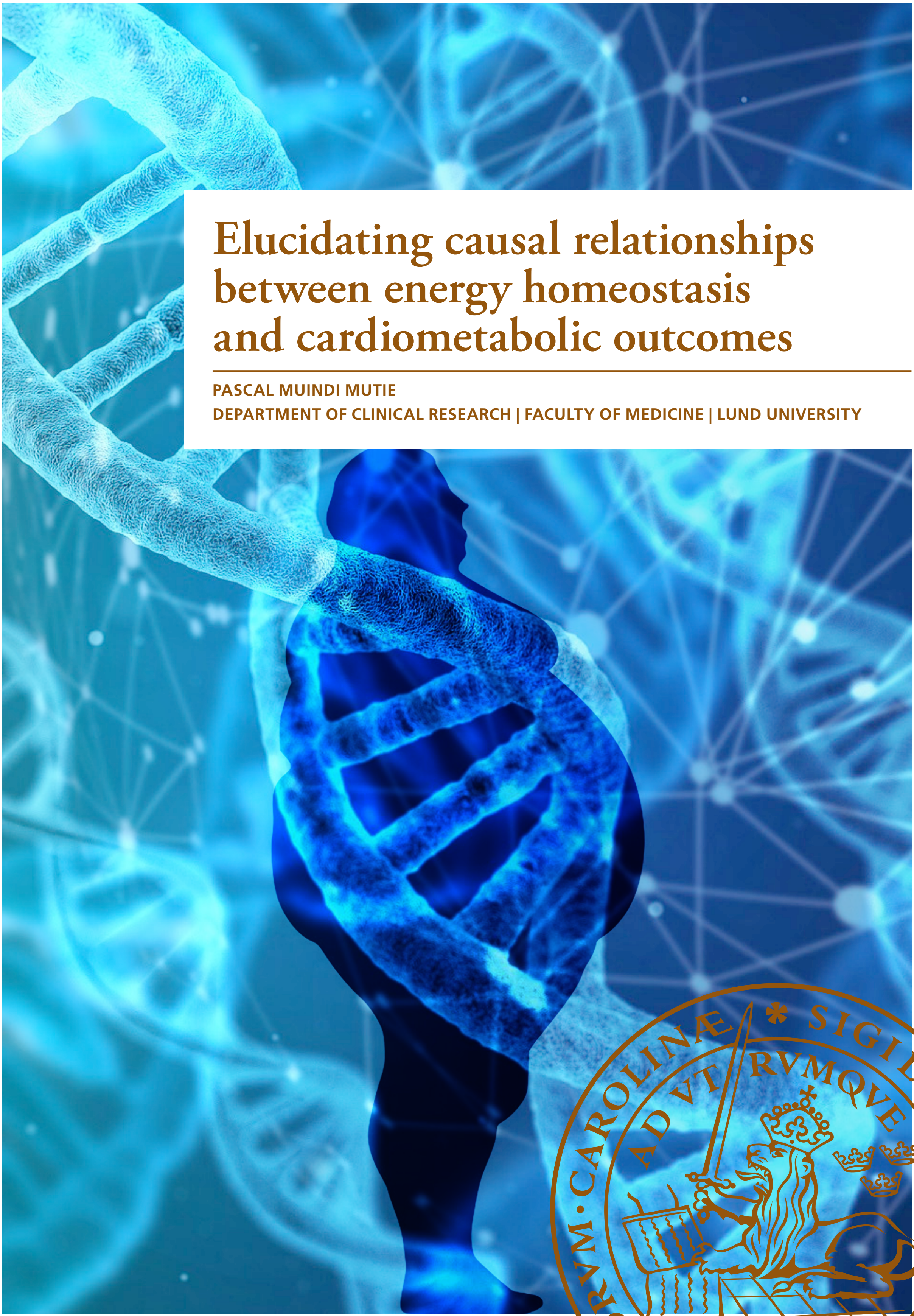
Total number of authors:
1

LUND UNIVERSITY

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

Elucidating causal relationships between energy homeostasis and cardiometabolic outcomes

PASCAL MUINDI MUTIE
DEPARTMENT OF CLINICAL RESEARCH | FACULTY OF MEDICINE | LUND UNIVERSITY



Elucidating causal relationships between energy homeostasis and cardiometabolic outcomes

Elucidating causal relationships between energy homeostasis and cardiometabolic outcomes

Pascal Muindi Mutie

LUND
UNIVERSITY

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 the 16th of June, 2022 at 13.00 in CRC Aula Hall, Department of Clinical Sciences,
Jan Waldenströms gata 35, Malmö
Thesis advisors: Giuseppe N. Giordano, Paul W. Franks
Faculty opponent: Dr Jess Tyrell – University of Exeter, UK

Organization LUND UNIVERSITY	Document name: DOCTORAL DISSERTATION
	Date of disputation: 2022-06-16
Author: Pascal Muindi Mutie	Sponsoring organization: RHAPSODY, IMI-DIRECT, NASCENT, SOPHIA

Elucidating causal relationships between energy homeostasis and cardiometabolic outcomes

Abstract

Energy metabolism dyshomeostasis is associated with multiple health problems. For example, abundant epidemiological data show that obesity and overweight increase the risk of cardiometabolic diseases and early mortality. Type 2 diabetes (T2D), characterized by chronically elevated blood glucose, is also associated with debilitating complications, high healthcare costs and mortality, with cardiovascular complications accounting for more than half of T2D-related deaths. Prediabetes, which is defined as elevated blood glucose below the diagnostic threshold for T2D, affects approximately 350M people worldwide, with about 35-50% developing T2D within 5 years. Further, non-alcoholic fatty liver disease, a form of ectopic fat deposition as a result of energy imbalance, is associated with increased risk of T2D, CVD and hepatocellular carcinoma.

Determination of causal relationships between phenotypes related to positive energy balance and disease outcomes, as well as elucidation of the nature of these relationships, may help inform public health intervention policies. In addition, utilizing big data and machine learning (ML) approaches can improve prediction of outcomes related to excess adiposity both for research purposes and eventual validation and clinical translation.

Aims

In paper 1, I set out to summarize observational evidence and further determine the causal relationships between prediabetes and common vascular complications associated with T2D i.e., coronary artery disease (CAD), stroke and renal disease. In paper 2, I studied the association between *LRIG1* genetic variants and BMI, T2D and lipid biomarkers. In paper 3, we used ML to identify novel molecular features associated with non-alcoholic fatty liver disease (NAFLD). In paper 4, I elucidate the nature of causal relationships between BMI and cardiometabolic traits and investigate sex differences within the causal framework.

Results

Prediabetes was associated with CAD and stroke but not renal disease in observational analyses, whilst in the causal inference analyses, prediabetes was only associated with CAD. Common *LRIG1* variant (rs4856886) was associated with increased BMI and lipid hyperplasia but a decreased risk of T2D. In paper 3, models using common clinical variables showed strong NAFLD prediction ability (ROCAUC = 0.73, p < 0.001); addition of hepatic and glycemic biomarkers and omics data to these models strengthened predictive power (ROCAUC = 0.84, p < 0.001). Finally, there was evidence of non-linearity in the causal effect of BMI on T2D and CAD, biomarkers and blood pressure. The causal effects BMI on CAD were different in men and women, though this difference did no hold after Bonferroni correction.

Conclusion

We show that derangements in energy homeostasis are causally associated with increased risk of cardiometabolic outcomes and that early intervention on perturbed glucose control and excess adiposity may help prevent these adverse health outcomes. In addition, effects of novel *LRIG1* genetic variants on BMI and T2D might enrich our understanding of lipid metabolism and T2D and thus warrant further investigations. Finally, application of ML to multidimensional data improves prediction of NAFLD; similar approaches could be used in other disease research.

Key words: Adiposity, causal inference, cardiometabolic disease, Mendelian randomization

Classification system and/or index terms (if any)

Supplementary bibliographical information		Language
ISSN and key title : 1652-8220 Lund University, Faculty of Medicine Doctoral Dissertation Series 2022: 92		ISBN 978-91-8021-253-3
Recipient's notes	Number of pages 89	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 2022-05-13

Elucidating causal relationships between energy homeostasis and cardiometabolic outcomes

Pascal Muindi Mutie

LUND UNIVERSITY

Cover design: Pascal M. Mutie. Images courtesy of www.pixabay.com

Faculty of Medicine
Department of Clinical Research

ISBN 978-91-8021-253-3
ISSN 1652-8220

Lund University, Faculty of Medicine Doctoral Dissertation Series 2022: 92

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

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

Dedication: To my parents and family

Table of Contents

List of publications

Mutie, P.M., Pomares-Millan, H., Atabaki-Pasdar, N. et al. **An investigation of causal relationships between prediabetes and vascular complications**. Nat Commun 11, 4592 (2020). https://doi.org/10.1038/s41467-020-18386-9

Herdenberg, C., Mutie, P.M., Billing, O. et al. **LRIG proteins regulate lipid metabolism via BMP signaling and affect the risk of type 2 diabetes**. Commun Biol 4, 90 (2021). https://doi.org/10.1038/s42003-020-01613-w

Atabaki-Pasdar N, Ohlsson M, Viñuela A, Frau F, Pomares-Millan H, Haid M, Jones AG, Thomas EL, Koivula RW, Kurbasic A, Mutie PM, Fitipaldi H, Fernandez J, Dawed AY, Giordano GN, Forgie IM, et al. **Predicting and elucidating the etiology of fatty liver disease: A machine learning modeling and validation study in the IMI DIRECT cohorts.** PLoS Med. 2020 Jun 19;17(6):e1003149. doi: 10.1371/journal.pmed.1003149. PMID: 32559194; PMCID: PMC7304567.

Pascal M. Mutie[1], Hugo Pomares-Milan[2], Naeimeh Atabaki-Pasdar[2], Daniel Coral[2], Hugo Fitipaldi[2], Neli Tsereteli[2], Juan Fernandez Tajes[2], Paul W. Franks[2], Giuseppe N. Giordano[2]. **Investigating the causal relationships between excess adiposity and cardiometabolic health in men and women**.

Submitted paper

Publications not included in this thesis

Bergwall S, Acosta S, Ramne S, Mutie P, Sonestedt E. **Leisure-time physical activities and the risk of cardiovascular mortality in the Malmö diet and Cancer study**. BMC Public Health. 2021;21(1):1948. Published 2021 Oct 26. doi:10.1186/s12889-021-11972-6

Mutie, P.M., Drake, I., Ericson, U. et al. **Different domains of self-reported physical activity and risk of type 2 diabetes in a population-based Swedish cohort: the Malmö diet and Cancer study**. BMC Public Health 20, 261 (2020). https://doi.org/10.1186/s12889-020-8344-2

Mutie, P.M., Giordano, G.N. & Franks, P.W. **Lifestyle precision medicine: the next generation in type 2 diabetes prevention?** BMC Med 15, 171 (2017). https://doi.org/10.1186/s12916-017-0938-x

Acknowledgements

First is to sincerely thank my supervisors, **Paul W. Franks** and **Giuseppe N. Giordano (Nick). Paul**, you have been a very caring, kind and forward-looking supervisor who was genuinely concerned with my development as a researcher and everyone else in the group, all the time being humble and modest about it. Your teams have always been amazing and I am glad to be part of the GAME unit. **Nick,** you have been helpful and caring as well, not forgetting how you made sure the transition period was smooth. Thank you so much.

I am indebted to the folks I found in the group when I joined, who have since moved on with their careers elsewhere. **Robert Koivula**, I bumped in to him when he was presenting about his physical activity project at CRC and we had a heated debate on objective measurements of PA, little did I know I would be joining his group. He gave me tips and encouraged me to remain curious. Thanks Robo! **Tibor Varga, Angela Estampador** and **Alaitz Poveda**, the folks who gently introduced me to R (and other tools), and pointed me to unlimited resources. Thanks.

The amazing and crazy colleagues I share the office with, words can never do justice to how grateful I am to have met you bunch! Even with COVID-19 upending the society, we still found time and space to be cordial to each other! So, **Naeimeh Atabaki-Pasdar** – for being cheerful, encouraging and always seeing the fun side of life and still being an excellent scientist, *mamnoun hastam*! **Hugo Pomares-Milan** (Pommes) – I collaborated with you on the first project; for being the cool guy with the darkest jokes and helpful all round, *gracias amigo!* **Hugo Fitipaldi** (Vivaldi) – ever resourceful and caring yet you also serve humour unprovoked, *obrigado!* **Daniel Coral** – nice chap, super resourceful and a great human being, not one to miss out in the humour hall of fame! It's great knowing and working with you, *muchas gracias!* And of course, **Sebastian Kalamajski** – brilliant scientist (we had amazing discussions), awesome fellow and the only wet lab person I know who can tell a dry joke! Thanks Seb! **Mi Huang** – calm and friendly and also very helpful. Amazing how you took all the banter from us with such calmness. Thanks Mi.

I had the opportunity to work closely with **Juan Fernandez** and **Neli Tsereteli** who came to the team to offer technical support. **Juan,** your help was invaluable and it was a pleasure working with you, I learned a lot. *Gracias Juan!* **Neli**, you were amazing with all the data support and the extra hours you put just to make sure that things worked. Thank you so much.

There are always unsung heroes behind the scenes who make things work smoothly. **Pernilla Siming**, many thanks for making navigation through many administrative issues so easy. To the IT duo, **Mattias Borell** and **Johan Hultman** -thank you for making work in the clusters manageable, for helping with troubleshooting the many times things weren't working out and not 'killing' our jobs when we nearly hogged up all the RAM!

And to the many people I have closely interacted with in my research journey, you are not forgotten, thank you very much.

Finally, to my family. **Linet**, my partner and love of my life, this journey would not have been possible without your constant support. Thank you so much. **Nathan** and **Liam**, you brought so much joy to our lives and made every day worth the effort. To my siblings, **Peter, Maggie,** and **Trieza** – thanks for walking this journey with me, for the encouragement and the humour. To my Mum – thanks for understanding and your unwavering support. To my Dad (RIP) – Look papa, we became scientists! I know we'd be *chillin'* and laughing so hard over drinks and barbeque, so I forever hold you in my heart. I know you smiling down on me saying everything is OK.

Popular summary

Obesity is correlated with many diseases including cardiovascular disease, type 2 diabetes (T2D), kidney diseases, and some cancers. In addition, obesity can reduce quality of life, is often stigmatized, lowers self-esteem and adversely affects many other aspects of mental health. Obesity and T2D are two of the greatest public health challenges of recent times; prediabetes, which is elevated blood sugar levels below the diagnosis threshold for diabetes, is another major concern, which is often unrecognized, as moderately elevated blood sugar concentrations yield few, if any, perceptible symptoms. This is also true during the early stages of diabetes, often leaving the disease undiagnosed until it has worsened, and symptoms (frequent urination, excessive thirst, fatigue etc) are apparent.

Correlation does not always mean causation, and to effectively prevent the complications of T2D or obesity it is important to establish whether excess body fat or elevated blood sugar actually "cause" the development of other diseases. To be confident that there is a causal relationship between obesity, prediabetes, and the diseases with which they correlate, we must be sure that these relationships are not influenced by other factors, here called 'confounders'. Further, it is important to establish the nature of these relations in the sense that, does risk increase gradually and proportionately according to the level of obesity or prediabetes, or do the outcomes change in some other 'non-linear' way?

Accumulation of fat in the liver, non-alcoholic fatty liver disease (NAFLD), is one of the consequences of obesity and can progress to extensive liver damage and liver cancer. Early detection of NAFLD is not easy and confirmation of diagnosis requires invasive biopsy, which is associated with complications like infections and bleeding. Even when imaging methods like ultrasound are used, they do not tell us who is likely to progress to the severe form of liver damage. By using information that is easily accessible in the clinical environment, it is possible to construct statistical methods that can predict NAFLD and identify those at risk early. These methods can be further improved by additional 'omics' data which refers to novel assays of genetic and metabolites found in the blood. Such prediction models show a strong prediction ability and with further model refinement and as such data becomes more available, this offers possibilities of more accurate early detection of NAFLD and other diseases.

There has been immense growth in the field of genetics with subsequent identification of genetic variants associated with many diseases, through genome-wide association studies (GWAS). While these are important developments, the identified genetic variants are important to the biomedical field if their exact roles can be determined. To date, a modest number of the millions of genetic variants identified have known functions, and therefore research is still ongoing to determine their role. In this project, a polymorphism (a piece of DNA code that differs between people within the same population) in the *LRIG1* gene was shown to influence obesity and T2D and therefore will be of great interest for further investigation.

Overall, in this project, we found that prediabetes is a likely cause of heart disease irrespective of T2D diagnosis, but we did not find sufficient evidence to support prediabetes as a likely cause of stroke or kidney disease. This implies that early detection and correction of impaired blood glucose control is crucial in preventing heart disease, emphasizing the importance of maintaining a healthy lifestyle to keep blood sugar under control. We also showed that obesity is a probable cause of T2D, heart disease and hypertension, as well as variation in levels of blood glucose, blood lipids and blood pressure. These associations are largely non-linear. Of interest, we showed that the risk to cause heart disease conferred by obesity differs between men and women and that menopause status and age in women has substantial influence on the effect obesity has on a woman's health. Again, this underscores the fact that maintaining a healthy weight is highly beneficial in preventing non-communicable diseases and the often-associated premature death.

Using data that can be availed in the clinic, we created models for predicting NAFLD which showed good performance even when tested on another cohort. Further, addition of omics data, availed by more sophisticated assay methods, improved the prediction ability of the models. These methods can be used to identify people at risk who can benefit from timely interventions. In addition, as omics data become more accessible, it will lead to improved prediction tools, not only for NAFLD, but also other diseases. Lastly, we showed that a variant of the *LRIG1* gene was associated with increased BMI but reduced risk of T2D, and was also associated with the size of fat cells (adipocytes). This variant seems to determine condition of "healthy obesity", but further research is needed to fully understand its biological functions.

In conclusion, prediabetes causes heart disease, independent of T2D. Obesity also causes T2D, heart disease and hypertension, as well as variation in blood levels of glucose and lipids, and blood pressure. These effects of obesity are largely non-linear and it is possible to estimate the impact of weight reduction on causal risk of a disease. Further, effects of obesity on causation of heart disease differ in men and women, and in women differ between younger and older women. While general recommendations encourage maintaining a healthy weight, further research is needed to understand interventions that are best suited for men and women, as well as understanding the roles of newly identified genetic variants.

Abbreviations

AgRP	Agouti-related peptide
ALT	Alanine aminotransferase
ANS	Autonomic nervous system
AST	Aspartate aminotransferase
AUC	Area under the curve
BAT	Brown adipose tissue
BMI	Body mass index
BMP	Bone morphogenic protein
BP	Blood pressure
CAD	Coronary artery disease
CART	Cocaine- and amphetamine-regulated transcript
CCK	Cholecystokinin
CKD	Chronic kidney disease
CNS	Central nervous system
CVD	Cardiovascular disease
DNA	Deoxyribonucleic acid
DPP	Diabetes Prevention Program
DPS	Diabetes Prevention Study
ECG	Electrocardiogram
EGFR	Estimated glomerular filtration rate
FFA	Free fatty acid
fsOGTT	Frequent sample oral glucose tolerance test
FTO	Fat mass and obesity-associated
GAD	Glutamic acid decarboxylase
GCK	Glucokinase
GIP	Glucose-dependent insulinotropic peptide
GLP-1	Glucagon-like peptide 1
GWAS	Genome-wide association study
HDL	High density lipoprotein
HRC	Haplotype Reference Consortium

IFG	Impaired fasting glucose
IGR	Impaired glucose regulation
IGT	Impaired glucose tolerance
IV	Instrumental variable
LASSO	Least absolute shrinkage and selection operator
LDL	Low density lipoprotein
LPA	Lipoprotein(a)
MC4R	Melanocortin receptor
MEF	Mouse embryonic fibroblast
ML	Machine learning
MMTT	Mixed meal tolerance test
MR	Mendelian randomization
MRI	Magnetic resonance imaging
mTOR	Mechanistic target of rapamycin
NCD	Non-communicable diseases
NLMR	Non-linear Mendelian randomization
NPY	Neuropeptide Y
OLS	Ordinary least squares
OR	Odds Ratio
PAF	Population attributable fraction
PCSK1	Proprotein Convertase Subtilisin/Kexin Type 1
POMC	Pro-opiomelanocortin
PRS	Polygenic Risk Score
PYY	Peptide YY
RCT	Randomised controlled trial
RMR	Resting metabolic rate
ROC	Receiver-operator characteristic curve
SAT	Subcutaneous adipose tissue
SNP	Single nucleotide polymorphism
T2D	Type 2 diabetes
TEF	Thermic effect of food
TG	Triglyceride
TSLS	Two stage least squares
WES	Whole-exome sequencing
WHO	World Health Organization
WHR	Waist-hip ratio

Chapter 1

Introduction

Living organisms sustain their core biological functions by intake, expenditure and storage of energy[1]. In humans, energy intake involves ingestion of carbohydrates, fat, proteins and alcohol, while expenditure includes resting metabolic rate (RMR), thermic effect of food (TEF), and physical activity (PA)[2]. If energy intake equals expenditure, fluctuation in body weight is negligible. In positive energy balance, energy intake exceeds expenditure and leads to excess calories being stored as fat; when sustained for long enough this results in overweight (BMI 25 – 29.9 kg/m^2) or obesity (BMI ≥ 30kg/m^2). When energy expenditure exceeds intake, the result is a state of negative energy balance and loss of body weight[3]. Disturbances in this elaborate system result in diseases mainly associated with energy metabolism and its complications.

The aim of aetiological studies is to establish causality i.e., cause-effect inference. Observational studies have major drawbacks namely confounding, bias and reverse causality that render causal inference unreliable. Randomized controlled trials (RCTs), considered the gold-standard for testing causality, are expensive, some take time, or are unfeasible in some situations making epidemiological studies the preferred design. Further, co-occurrence of exposures make it impossible to entirely attribute outcomes to an intervention. For instance, it is uncertain whether the benefits of weight loss are specifically due to reduction in adipose tissue or alterations in other risk factors that coincide with weight loss or other risk exposures that are attenuated by the intervention[4]. Different study situations, therefore, call for different designs. To answer the question of causality in epidemiological contexts, Mendelian randomization (MR) offers a rather powerful solution.

Background

Obesity and T2D are complex and inter-related conditions. The progressive increase in the prevalence of obesity over time has hugely contributed to the rise of T2D[5]. According to the World Health Organization (WHO), obesity has nearly tripled since 1975. In 2016, more than 1.9 billion adults (>18yrs) were overweight, out of which 650 million were obese. In the same year, 39% of adults were overweight and

13% were obese[6]. Between 1975 and 2014, the global age-standardized BMI increased from 21.7 kg/m^2 (CI: 21.3, 22.1) to 24.2 kg/m^2 (CI: 24.0, 24.4) in men, and from 22.1 kg/m^2 (CI: 21.7, 22.5) to 24.4 kg/m^2 (CI: 24.2, 24.6) in women. Over the same period, the prevalence of obesity increased from 3.2% to 10.8% in men and from 6.4% to 14.9% in women[7].

The prevalence of obesity peaked 10 years earlier in men (between 50-54 years) than women (60 -64 years)[8]. On average, the global population became >1.5 kg heavier each decade. Obesity was previously deemed a problem in high-income countries, but is now on the rise in low- and middle-income countries and if the secular trends in overweight and obesity continue, about 60% of the global population will be overweight or obese by the year 2030[9].

T2D is the commonest form of diabetes and accounts for >90% of all diabetes globally. It develops slowly and the exact time of onset is impossible to determine[10]. Globally, an estimated 30-50% of those with T2D may be unaware of their disease or undiagnosed[11]. About 537 million adults (20-79 years) were estimated to have T2D globally in 2021 and this number is expected to be about 643 million by 2030 and 783 million by 2045. In fact, in the last 20 years, the number of people with T2D has risen steadily from about 151 million in the year 2000 to 537 million in 2021[11].

Prediabetes is a state of hyperglycaemia which is below the diagnosis threshold for T2D. Two distinct states define prediabetes, impaired fasting glucose (IFG) and impaired glucose tolerance (IGT), the co-occurrence of both is referred to as impaired glucose regulation (IGR). The WHO defines IFG as a fasting plasma glucose (FG) of 6.1- 6.9 mmol/L and IGT as 2-hr glucose of 7.8-11.0 mmol/L[12]. The American Diabetes Association (ADA) defines IFG as FG of 5.6-6.9mmol/L and uses the same threshold for IGT; it additionally includes HbA$_{1c}$ between 39-46 mmol/mol (or 5.7 - 6.4%) to define prediabetes[13].

Prevalence of prediabetes may vary because of the different criteria for IFG by ADA and WHO. However, in one meta-analysis the estimated prevalence of IFG in combined cohorts of Asians and Caucasians was 36% using WHO and 53% using ADA criteria respectively, while that of IGR was 15.8% (WHO) and 20.2% (ADA) respectively. In Caucasians the prevalence of IFG, IGT, and IGR was 43.9%, 41.0%, and 13.5%, respectively, while in Asians it was 29.2%, 49.4%, and 18.2%, respectively using WHO definition[14]. Globally, in 2021, about 541 million (10.6%) adults had IGT and, 319 million (6.2%) had IFG with these figures projected to rise to approximately 730 million (11.4%) and 441 million (6.9%) people by 2045, respectively[11].

Prediabetes is a high risk for progressing to overt T2D, and studies have reported annual conversion rates (from prediabetes to T2D) between 4-11%[15-17]. However, without interventions the long-term conversion rates are high as seen in the Da-Qing study, where the cumulative incidence rate of T2D was >90% after 20 years in the control group[18]. Despite this, interventions like lifestyle modification (exercise and diet), and/or medication can reduce the rate of progression to T2D[16,18,19].

By and large, obesity and dysglycaemia are attributed to environmental (including lifestyle) or genetic factors, or interaction of both, barring sickness or other causes like medication.

Environmental factors

Our societies have changed substantially over the last 50 years. Some of the most important changes, with regard to rise in global prevalence of obesity, are in the global food system. These changes, especially agricultural policies in rich countries, resulted in abundance of highly processed, affordable and aggressively marketed energy-dense foods[5,20]. In addition, "westernization" of lifestyles created conditions that promote development of obesity such as increase in sedentary time and decreased PA (more sedentary jobs, screen time, reduced walkability), inadequate sleep and increased consumption of high-energy/sugar snacks and sugar-sweetened beverages, in an environment of intensive marketing[20-22]. Additional local factors like density of fast-food outlets, limited access to nutritious food, reduced neighbourhood walkability/recreation facilities and low socio-economic status are associated with increased risk of obesity[5,20]. Indeed, change of local environment, from impoverished (with high obesity rates) to wealthier neighbourhoods (with low obesity rates) can reduce the prevalence of extreme obesity[23].

Food systems interact with environmental and individual factors leading to variation in obesity prevalence between and within populations, and between individuals[20]. Ultimately, obesity is a result of interactions between an individual's innate biology and environmental factors[24]. Most risk factors associated with of obesity are also increase T2D risk, with obesity itself being a major risk factor for T2D. Other risk factors associated with T2D are gestational diabetes in women, intrauterine environment (small gestational age, maternal obesity), smoking, low socio-economic status, and age. More recently, microbiota have also been linked to obesity and especially when dysbiosis is triggered by effects of the obesogenic environment[25,26].

Genetic factors

In any given environment, there are phenotypic variations between individuals and in the case of obesity/ or T2D, not everyone develops obesity/T2D despite being in the same risk environment. This implies that other individual-specific factors (genetics/innate biology) predispose some people to obesity/T2D and not others. Both obesity and T2D are fairly heritable with studies showing that they cluster within families. In twin, family and adoption studies, the heritability (proportion of phenotypic variance explained by heritable factors) of BMI is estimated to be 40-70%[27,28]. On average, the heritability of BMI adjusted for age and sex is estimated to be between 40-50%[29].

From a genetic perspective, obesity can either be monogenic (syndromic/non-syndromic) or polygenic. Monogenic obesity is inherited in a Mendelian pattern (i.e., traits are passed from parent to child either as autosomal dominant or recessive, or X-linked dominant or recessive) and is caused by chromosomal deletions or single gene defects. Typically, monogenic obesity presents as early onset and severe obesity and is rare in the general population. Examples of monogenic obesity include mutations in the leptin and leptin receptor genes and the melanocortin pathway (*PCSK1, MC4R* and *POMC*). Mutations in genes of the leptin-melanocortin pathway which regulates feeding behaviour causes hyperphagia and severe obesity[30].

Polygenic obesity, on the other hand, is a consequence of hundreds or thousands of gene variants, each with relatively small effects working together, and follows a pattern of heritability similar to other complex traits/diseases. GWA studies, scanning the genome for genetic variants associated with BMI and other obesity related traits, have discovered single nucleotide polymorphisms (SNPs) associated with polygenic obesity. Among the most studied loci, the fat mass and obesity-associated (*FTO*) gene is associated with BMI and other obesity traits in both adults and children in diverse populations[31-34]. More discoveries have been made by research consortia formed to take advantage of increased sample size and power to discover more genetic variants. The Genetic Investigation for Anthropometric Traits (GIANT) consortium has the most recent GWAS for BMI, which included about 800,000 participants and identified more than 750 loci[35]. While most of these SNPs are yet to be fully elucidated, computing polygenic risk scores weighted by the variants' respective effect sizes provides a way of assessing their combined effect on variation of BMI or the respective trait. More recently, whole exome sequencing (WES) has facilitated exome-wide discovery studies and for BMI, a recent study (N = 640,000) identified 16 genes with rare variants associated with BMI[36]. Epigenetic modifications(heritable changes in gene expressions without changes in DNA sequence, via DNA methylation and histone modification) have also been associated with both obesity and T2D[37]

Pathophysiological mechanisms in obesity

Obesity is a disorder of chronic positive energy balance which results in storage of excess energy in the form of lipids in adipocytes[38]. Energy homeostasis is regulated by a system of multiple biological processes that work in concert to regulate acquisition, metabolism and storage of energy. This system responds to stimuli of demand, availability, expenditure or storage and initiates processes to maintain the homeostatic set-point of energy.

The central (CNS) and autonomic (ANS) nervous systems play a major role in mediating energy balance by influencing food intake behaviour and functions like brown fat thermogenesis[39]. This short and long-term energy balance is controlled via a coordinated network of central mechanisms and peripheral signals from adipose tissue, pancreas, the gastrointestinal tract (including microbiome), liver and other organs. The hypothalamus plays a central role in integrating sensory inputs that relate to energy balance and initiating autonomic, endocrine and behavioural homeostatic responses[38,39]. Other regions outside the hypothalamus contribute to energy regulation through sensory-signal input, cognitive processes (self-control of eating, adhering to actions like exercise), hedonic effects of food consumption, memory and attention[39-43]. Disorders in these tightly regulated systems lead to responses that promote energy intake and subsequent weight gain. Figure 1 depicts a summary of mechanisms involved in weight regulation.

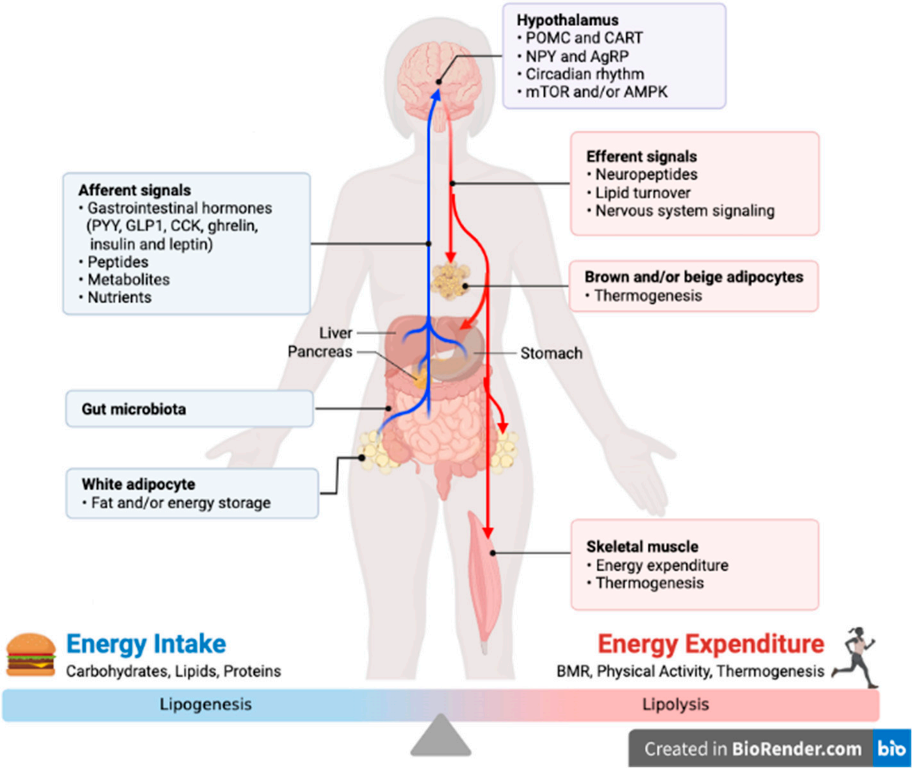
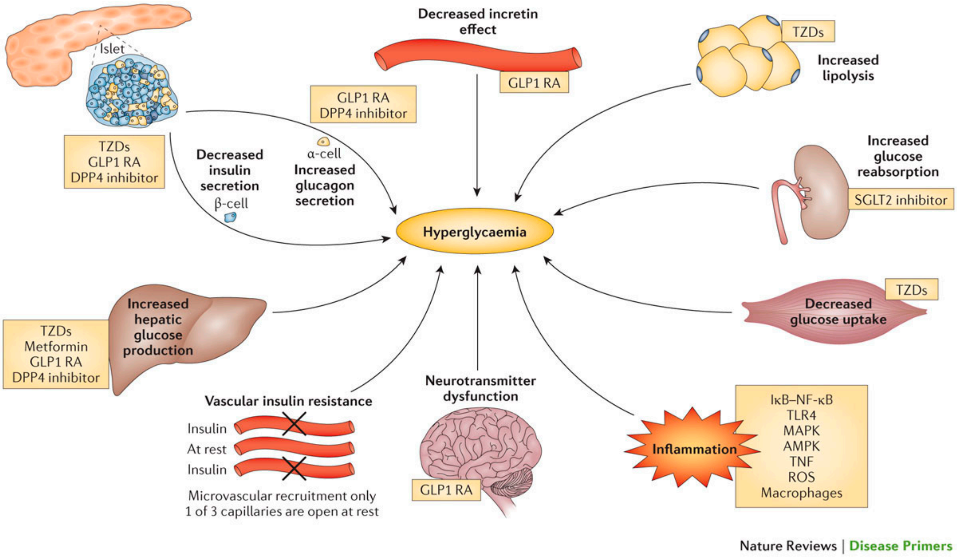


Figure 1
Factors related to weight regulation. POMC = Pro-opiomelanocortin, CART = cocain- and amphetamine-regulated transcript, NPY = neuropeptide Y, mTOR = mechanistic target of rapamycin, AgRP = Agouti-related protein, PYY = peptide YY, GLP1 = glucagon-like peptide-1, CCK = cholecystokinin, BMR = Basal metabolic rate. Created with www.BioRender.com

Pathophysiological mechanisms of dysglycaemia

The processes that lead to overt T2D are a continuum with earliest disorder of glycaemic control being insulin resistance. Reduced insulin sensitivity and the compensatory rise in insulin secretion occur years before T2D diagnosis[44,45]. T2D mainly results from progressive beta-cell failure (inability to secrete insulin), in the context of insulin resistance in the liver, adipose tissue and skeletal muscle[46]. DeFronzo described the "ominous octet" of eight pathophysiological processes that cause hyperglycaemia and eventual T2D [46], depicted in figure 2 below.



Islet
TZDs
GLP1 RA
DPP4 inhibitor
Decreased insulin secretion
β-cell
α-cell
Increased glucagon secretion
GLP1 RA
DPP4 inhibitor
Decreased incretin effect
GLP1 RA
TZDs
Increased lipolysis
Increased glucose reabsorption
SGLT2 inhibitor
Hyperglycaemia
TZDs
Metformin
GLP1 RA
DPP4 inhibitor
Increased hepatic glucose production
TZDs
Decreased glucose uptake
Vascular insulin resistance
Insulin
At rest
Insulin
Microvascular recruitment only
1 of 3 capillaries are open at rest
Neurotransmitter dysfunction
GLP1 RA
Inflammation
IκB–NF-κB
TLR4
MAPK
AMPK
TNF
ROS
Macrophages
Nature Reviews | Disease Primers

Figure 2.

Pathophysiological mechanisms in T2D. The components of the "ominous octate" include muscle insulin resistance which reduces glucose intake; hepatic insulin resistance with excessive gluconeogenesis, insulin resistance in adipocytes with increased lipolysis. Insulin promotes lipogenesis and inhibits lipolysis and thus when insulin resistance develops in adipose tissue the inhibition of lipolysis is impaired. Uncontrolled lipolysis leads to high FFA levels in circulation and when the liver and muscles are constantly exposed to these levels, there is increased uptake and storage of ectopic fat which further worsens insulin resistance in these organs[47]. Progressive beta-cell failure and apoptosis leads to decreased insulin production, and an increase in glucagon secretion and increased hepatic sensitivity to glucagon. Further there is blunted incretin effect due to beta-cell resistance to glucagon-like peptide (GLP-1) and glucose-dependent insulinotropic peptide (GIP). In the kidney there is increased gluconeogenesis and glucose reabsorption by renal tubules while in brain tissue insulin resistance and neurotransmitter dysfunction lead to impaired appetite stimulation and weight gain. To the ominous octet, two more mechanisms have been included namely vascular insulin resistance and inflammation[48]. Boxes represent therapeutic agents targeting that particular pathophysiological mechanism. Picture source[48]

Impaired insulin signal transduction, which leads to insulin resistance, results in decreased glucose transport into the cell and defective intramyocellular glucose metabolism in muscles, while in the liver, basal hepatic gluconeogenesis is unresponsive to insulin and overproduction of glucose occurs in the presence of elevated plasma insulin levels[46]. Further increase in gluconeogenesis is driven by elevated glucagon levels in circulation with enhanced hepatic sensitivity[49]. Beta-cell lipotoxicity and glucotoxicity due to elevated FFAs and hyperglycaemia respectively, impair insulin secretion which further promotes hepatic gluconeogenesis[50 51]. In addition, there is impaired uptake of glucose by the liver after glucose ingestion due to impaired incretin-induced potentiation[46]. This hyperglycaemic environment is accompanied by increased insulin production by beta-cells up to a point where the compensatory mechanisms cannot offset the effects of insulin resistance. At the same time, there is progressive decline in beta-cell mass effectively reducing the amount of insulin secreted, eventually leading to overt T2D disease[52,53].

Consequences of Obesity

Excess adiposity interferes with health and overall wellbeing in multiple ways, ranging from psychosocial to anatomic and metabolic effects. Obesity reduces quality of life, leads to low self-esteem, stigma, decreased productivity and increased risk of overall mortality[5,8,20]. Adipose exerts pressure on surrounding organs and structures in addition to increased weight-bearing on joints, especially the knees, and spine. Increase intra-abdominal pressure may cause reflux and associated chemical injury to the oesophagus[54].

Obesity substantially increases the risk of metabolic diseases, for example T2D and fatty liver disease, and remains one of strongest predictors of T2D. It is also linked to cardiovascular diseases (hypertension, myocardial infarction and stroke), musculoskeletal disease (osteoarthritis, owing to pressure degradation and inflammation), mental and neurological disease, depression, sleep apnoea, gallstones and some types of cancer (for example, breast, ovarian, prostate, liver, kidney and colon)[8,22,48,55-57]. Excess adipose tissue becomes inflamed with macrophages and other immune secreting proinflammatory cytokines (e.g., interleukin 6) in addition to peptides and metabolites released during adipose remodelling[58]. There is also a decrease in levels of adiponectin which is anti-inflammatory[59]. These processes create an environment of generalized inflammation, with elevated C-reactive protein levels, which leads to insulin resistance, endothelial damage, hypercoagulability, and subsequent cardiovascular diseases; essentially linking inflammation to the complications of obesity[60]. Figure 3 summarises the consequences of obesity and the associated intermediate processes.

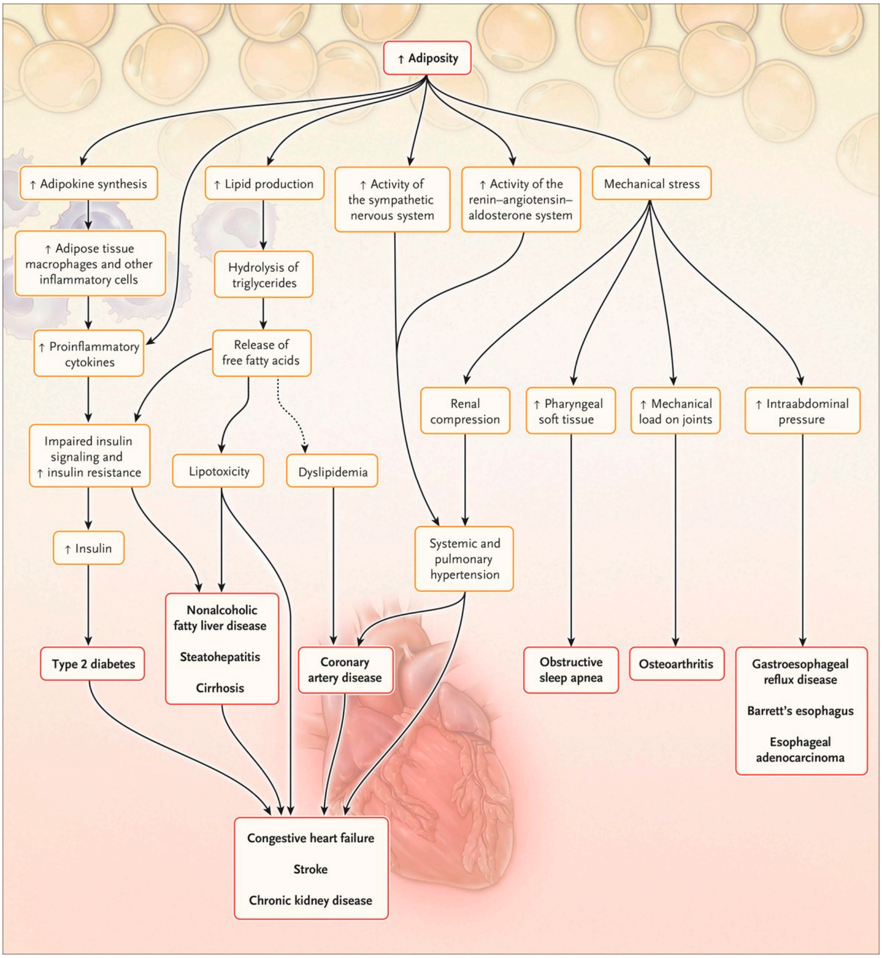

↑ Adiposity
↑ Adipokine synthesis
↑ Lipid production
↑ Activity of the sympathetic nervous system
↑ Activity of the renin–angiotensin–aldosterone system
Mechanical stress
↑ Adipose tissue macrophages and other inflammatory cells
Hydrolysis of triglycerides
↑ Proinflammatory cytokines
Release of free fatty acids
Renal compression
↑ Pharyngeal soft tissue
↑ Mechanical load on joints
↑ Intraabdominal pressure
Impaired insulin signaling and ↑ insulin resistance
Lipotoxicity
Dyslipidemia
↑ Insulin
Systemic and pulmonary hypertension
Nonalcoholic fatty liver disease
Steatohepatitis
Cirrhosis
Type 2 diabetes
Coronary artery disease
Obstructive sleep apnea
Osteoarthritis
Gastroesophageal reflux disease
Barrett's esophagus
Esophageal adenocarcinoma
Congestive heart failure
Stroke
Chronic kidney disease

Figure 3.
Diagram depicting the various consequences of obesity. Source[54]

Non-alcoholic Fatty liver disease (NAFLD)

NAFLD is spectrum of liver diseases manifesting as hepatic steatosis (NAFLD - the first stage recognizable when fat content >5% of liver volume), non-alcoholic steatohepatitis (NASH), fibrosis and eventual cirrhosis[61]. NAFLD affects about 20-25% of the general population globally with highest prevalence seen in obesity and T2D[61,62]. NASH, the second stage of NAFLD, is characterized by the occurrence of inflammation and affects about 3-5% of the general population, predisposing sufferers to progressive liver fibrosis, cirrhosis and hepatocellular carcinoma[63,64].

Diagnosis of NAFLD requires exclusion of other liver diseases like alcoholic liver disease, Wilson's disease, and infections like viral hepatitis.

Risk factors associated with NASH include diabetes, hypertension, dyslipidaemia and obesity (especially visceral obesity)[63]. NASH occurs as a result of increased lipid synthesis in the liver, decreased utilization of lipid stores and impaired oxidation of free fatty acids (FFAs), processes that promote macro-vesicular steatosis (lipid deposition in hepatocytes). Further, oxidative stress leads to hepatocyte injury and subsequent release of cytokines, which can lead to fibrosis over time [63,64].

Diagnosis of NASH is done via liver enzyme tests (alanine aminotransferase - ALT and aspartate aminotransferase - AST), liver biopsy (the gold-standard), ultrasonography or magnetic resonance imaging (MRI) (where feasible), though the accuracy of imaging is contested and invasive biopsies carry complications[61,64]. There are no proven treatments specifically for NAFLD and screening programs are not recommended, due to uncertainties of diagnostic tests and treatment options. However, lifestyle change with sustained weight loss are reported to reduce liver fat and improve insulin sensitivity and glycaemic control[61,63,65].

Observational and experimental studies

The associations between prediabetes and vascular outcomes, and those between excess adiposity and cardiometabolic outcomes have been studied extensively in epidemiology. In multiple cohorts, prediabetes has been associated with increased risk of vascular complications [66-69] with some studies demonstrating a direct relation between levels of baseline hyperglycaemia and risk of vascular complications[70,71]. Numerous studies have provided observational evidence of the multiple deleterious effects of obesity on different organ systems[22,55,56,72,73]. Further evidence from observational studies on effects of weight loss show reduced risk of cardiovascular disease (CVD) end points and other outcomes associated with obesity[74,75].

Inferring causality from observational studies is unreliable owing to their inherent shortcomings namely confounding, bias and reverse causality. To mitigate against these, RCTs are undertaken to establish causal relationships between an exposure and outcome. In one trial, lowering glucose levels in participants with IGT significantly reduced the relative risk of CVD events by 49% (CI: 5-72%) and 34% reduction in relative risk of hypertension (CI: 11-51%)[76]. In the DPP, Da Qing and the DPS studies, lifestyle and/or pharmacological interventions in participants with impaired glucose regulation resulted in improved glycaemia and delayed onset of T2D[16,18,19]. A more recent meta-analysis of RCTs showed decreased risk of CVD, mortality and cancer after weight loss interventions[77]. However, RCTs are not feasible in all situations and we are not sure whether the benefits of weight loss observed are due to actual reduction in adiposity levels or other risk factors that are correlated with adiposity. Therefore, alternative methods like MR that are robust to weaknesses of observational studies and are applicable to epidemiological data, offer a particularly useful design to investigate causality.

MR in causal research

MR is a method borrowed from econometrics that uses instrumental variables (IV) to estimate causal effects of an exposure on an outcome[78,79]. An instrumental variable is considered a proxy of the exposure which under certain conditions can be used to infer causality between exposure and outcome. To be considered an IV, the variable must be associated with the exposure, must exert its effects on outcome only through the exposure, and must not be associated with any confounders of the exposure–outcome relationship. MR uses genetic variants (SNPs) as instrumental variables and is therefore not affected by the weaknesses of observational studies (chance, confounding, bias and reverse causality) making it an ideal method to estimate causality, provided the SNPs meet the IV criteria[80]. More about this is discussed in the methods section.

Causal effects of prediabetes have been investigated previously using MR though studies used fewer instruments (SNPs) or exclusion of T2D was unclear[81,82]. In obesity causal studies using MR, different numbers and types of instruments (SNPs) were used as well different cardiometabolic outcomes[83-85]. In a recent meta-analysis of MR studies, BMI was associated with T2D, circulatory diseases neoplasms and NFLD[86]. Sometimes, we want to assess the nature of the causal relationships identified, i.e., if the causal relationship between exposure and outcome linear or non-linear. There is scarcity of literature on this topic, but with new methodologies, studies investigating these aspects are now being published. One study that investigated nature of causal effect of BMI on mortality found a J-shaped causal relationship[87] and another found a non-linear causal association between BMI and CKD[88].

Sex differences in adiposity and cardiometabolic risk

Sexual dimorphism in metabolism and cardiometabolic risk profile has been reported in observational studies. Women tend to store fat in gluteal-femoral subcutaneous adipose tissue (SAT) while men store fat more centrally; and on average women have higher body fat mass compared to men, who on the other hand have proportionately more muscle mass[89,90]. In addition, women tend to have higher levels of serum FFAs, intramyocellular fat and differ from men in terms of energy substrate preference, at rest and during activity[91,92]. Whether these differences are reflected in causal associations or whether they influence the nature of causal relationships between BMI (adiposity) and cardiometabolic disease is yet to be fully understood. At least one study assessed sex differences of causal effects of BMI on cardiometabolic outcomes and other leading causes of death but did not find a difference[93].

Aims of the thesis

Paper 1

T2D is a chronic disease associated with severe debilitating complications, the commonest being micro- and macrovascular, especially coronary artery disease (CAD), that lead to high morbidity and mortality. Prediabetes is characterized by hyperglycaemia that is below the threshold for T2D diagnosis. Whether prediabetes is causally linked to these vascular complications independent of T2D status, or a mere antecedent event to the diagnosis of T2D is not clear. We investigated whether prediabetes is a non-causal prelude to disease, or whether it is a causal factor of T2D complications. Therefore, the aim of this study was to investigate the association between prediabetes and common micro- and macrovascular complications of T2D by first summarizing the observational evidence via meta-analysis and then using MR to conduct causal inference analyses. Here, we utilized summary GWAS data from various consortia to estimate causal effects of prediabetes on CAD, CKD and stroke.

Paper 2

In mammals, the leucin-rich repeat immunoglobulin-like domains (*LRIG)* are transmembrane proteins putatively associated with cancer (as aetiological and prognostic factors), of which three subtypes, *LRIG1, LRIG 2* and *LRIG3* are found in vertebrates. Functional studies of these proteins in mouse models have been hindered by inviability of *Lrig*-null mice. In the nematode *C. elegans*, the homolog of *LRIG, Sma-10,* regulates body size (mutant worms are smaller than wild-type) via bone morphogenic protein (BMP) signalling, and may also regulated lipid metabolism[94]. Little is known about the effects of this gene in humans, specifically adiposity phenotypes, T2D risk and markers of lipid metabolism.

This study combined molecular biology, adipocyte biology and epidemiological aspects, essentially linking results from functional studies to population level effects. The aims of this study were to analyse the physiological and molecular functions of LRIG proteins in isogenic cells (cells with identical genes); investigate the effects of *Sma-10/LRIG* mutation in *C. elegans*; investigate the relationships between *LRIG1* SNPs on human adipocyte morphology and metabolic traits (BMI, T2D and lipid biomarkers). We used data from the GENiAL cohort for adipocyte morphology investigations, and data from the UKB for metabolic phenotypes relationships with *LRIG1* SNPs. My role in this study was in designing and conducting the epidemiological part of the project in the UKB. For purposes of this thesis, I will describe the epidemiological methods and refer the reader to the main paper (included at the end of the thesis) for wet lab methods.

Paper 3

One of the consequences of energy dyshomeostasis is ectopic deposition of lipids in the liver, leading to development of NAFLD. Definitive diagnosis of NAFLD is via liver biopsy, which is invasive and carries risk of complications like bleeding, infections and pain days after the procedure. Imaging modalities like MRI are expensive and not routinely available, while the more common ultrasonography has limitations: inter-operator variability, challenges scanning obese patients, and the inability to distinguish between different stages of NAFLD (which is not made any better by the lack of a standard grading system)[61]. Current prediction models vary in their performance and use different parameters and therefore there is opportunity to improve prediction of NAFLD by leveraging on data from commonly collected clinical variables and omics data. The aim in this study was thus to develop prediction models for NAFLD via machine learning (ML) using common clinical variables and omics data. In this study we used data from the IMI-DIRECT cohort to develop prediction models and the UKB for validation. My role in this project was dimension reduction and feature selection through the least absolute shrinkage and selection operator (LASSO), evaluating performance metrics and data visualization.

Paper 4

This was my most independent paper, which I worked on in the final phases of my PhD and I was responsible for everything. It is part of a project I designed to explore causal effects of adiposity on cardiometabolic outcomes, and to unravel the nature of these causal relationships.

Most MR studies assume a linear relationship between exposure and outcome, which may not be the case. For instance, observational studies have reported J-shaped relationship between BMI and mortality. Further, men and women differ in their body fat distribution and energy metabolism, which drive a sex dimorphic risk profile of cardiometabolic disease. Whether causal relationships between BMI and cardiometabolic outcomes are linear or whether there are significant differences between men and women in these causal associations is yet to be extensively investigated. In this paper, therefore, the aim was to elucidate the nature of the causal effects of BMI on cardiometabolic outcomes and risk factor biomarkers and investigate sex differences within the same causal framework. In this study, I used data from the UKB in all the analyses. In addition, effect sizes for computing the BMI polygenic risk score (PRS), used as an instrument in MR analyses, were obtained from the latest GWAS of BMI that did not include UKB participants[95].

Chapter 2: Cohorts used

GWAS Summary data

In causal inference analyses for prediabetes (paper 1) we used summary GWAS data that is publicly available from different consortia. For fasting glucose (FG), we used data from the Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC). The MAGIC GWAS meta-analysis included 32 cohorts with 133,010 participants of European descent[96]. HbA_{1c} GWAS summary data were obtained from the most recent MAGIC transethnic GWAS meta-analysis of HbA_{1c} which comprised of 82 cohorts with 159,940 participants of different ancestries (European, South and east Asian, and African)[97]. For our analysis we used data from individuals of European descent only (n = 120, 962) from the transethnic meta-analysis.

GWAS summary statistics for CAD were obtained from the latest cardiomics GWAS meta-analysis[98] which consisted of 34,541 cases and 261,984 controls from the UKB and was replicated in 88,192 cases and 162,544 controls from the Coronary Artery (C4D) Genetics consortium (CARDIoGRAMplusC4D)[99,100]. Summary statistics for stroke and stroke subtypes were obtained from the MEGASTROKE consortium which is a meta-analysis of 40,585 cases and 406,111 controls of European ancestry[101]. Renal disease summary data were obtained from the CKDgen consortium which is a meta-analysis for CKD (eGFRcrea <60ml per min per $1.73m^2$) performed on 745,348 participants and replicated in a further 280,722[102]. T2D data, used to excluded T2D-associated SNPs from the FG and HbA_{1c} data, were obtained from the latest T2D GWAs meta-analysis of 81,412 cases and 370,832 controls of diverse ancestries followed by fine-mapping in 50,160 T2D cases and 465,272 controls of European ancestry[103]. From these data, we used summary statistics from participants of European ancestry only.

GENiAL cohort

The GENiAL cohort participants were enrolled between 1986-2016 via local advertisements in Stockholm, Sweden. A total of 939 participants with adipocyte lipolysis measurements were included of which 57% were obese, 194 had T2D, hypertension or dyslipidaemia, alone or in different comorbid combinations. All participants gave written informed consent and the study was approved by the local ethics board[104].

Following an overnight fast, participants presented at the Karolinska clinical research centre for anthropometric measurements including body fat content (using bioimpedance). Blood samples (venous) were obtained for genotyping and biochemical assays. Needle aspiration biopsy was used to obtain SAT samples next to the umbilicus[104].

SAT samples were cleaned of blood vessels and cellular debris in sodium chloride and treated with collagenase to isolate adipocytes which were incubated using a previously described protocol[105]. For genetic analysis, the UKB Axiom Array r3 platform was used for genotyping samples and the Axiom analysis suite for genotype calls. Samples with cryptic relatedness, ambiguous sex, or call rates <95% were excluded while SNPs were excluded based on Hardy-Weinberg equilibrium (HWE) $p < 5x10^{-6}$, MAF < 1% and SNP call rates <95%. Imputation was done using the haplotype reference consortium (HRC) panel and 1000G phase 3 panel when variants were missing. In further quality control (QC) after imputation, SNPs were excluded if minor allele counts were < 3 and INFO score <0.4; in addition to related participants – only one in a pair of 1st or 2nd degree relatives[104].

UK Biobank (UKB)

The UKB is an open-access resource accessible to researchers from all-over the world, after application and approval, to conduct health research of public interest[106]. It is a prospective cohort of approximately 500,000 participants of mixed ancestries (European, Asian and African), aged between 40-69 years at the time of enrolment, recruited and assessed across 22 centres in the United Kingdom from 2006 to 2010. Response rate was 5.4% (with regional differences) and participants were more likely to be women, older, more affluent and healthier than the general population. In general, the UKB is not representative of the general UK population (though it is ethnically representative) but nonetheless its large size and extensive data would provide useful inferences, albeit generalizable to other populations[107].

Participants provided electronic signed consent at enrolment, including consent for follow-up via linkage to their health records, and answered interview questions on socio-demographic, lifestyle and health-related factors. Standard anthropometric and other physical measurements were taken, in addition to biological samples (urine, blood and saliva) for biochemical assays and preservation for future assays. When it was determined that recruitment was progressing well, further assessments were included in the visits like electrocardiogram (ECG), arterial stiffness test, hearing test and a range of eye measures. To account for measurement calibration, correct regression dilution and estimate longitudinal changes, repeat assessments were undertaken in subset of the participants every few years. In 100,000 participants, objective measures of PA were collected between 2013-2014 with repeat measures taken from 2500 of them[107].

The UKB study received approval from the Multi-centre Research Ethics Committee (REC ref: 16/NW/0274) and all participants gave informed consent[108]. Information about recruitment and data collection has been detailed elsewhere[106]. In Sweden, use of UKB data was approved by the Swedish Ethics Approval Authority (*Etikprövningsmyndigheten, EPM*), application number 2021-03174.

Genetic data

The UKB cohort contains genotypes from 488,377 subjects assayed using two closely related genotyping arrays. A sub-cohort of 49,950 participating in the UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) study were genotyped using the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix (now part of Thermo Fisher Scientific), at 807,411 markers. Subsequently, 438,427 participants were genotyped using the Applied Biosystems UK Biobank Axiom Array (825,927 markers). Genotypes that were not directly assayed were imputed using Haplotype Reference Consortium (HRC) data as the main panel, or a combination of UK10K and 1000 Genomes Phase 3 reference panels. Imputed data from both imputation processes were combined and in cases where a SNP was present in both panels, HRC imputation was used. Quality control (QC) was conducted by the UKB data team and the results were shared with researchers for their downstream analyses. For this project, we used the imputed data release version 3 (v3) from the UKB. Details of enrolment and genetic data handling are further explained in Bycroft *et al.*[106].

In our analyses, we excluded samples with less than 99% genotype call rate, SNPs with Hardy-Weinberg equilibrium $P < 1 \text{ x}10^{-10}$, those with less than 80% imputation score and any duplicated SNPs. We further selected SNPs with a minor allele frequency (MAF) > 0.01, and excluded individuals who were outliers for heterozygosity, those with indeterminate sex and aneuploidy, and one of a pair of related individuals (up to 3rd degree relatedness, kinship coefficient 0.0442 – 0.0882).

IMI-DIRECT consortium

The Innovative Medicines Initiative (IMI) is a collaboration between the European Union (EU), twenty European academic institutions and five pharmaceutical companies in Europe. The DIRECT consortium, formed under the IMI, aims to discover potential biomarkers of disease progression or response to T2D therapies; which could address challenges in drug development and help develop stratified approaches to T2D management[109]. Towards this end, DIRECT initiated two multi-centre studies focused on discovering novel biomarkers of glycaemic deterioration in persons at high risk, or newly diagnosed with T2D.

The first cohort (cohort 1) enrolled participants at risk of dysglycaemia from four existing cohorts and focused on glycaemic deterioration before T2D onset. Persons at risk of dysglycaemia were identified using a sex-specific validated prediction tool (DIRECT-DETECT) based on baseline age, BMI, waist circumference, antihypertensive medication, smoking, parental diabetes, and change in HbA_{1c} during follow-up. Identified participants were included for screening if they were white European, aged between 35-75 years, had baseline FG (capillary) of <10mmol/L and not on T2D treatment. Exclusion was based on: previous diagnosis of any form of diabetes, use of a pacemaker, medical reasons, and for women if pregnant (or planned to be during the study) or lactating[109,110].

Cohort 2 (T2D) participants were identified through clinical practice, existing databases, educational clinics, routine retinal screening programmes and other registries. Inclusion criteria were T2D diagnosis strictly in the last 6-24 months before baseline examination, on lifestyle management of T2D with or without metformin, all Hba_{1c} <7.6% in the previous 3 months, and estimated glomerular filtration rate (EGFR) > 50 ml/min. Age, ethnicity and women-specific considerations were as for cohort 1. Participants were excluded based on; a diabetes diagnosis other than type 2, previous HbA_{1c} > 9.0%, prior insulin use or antidiabetic medication apart from metformin, BMI <20 kg/m^2 or >50 kg/m^2, or medical reasons. The final sample sizes after applying inclusion and exclusion criteria and data quality control, were 2127 and 789 participants for cohort 1 and 2 respectively[109,110]. Cohort 1 participants were followed-up at 18 and 36 months while cohort 2 were followed-up at 9 and 18 months after baseline visit[110].

In both cohorts, baseline measurements were done after a 10-hour overnight fast. Anthropometric and blood pressure (BP) measurement procedures were standardized across all study centres and performed by trained personnel. Blood samples were collected for omics (metabolomic, proteomic, genomic, epigenomic, and transcriptomic), beta-cell and insulin sensitivity indices, lipids (TG, cholesterol, HDL, LDL), liver enzymes (ALT, AST). Abdominal MRI scans were performed and local protocols were standardized across centres to harmonize scanning methodology.

Dietary assessment was done a day prior to visiting the study centre using a 24-hour multi-pass dietary record, a validated method with three levels of questioning[110]. Objective assessment of PA, sedentary time and sleep was assessed using a triaxial accelerometer (ActiGraph GT3X+; Actigraph LLC, Pensacola, FL, USA) worn in the participant's non-dominant wrist, worn 10 days in a row without removing. Additional information collected in questionnaires included quality of life, dental health, T2D family history and medication history.

Cohort 1 participants had frequently sampled oral glucose tolerance test (fsOGTT) at baseline, and at the 18 and 36-month follow-up visits, same measurements/assessments as baseline were conducted. Cohort 2 participants remained on their usual lifestyle treatment but if on metformin they paused 24 hours before baseline assessment and restarted immediately thereafter. Blood samples were collected for glutamic acid decarboxylase (GAD) and islet antigen-2 antibodies, GLP-1, glucagon, insulin, C-peptide, metabolomics, proteomics, HbA_{1c}, DNA and RNA, specifically for this cohort. In addition, a mixed meal tolerance test (MMTT) was performed for glucose, insulin and C-peptide analysis; and a further postprandial urinalysis. The 9 and 18 months follow up visit assessments were similar to baseline with the exception that blood samples for RNA analysis were not collected[110].

All study centres in DIRECT received ethical approval from their country's respective ethical review boards (there was no pan-European research ethics approval body), and all participants in each respective cohort gave written informed consent.

Genetic data

Genotyping was done using the Illumina HumanCore array (HCE24 v1.0) and genotypes called using Illumina's GenCall algorithm. Samples with a call rate <97%, heterozygosity, sex discordance and monozygosity were excluded. In further QC, samples were restricted to a genotype call rate >99%, HWE exact $p < 0.001$, MAF >1% and variants mapped to human genome build GRCh37. Duplicated variants were excluded[110].

Chapter 3: Methods

Linear and logistic regression

Regression modelling techniques are widely used in clinical and biomedical research to investigate associations between an exposure and outcome or to make predictions about the outcome. These inferential methods require several assumptions about the data and nature of relationships between exposure and outcome to ensure valid conclusions. The most widely used are the linear and logistic regression, with choice of the method determined mainly by the nature of the outcome being investigated.

Linear regression is used to examine the linear association between a continuous outcome and one or more independent variables. The pertinent assumptions are: normally distributed model residuals with a mean of zero, constant residual/model error variance, (homoskedasticity) and independence of observation units. A simple linear regression model can be represented as

$$y = \alpha + \beta x + \varepsilon \tag{1}$$

Where y represents the outcome, α is an intercept term, β is the regression coefficient of the exposure and ε is an error term.

In logistic regression the examined association is between a categorical outcome and one or more independent variables. Logistic regression uses a link function to connect the outcome variable to a linear function of the exposure variable. This method assumes the association between log-outcome and the exposure is linear, in addition to the assumptions of linear regression. Both methods are also used to predict the future values or odds/ probability of an outcome based on one or more exposure variables[111]. I used these methods in paper 2 to assess the relationships between the *LRIG1* gene variants and BMI, T2D and lipid metabolism biomarkers.

Systematic review and Meta-analysis

Sometimes we want to summarize the evidence about a particular topic in a particular field; of which systematic reviews with/or meta-analyses are the main methods used. Systematic reviews involve summarising and synthesis of evidence from individual studies that are selected using predefined, transparent and reproducible criteria. The studies are then screened for content relevance and those that meet the criteria are retained for detailed perusal and extraction of evidence. Meta-analysis differs from systematic reviews in that meta-analysis aims to quantitatively combine and summarise results from different studies to a point estimate. Usually, the scope of a meta-analysis is predefined and study selection is done systematically and reproducibly, and the evidence validity is assessed using standard methods[112].

It is important that effect sizes of the individual studies are comparable, computable, reliable and interpretable so as to summarise them in to a single estimate. The effect sizes of individual studies are weighted by the inverse of the variance to account for each individual study's strength of effect. For binary outcomes, alternative methods can be used to compute weighted estimates e.g., Mantel-Haenszel[113]. Meta-analyses can be performed assuming either fixed or random effects.

In a fixed effects meta-analysis, the underlying assumption is that effects come from a single homogeneous population and share the same true effect size. In random effects meta-analysis, the assumption is that there is no one true effect and that a distribution of effects exists, resulting in heterogeneity of study results, owing to additional variance because studies are not from the same single population[112]. The sources of heterogeneity can be clinical (participants, intervention, and outcome variation), methodological (study design and bias) or statistical (due to clinical or methodological heterogeneity, or both). Heterogeneity affects generalization of results from the summary estimate, but understanding its sources helps in targeting interventions[114].

Finally, bias in meta-analysis can arise from publications where studies available in the literature search pool are usually large studies reporting positive results and this publication bias can be assessed using a funnel plot or Egger's test[115].We used meta-analysis in paper 1 to summarise the observational evidence of association between prediabetes and vascular complications, and in paper 4 to summarise the causal effect of obesity on *any* cardiometabolic disease.

Causal inference analyses

In observational studies, the confidence with which one can infer causal relationships is often diminished owing to concerns about confounding, reverse causality and bias which are characteristic weaknesses of such studies[116]. The question of causality in epidemiological research is therefore an important one given that the gold standard for causality, RCTs, are not always feasible in all settings, are expensive, undertaken in selected subgroups that are not necessarily free-living. In some settings, epidemiological studies are the best choice. Therefore, methods that can bypass these obstacles and at the same time address the core weaknesses of observational studies provide a powerful tool for use in causal inference. If applied correctly, MR can meet these criteria.

Mendelian Randomization (MR)

MR takes advantage of the random assignment and independent assortment of homologous chromosomes during meiosis, which is unaffected by any confounders of the exposure outcome relationship or reverse causality. In MR, SNPs associated with an exposure are used as instrumental variables to test the causal relationship between the exposure and an outcome of interest. To be a valid instrument, a SNP should fulfil the criteria for instrumental variables, here referred to as the MR assumptions/conditions.

Let G be the genetic variant used as the instrument, X and Y be the exposure and outcome, respectively, and we are interested in the causal effect of X on Y. Also, let U represent confounders of the association between X and Y. The instrument, G, must satisfy the following conditions:

G is independent of U, i.e., it is not associated with any of the confounders of X and Y; G is associated with X; and G is not directly associated with Y but only through X (Figure 3).

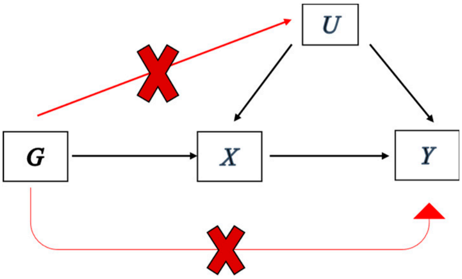


Figure 3
MR assumptions. Red path and cross indicate violations of the instrumental variable conditions.

One-sample MR (Two Stage least squares, TSLS)

TSLS involves two sequential stages and utilizes individual-level data to estimate causal effects. The instrumental variable used could be a single SNP or multiple SNPs combined in a PRS. For purposes of this thesis, I used a BMI PRS as the instrumental variable. In addition to the notation introduced above, let *L* represent a vector of covariates to adjust for. The first stage of TSLS involves regressing the exposure *X* on the instrumental variable, *G,* while adjusting for relevant covariates, *L*. Then, fitted values of the exposure, $\hat{X}$, 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. Depending on the nature of the outcome, the second stage can be a linear or logistic regression.

The first stage is represented as:

$$X_i = \alpha_0 + \sum_{k=1} \alpha_k G_{ik} + \alpha_1 L_i + \varepsilon_{Xi} \tag{2}$$

and second stage as:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \beta_2 L_i + \varepsilon_{Yi} \tag{3}$$

Where *i* is the individual index, *i* = 1, …, N, *k* is the number of instrumental variables and α and β represent regression coefficients and ε represents the error term. I used TSLS in paper 4 to estimate causal effects of BMI on cardiometabolic outcomes since I had access to individual-level data. TSLS provides a consistent method of estimating causal effects while allowing for covariate adjustment and stratified analyses.

Two Sample MR (TSMR)

In some situations, individual-level data are not available due to practical or legal/confidentiality issues related to data sharing or archiving. Two sample MR (TSMR) obviates the absence of these data and uses GWAS summary statistics (SNPs and their respective effect sizes) to estimate causal relationships between an exposure and outcome of interest. The SNPs must meet the IV conditions detailed above. Assuming all the genetic variants are instrumental variables, these summary data (effect size coefficients and standard errors), can be combined to generate causal effect estimates[117,118]. Using multiple variants (SNPs) increases the power of an MR study compared to using a single genetic variant[119].

Let *k (*indexed *k = 1,…,k)* represent a SNP associated with observed mean change in exposure, X_k, per additional allele with standard error σ_{Xk} and a corresponding change in the outcome, Y_k , with standard error σ_{Yk}. For binary outcomes, Y_k can be represented as per-allele change in the log-odds/probability of an outcome[118].

The causal effect can be estimated using the inverse-variance weighted (IVW) method which combines ratios of estimates (Y_k/X_k) weighted by the inverse of the variance. The IVW (β_{IVW}) estimate is computed by:

$$\beta_{IVW} = \frac{\sum_k X_k \, Y_k \, \sigma_{Yk}^{-2}}{\sum_k X_k^2 \, \sigma_{Yk}^{-2}}$$

(4)

And the standard error is estimated by:

$$se(\beta_{IVW}) = \sqrt{\frac{1}{\sum_k X_k^2 \, \sigma_{Yk}^{-2}}}$$

(5)

I used TSMR in paper 1 to compute the causal estimates of relationship between prediabetes and vascular complications. With no access to individual-level data, this method offered a reliable and consistent alternative for estimating causal effects.

Sensitivity analyses for TSMR

In the event that even only one SNP is not a valid instrument, the causal estimate based on all variants will be biased with inflated type 1 error rates, that is, concluding a significant effect is there when it is in fact not[120]. To triangulate the evidence, other MR methods that relax the MR assumptions to various degrees or those that identify outliers/pleiotropic instruments, are used in addition to the main IVW analysis. These methods can be used to perform sensitivity analyses to assess consistency of the causal estimates[121]. The following methods were used for sensitivity analyses in paper 1.

MR-Egger

This is a method that tests for directional horizontal pleiotropy without making MR assumptions about the genetic variants. The method was originally used to assess small study bias in meta-analysis and has been applied to MR where it is used to show that bias due to pleiotropy is similar to small study bias in meta-analysis[115]. Under this method, the intercept is included as part of the regression (unlike conventional IVW where the intercept is forced to be zero), and the slope coefficient from the regression gives an asymptotically consistent estimate of the causal effect, even if all SNPs have pleiotropic effects on the outcome[121]. This assumes that the association between the instrument and the exposure is independent from its pleiotropic effects on the outcome, referred to as the InSIDE assumption (Instrument Strength Independent of Direct Effect)[122]. The idea behind the Egger regression is that stronger genetic variants ought to have causal effect estimates that are reliable than weaker ones, if the InSIDE assumption holds. Evidence of a causal effect in Egger regression is indicated by any residual genetic associations (dose-dependent) after the average pleiotropic effects of the genetic variants have been accounted for via the Egger intercept[121].

Median-weighted MR

The IVW method is said to have a 0% break-down level, because its estimates will be biased even if only one genetic variant is invalid. A simple median estimator has a 50% breakdown level and provides a consistent estimate when up to (but not including) 50% of the genetic variants are valid. This method estimates the simple median of ratio estimates but can be inefficient especially when precision of individual estimates vary considerably. However, a weighted version of the estimator can account for these variations. The weighted median method provides a consistent estimate of causal effect if at least 50% of the weight comes from valid IVs[123]. A weighted median estimator is the median of a distribution having β estimates ranked by percentiles defined by the difference between sum of the weights and half the weight of the respective genetic variant. The weights are proportional to the contribution of the genetic variant and are derived from the inverse of the variance of the ratio estimates[123].

MR-PRESSO

The MR-PRESSO (Mendelian Randomization pleiotropy residual sum and outlier) method performs three core functions: 1) detection of horizontal pleiotropy global test, 2) outlier test, and 3) distortion test - differences in causal estimates before and after correcting for outliers[124]. In the conventional IVW regression, if there is no horizontal pleiotropy then all variants are expected to have small residuals and hence closer to the regression line. Pleiotropic variants can deviate from the slope of the regression line owing to larger or smaller effect sizes than what is mediated by the exposure in question[124]. The MR-PRESSO outlier test needs at least 50% of instruments to be valid and the InSIDE assumption to hold[122]. In the global test, regression models are fitted excluding each variant in turn, here referred to as *j*, to produce causal effect estimate without the variant. Then observed residual sum of squares (RSS) is calculated as the squared difference between the observed effect of variant *j* on the outcome and the effect size estimated without *j*. The observed RSS is then compared to expected distribution of RSSs simulated with no outliers and an empirical P value calculated[124]. An outlier test compares observed RSSs of variants with the distribution of expected RSSs, and the distortion test is computed as a percentage of causal effect estimate that is attributed to significant pleiotropic SNPs.

Non-linear MR (NLMR)

Sometimes the purpose of a study is to assess the nature of the relationship between an exposure and outcome, i.e., whether it is linear or non-linear. In MR causal analyses, the underlying assumption is that the relationship between exposure and outcome is linear, which may not be the case. NLMR investigates whether the estimated causal relationship between an exposure and outcome is linear or not. The method involves calculating the local average causal effect (LACE) as a ratio of coefficients in quantiles that are based on the instrumental variable-free (IV-free) distribution of the exposure. This IV-free distribution solves the problem of inducing a spurious association between the IV and outcome where none exists if the population is stratified on the exposure directly, which would invalidate the mandatory IV conditions[125].

The IV-free exposure is calculated as the residuals of a model in which the exposure is regressed on the IV, with the IV value set to zero. This represents the expected value of the exposure if one had an IV value of zero and be viewed as the non-genetic component of the exposure[126,127]. To assess the nature of the causal relationship between exposure and outcome, quantiles of the IV-free exposure are computed and within each quantile a LACE is calculated. From these LACE values, the relationship between exposure and outcome can be estimated using fractional polynomials or piecewise linear functions[127]. In the fractional polynomial method, the LACE estimates are meta-regressed against the mean exposure in each quantile in a flexible semiparametric framework. Tests of nonlinearity are then applied to test the null hypothesis that the resultant non-linear model is no different from a linear model. Fractional polynomials are described as set of functions for fitting nonlinear relationships between of covariates which mitigate against the shortcomings of both high and low order polynomials[128]. The powers used for fractional polynomials are $P = (-2, -1, -0.5, 0, 0.5, 1, 2, 3)$, which are different from mathematical powers. Here, a power of 0 represents the natural logarithm function, and conventional powers are a subset the these chosen powers[127,128].

We define fractional polynomials of degree 1 as[127]

$$f(x) = \beta_0 + \beta_1 x^p \tag{6}$$

Where $p \in P$.

A fractional polynomial of degree 2 is defined as

$$f(x) = \beta_0 + \beta_1 x^{p1} + \beta_2 x^{p2}, \qquad if\ p1 \neq p2$$

Or

$$f(x) = \beta_0 + \beta_1 x^p + \beta_2 x^p \log(x), \qquad if\ p1 = p2 = p \tag{7}$$

Where β represents regression coefficients and p represents fractional polynomial power.

In piecewise polynomial method, a linear function is fitted for each quantile to estimate the exposure-outcome relationship with the gradient of each line segment representing the LACE estimate of the respective quantile. To ensure that each line segment starts where the previous one ends, the function is constrained to be continuous[127]. I used NLMR to elucidate the nature of causal relationships between BMI and cardiometabolic outcomes in paper 4.

Machine learning approaches

Machine learning (ML) involves techniques that apply computation algorithms to identify patterns within data, between variables or subsets or variables. The commonest use of ML in biomedical research is in prediction of outcomes or finding patterns within a data sample that represent shared uniqueness in features. The two commonly used types are supervised and unsupervised ML. In supervised ML, the pattern or solution is known and algorithms are trained to be able to make the right "decision". In unsupervised ML, there is no prior known pattern and algorithms identify groups or clusters of shared features. Supervised ML is commonly used in classification and regression problems. Classification involves predicting the correct group/class an observation belongs to given a set of predictors while regression involves predicting a continuous value[129].

With increased computation power and availability of big data (or high-dimensional data), use of ML has increased substantially in biomedical research. However, not all variables/features in a given dataset are meaningful (for predicting some group or value of interest) and therefore selection of the most informative ones is a prerequisite to building reliable ML models. This can be done via dimensionality reduction and/or feature selection, as dictated by the question at hand.

Dimensionality reduction and feature selection

A regular dataset is generally composed of observations represented as rows (n), and features or observed variables (i.e., attributes of interest for each observation) represented as columns (p). In high dimensional data, $p > n$ and this poses several important problems. Consider a data space where dimensions are represented by numerous axes. For a dataset with p variables, each is an axis in a p-dimensional space and as p increases data points within the space diminish (become sparse). That is, as the dimensions increase so does the probability of observations without similar values of a variable, and this, among other problems of high dimensionality, constitute the curse of dimensionality[130]. When $p \gg n$ (meaning p much greater than n), the normal distribution assumption is invalidated, which may lead to unreliable scientific conclusions[131]. Further, there is perfect multicollinearity and overfitting making it difficult to detect real associations/patterns in the data or determine the most important predictors (makes predictions difficult and uninterpretable). While many variables may be ideal for exploratory studies, confirmatory studies for scientific discovery require more focused variables with reduced dimensions[130].

Therefore, to overcome the curse of dimensionality, non-informative or least informative features in a dataset are removed while the most informative are retained. This is commonly done using regularization and variable selection, or dimensionality reduction techniques. The latter can be divided into linear and non-linear methods. Linear methods include Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS), Singular Value Decomposition (SVD), Independent Component Analysis (ICA), and Non-negative Matrix Factorization (NMF) among others while some examples of non-linear methods are T-distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) [131-134]. These methods are however outside the scope of this thesis and will not be discussed further. I will focus on regularization and variable selection which we used in paper 2.

In regularization and variable selection, a penalty is included in a regression model such that coefficients (effects) of least informative predictors are reduced, sometimes to zero, therefore reducing their influence on the results. These regularization techniques include LASSO, ridge regression and elastic net regression[135]. If we consider ordinary least squares (OLS) regression, the aim is to minimise the RSS, which reflects how well predictors estimate the outcome. OLS estimates have a large variance and when there are many predictors, it is imperative to identify the most important via exclusion or attenuating effects of the least informative ones.

The LASSO attenuates some of the coefficients (regularization) and sets others to 0 , by minimizing the RSS subject to the sum of the absolute value of coefficients being constrained by an upper limit, τ[136]. In feature selection, the LASSO includes variables that have non-zero coefficients and drops those with zero coefficients, effectively retaining in the model only coefficients that are most informative about the outcome[137]. The aim of LASSO, is therefore to reduce RSS, while retaining the most informative features. This can be represented as follows:

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

subject to:

$$\lambda(\alpha \sum_{i=1}^{p} |\beta_i| + (1 - \alpha) \sum_{i=1}^{p} \beta_i^2) \leq \tau \tag{8}$$

Where i represents the i^{th} individual, β is the regression coefficient and λ is the tuning parameter, which controls the amount of shrinkage applied coefficient estimates. As shown in equation, the desired value of λ minimizes the first part of the equation subject to the second constraint. α is a model selection parameter, such that when $\alpha = 0$, the model becomes a ridge regression, when $\alpha = 1$ the model is LASSO and when $\alpha = 0.5$ its elastic net. λ can be determined by cross validation where a range of values is run several times (folds) and the best value is assessed using the RSS as the accuracy metric.

Chapter 4: Results and discussions

Paper 1

In this study, which combined a meta-analysis of observational studies and causal inference (TSMR) analyses, we investigated the relationship between genetically determined non-diabetic hyperglycaemia and common micro and macro-vascular diabetic complications i.e., coronary artery disease (CAD), stroke and kidney disease. In the observational meta-analysis of 37 studies comprising a total of 1,326,915 participants, prediabetes was significantly associated with CAD (RR = 1.16; 95% CI: 1.09, 1.23; Q = 52.5, $P_{Qstat} = 0.058$; $I^2 = 27.7\%$) and stroke (RR = 1.11; 95% CI: 1.03, 1.18; Q = 28.5, $P_{Qstat} = 0.23$; $I^2 = 16\%$) but not CKD (RR = 1.05; 95% CI: 0.98, 1.12; Q = 27.2, $P_{Qstat} = 0.002$; $I^2 = 63.3\%$). These results were unchanged in sub-group analyses.

From the TSMR analyses, prediabetes was significantly associated with CAD, but not stroke (or any stroke subtype) or CKD, without evidence of directional horizontal pleiotropy (Egger intercept $P > 0.05$). Sensitivity analyses (MR-Egger and weighted median regression) yielded consistent results. Prediabetes instruments based on HbA_{1c} with only 8 SNPs were not significantly associated with CAD and there was evidence of directional horizontal pleiotropy. Table 1 below shows details of the results.

Table 1.
Estimates of causal association between genetically determined prediabetes and vascular outcomes

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

IVW inverse-variance weighted, CAD coronary artery disease, AIS any ischemic stroke, LAS large artery stroke, CES cardioembolic stroke, SVS small vessel stroke, CKD chronic kidney disease.
[a]TSMR results of the association between genetically determined HbA1c levels and CAD using robust IVW.

In further analyses assessing robustness of our prediabetes instrument, no association was found between the prediabetes instrument and T2D. In addition, relaxing the QC criteria (taking all FG SNPs with $P < 5 \times 10^{-8}$, without clumping or harmonization, regardless of their nominal association with T2D), showed associations with vascular complications but with high degree of directional horizontal pleiotropy. Lastly, using MR-PRESSO to correct for outliers did not change the results and a leave-one-out analysis did not show evidence that results were driven by one or more influential SNPs.

Paper 1 discussion

In this study, observational evidence showed that prediabetes was associated with stroke and CAD but not CKD while causal inference analyses showed a significant association with CAD but not stroke or CKD. By selecting instruments specifically associated with prediabetes only, we isolated the causal effects of prediabetes from those of diabetes, and by using MR we estimated causal effects robust to confounding, bias or reverse causality. These results show that glycaemic perturbations are likely causal of vascular complications, specifically those related to coronary disease.

Causal effects of prediabetes on cardiovascular outcomes have been investigated before though other studies used fewer instruments or were not clear about exclusion of T2D[81,82]. The clinically used threshold for T2D diagnosis does not determine onset of vascular damage but hyperglycaemia seems to, in a dose-dependent manner. Despite the lack of approved therapeutic agents for prediabetes, studies show that lifestyle and/or therapeutic interventions are beneficial[16,19]. Perhaps it is intriguing that in naturally occurring prediabetes of MODY2 (mutations of *GCK* gene) patients do not develop vascular complications or insulin resistance, and have normal post-prandial glycaemic responses and cardioprotective lipid profiles indicating a higher set-point for glucose homeostasis, since there is no progressive deterioration of glycaemic control[138,139]. The differences in glucose homeostasis set-points in MODY2 patients and the general population may explain why in the former elevated glucose levels are not detrimental while in the latter they are. This contrast shows that hyperglycaemia in the general population is a probable cause of CAD.

Further these findings may explain why cardiovascular complications are the commonest in T2D, hard to treat and the leading cause of death. It also shows that by the time T2D is diagnosed, hyperglycaemia-related pathogenesis of CAD has already begun, which probably explains the difficulties of preventing CAD in already established T2D. In the LOOK-AHEAD trial, aggressive lowering of glucose levels did not make a difference in CAD outcomes in those who had T2D, likely due to the effects of prolonged hyperglycaemia on the coronary vasculature. Glucose levels were lowered but not the risk of mortality or CVD events[140]. The overall implication of our study is that intervening on prediabetes early, before the threshold for T2D diagnosis, may prevent T2D-related CAD. In fact, there have been calls to consider individuals with IGT as diabetic since they have lost about 80% beta-cell function and are highly insulin resistant[46].

Paper 2

In functional analyses for adipogenesis in mouse embryonic fibroblasts (MEF), *Lrig*-null MEFs had impaired adipogenesis compared to wild type after adipogenic stimulation. BMP inhibitors prevented adipogenesis in wild-type MEFs while high doses of BMP4 enhanced adipogenesis in the wild-type and restored adipogenesis in the *Lrig*-null MEFs. Investigations of BMP signalling showed that *Lrig*-null MEFs had lower sensitivity to BMP4 than wild-type but in *Lrig1* and *Lrig3* single-knockout MEFs this sensitivity was apparently unaltered. These differences were not attributed to differential receptor expression.

To investigate whether individual *LRIG* alleles could rescue the *Lrig*-null phenotype, an *Lrig*-null MEF line was transduced with the inducible human alleles *LRIG1, LRIG2*, or *LRIG3*, with an empty vector serving as a control. Induction of *LRIG1* or *LRIG3* expression rescued the regular BMP sensitivity phenotype of the *Lrig*-null MEFs, whereas the induction of *LRIG2* expression, or vector control, did not. The induced expression of *LRIG1* or *LRIG3* also enhanced the adipogenesis rate of the MEFs.

In investigations of *LRIG/Sma-10* effects on lipid metabolism in *C. elegans*, mutant worms showed shorter body lengths and lower lipid accumulation compared to wild-type worms (Figure 4, panels a, b and c). Lipid accumulation, likely mediated via BMP signalling, appeared to be independent of body size regulation.

The *LRIG1* gene variant with the strongest effect on BMI was rs4856886(G). In epidemiological analyses, this SNP significantly increased BMI by approximately 0.05 kg/m^2 (95% CI: 0.03, 0.08) per each copy of the minor allele (Figure 4, panel d). The variant was also associated with decreased risk of T2D which was strengthened after adjusting for BMI, but did not show a significant association with liver fat percentage, p = 0.69. The variant was negatively associated with triglyceride (TG) levels, β = -0.007; 95% CI: -0.013, -0.002, p = 0.01.
Adipocyte analyses in the GENiAL cohort (n = 948) showed that the two strongest *LRIG1* signals from the UKB were associated with adipose hyperplasia; rs4856886 (P = 0.039) and rs9840088 (P = 0.014).

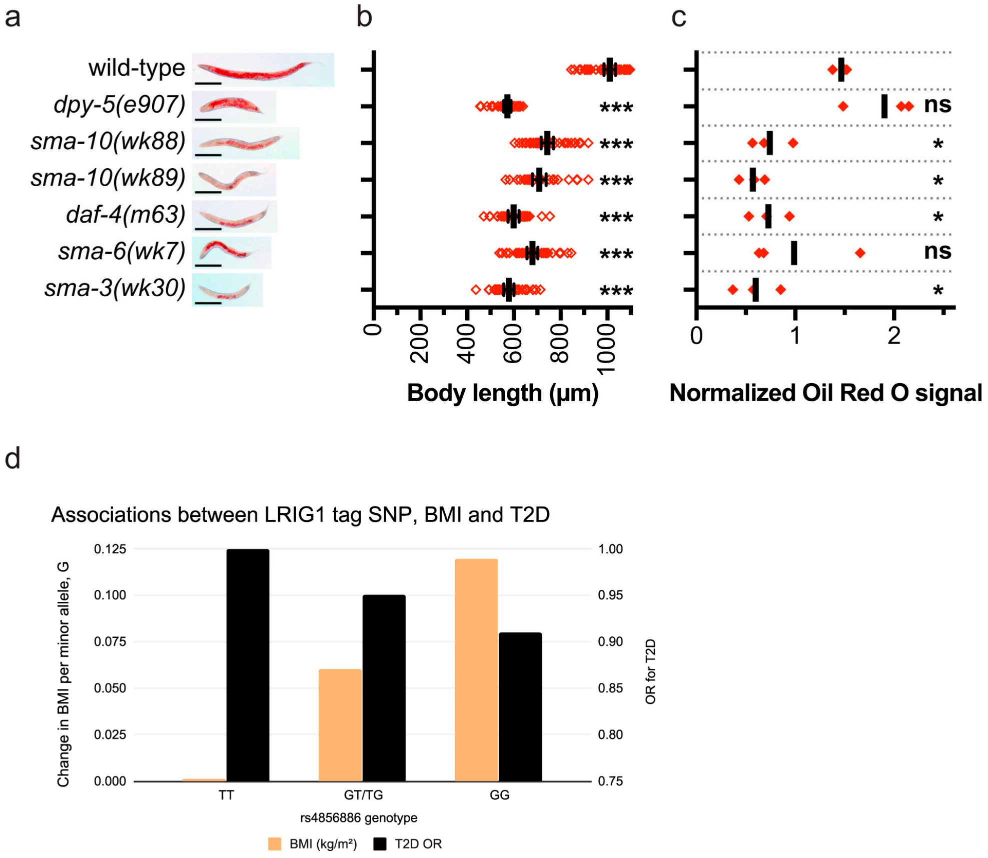


Figure 4

Panel **a:** Representative whole-body images of Oil Red O-stained adult hermaphrodite worms. Scale bars, 200 µm. **b:** Adult body lengths of wild-type animals (n = 46), dpy-5(e907) (n = 37), sma-10(wk88) (n = 41), sma-10(wk89) (n = 32), daf-4(m63) (n = 32), sma-6(wk7) (n = 41), and sma-3(wk30) (n = 37). The body length of each individual animal is plotted as a red square. Solid lines and error bars indicate the means and 95% confidence intervals, respectively. The order of the dot plots, from top-to-bottom, is the same as that for the images in **a**. **c**: Oil Red O signal intensities from three independent experiments. Each experiment was normalized to its combined mean signal intensity across all genotypes. For each genotype, solid, red squares indicate the mean normalized signal in each independent experiment. Solid lines indicate the combined means from three experiments. The order of the dot plots, from top-to-bottom, is the same as those for the images in a. Statistical significance versus wild-type was determined with multiplicity-adjusted P-values, calculated using Holm-Sidak multiple comparisons tests. *P < 0.05, ***P < 0.001. **d**: Plot illustrating the difference in predicted BMI (least square means, LSMs) across the genotypes of rs4856886 (minor allele = G, major allele = T) and odds ratio for type 2 diabetes (T2D) across genotypes. The x-axis represents the rs4856886 genotypes compared. The y-axis on the left represents the difference in BMI LSMs per single minor allele across the genotypes, while the y-axis on the right represents the odds ratios for T2D. In this study, the minor allele was associated with an increase in BMI and a lower odds of T2D risk.

Paper 2 discussion

In this integrated study, we found that BMP signalling was the main mechanism via which *LRIG* proteins regulate lipid metabolism, differentiation of adipocytes in MEFs and lipid storage in *C. elegans*. In population-level analyses, human *LRIG1* gene variants were associated with increased BMI, decreased risk of T2D and adipocyte size. Taken together, these results show that *LRIG* proteins regulate BMP signalling and lipid metabolism, and are implicated in metabolism of lipids in humans.

These results may indicate a metabolically favourable adipose tissue phenotype characterised by hyperplastic adipose morphology and low triglyceride levels. It is possible that enhanced BMP signalling mediates *LRIG1*-associated adipogenesis. Hyperplastic adipose tissue has numerous metabolically active, hence energy-efficient adipocytes. This energy efficiency attributed to a high number of adipocytes could explain the T2D protective effect we observed in this study. This is among the first studies to describe these associations and therefore offers opportunity for further investigations of *LRIG1* gene variants in human energy metabolism.

Paper 3

In this study, we developed and evaluated the performance of different models in predicting NAFLD. Different models contained different features which are summarised in table 3. Briefly, the variables included clinical/ anthropometric variables, and omics (transcriptomic, proteomic, genetic and metabolomic). Prediction models were constructed based on variables commonly available within clinical settings and also those not routinely available (Table 3). Models 1-3 included clinically accessible variables that are also known to be associated with NAFLD and these were not subjected to dimension reduction. In model 4, the most accessible of highly correlated variables ($r > 0.8$, Pearson's) were selected from the combined IMI-DIRECT cohorts 1 and 2. For high dimensional omics data, dimensionality reduction and feature selection was performed using LASSO in a 70:30 train-test ratio with a 10-fold cross-validation, before using the data for model construction. Prior to LASSO reduction, a GWAS for liver fat was performed on the genetic data and associated SNPs ($p < 5 \times 10^{-6}$) selected. With the 30% test data, prediction models for fatty liver were developed based on features for each respective model, using random forest algorithm. Model evaluation was performed by computing the receiver operating characteristic area under the curve (ROC-AUC) for each model and comparing the results. Different cut-off values for classification and their impact on the model performance evaluation measures were explored. Finally, we also compared our models with existing liver fat calculation formulae namely the fat liver index (FLI), hepatic steatosis index (HSI) and the NAFLD liver fat score (NAFLD-LFS). Analyses were performed on the combined, no diabetes (cohort 1) only and diabetes only (cohort 2) cohorts in the IMI-DIRECT consortium. Models 1 and 2 were then externally validated in the UKB cohort.

Table 2

Features used in constructing different prediction models

Feature	Model	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
BMI (kg/m^2)		×	×	×	×		×		×		×		×		×		×		×
Waist (nearest cm)		×	×	×	×		×		×		×		×		×		×		×
SBP (mm Hg)		×			×		×		×		×		×		×		×		×
DBP (mm Hg)		×			×		×		×		×		×		×		×		×
Alcohol intake ("never", "occasionally", "regularly")		×	×	×	×		×		×		×		×		×		×		×
Diabetes status ("Non-diabetes", "diabetes")		×	×	×	×		×		×		×		×		×		×		×
Triglyceride (mmol/L)			×	×	×		×		×		×		×		×		×		×
ALT(U/L)			×	×	×		×		×		×		×		×		×		×
AST (U/L)			×	×	×		×		×		×		×		×		×		×
HbA1c (mmol/mol)			×	×	×		×		×		×		×		×		×		×
Fasting glucose (mmol/L)			×	×	×		×		×		×		×		×		×		×
Fasting insulin (pmol/L)				×	×		×		×		×		×		×		×		×
2hr glucose (mmol/L)					×		×		×		×		×		×		×		×
2hr insulin (pmol/L)					×		×		×		×		×		×		×		×
GGTP (U/L)					×		×		×		×		×		×		×		×
HDL (mmol/L)					×		×		×		×		×		×		×		×
BasalISR (insulin secretion at the beginning of the OGTT/MMTT)					×		×		×		×		×		×		×		×
OGIS (ml×min^(-1)×m^(-2))					×		×		×		×		×		×		×		×
Clins (mean insulin secretion)/(mean insulin concentration)					×		×		×		×		×		×		×		×
Glucagonmin0 (pg/ml)					×		×		×		×		×		×		×		×
TotGLP1min0 (pg/ml)					×		×		×		×		×		×		×		×
PA_intensity_mean (Mean high-pass filtered vector magnitude)					×		×		×		×		×		×		×		×
Genetics						×	×							×	×				
Transcriptomics								×	×					×	×				
Exploratory proteomics										×	×			×	×				
Targeted metabolomics												×	×	×	×				
Targeted proteomics																×	×		
Untargeted metabolomics																		×	×

In clinical models 1-3, model 3 had the best ability to predict NAFLD in the combined and non-diabetic cohorts (ROC-AUC = 0.82, 95% CI 0.81, 0.83; p < 0.001) but there was no difference when compared to the NAFLD-LS model in the diabetic cohort. Clinical model 1 performed below all other models. Figure 5 shows a summary of the ROC-AUCs and 95% CI of the different models per cohort.

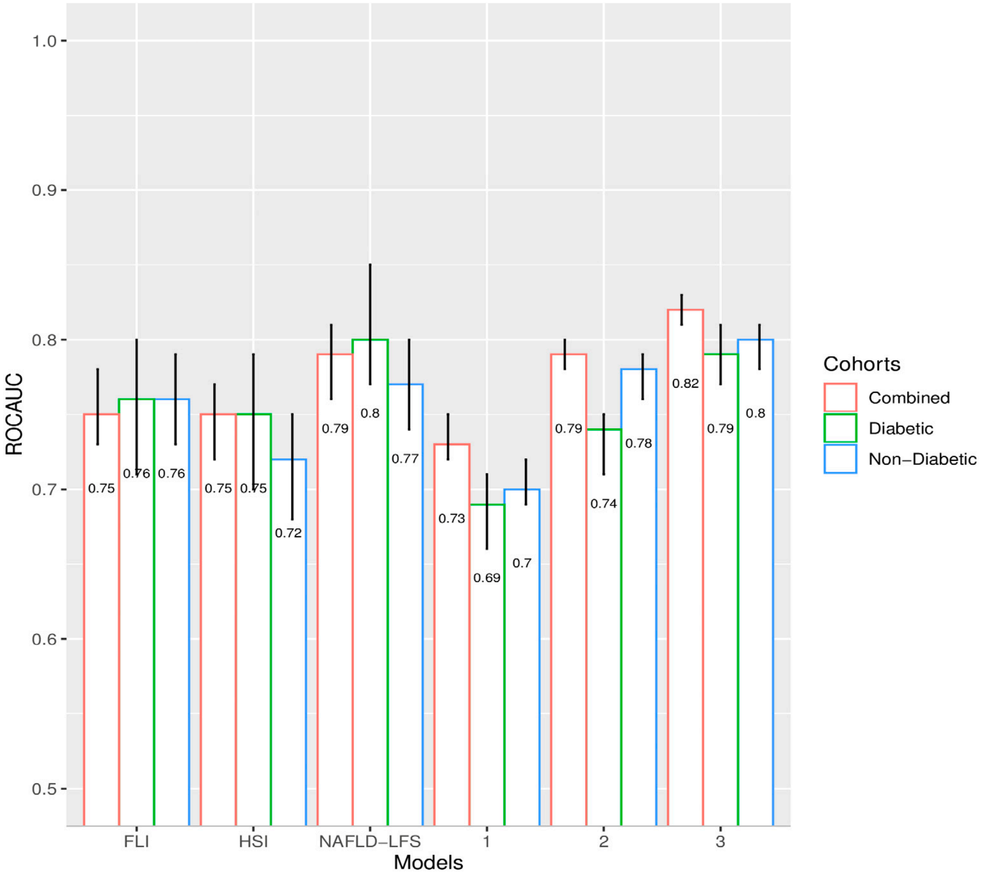


Figure 5
Model performance (ROC-AUC with 95%) for clinical models 1-3 and the FLI, HIS and NAFLD-LFS.

Table 3 shows the optimum values for each performance metric for the different models including results for validation of models 1 and 2, the FLI and HIS in the UKB.

Table 3
Performance metrics of different models under optimum cut-off points

Non-diabetes (IMI-DIRECT)	Cuttoff	Sensitivity	Specificity	F1 score	Balanced accuracy
Model 1	0.4	0.51	0.75	0.51	0.63
Model 2	0.4	0.60	0.79	0.59	0.69
Model 3	0.4	0.64	0.80	0.63	0.72
FLI	60	0.89	0.41	0.58	0.65
HSI	36	0.62	0.68	0.55	0.65
NAFLD-LFS	-0.64	1	0.04	0.51	0.52
Diabetes (IMI DIRECT)					
Model 1	0.6	0.63	0.64	0.67	0.64
Model 2	0.6	0.65	0.68	0.69	0.67
Model 3	0.6	0.69	0.75	0.74	0.72
FLI	60	0.77	0.54	0.73	0.66
HSI	36	0.83	0.48	0.75	0.65
NAFLD-LFS	-0.64	1	0.01	0.73	0.50
Combined (IMI DIRECT)					
Model 1	0.4	0.67	0.65	0.62	0.66
Model 2	0.4	0.72	0.69	0.67	0.71
Model 3	0.4	0.74	0.73	0.70	0.74
FLI	60	0.84	0.44	0.64	0.64
HSI	36	0.71	0.63	0.64	0.67
NAFLD-LFS	-0.64	1	0	0.58	0.50
UK Biobank					
Model 1	0.4	0.49	0.78	0.43	0.63
Model 2	0.4	0.67	0.74	0.52	0.71
FLI	60	0.62	0.76	0.50	0.69
HSI	36	0.66	0.72	0.50	0.69

Model 4, which had the highest number of clinical variables yielded a ROC-AUC of 0.79 (95% CI 0.76, 0.81; $p < 0.001$). Models that contained omics data only had poorer predictive ability than the clinical or combined models i.e., clinical with omics data. Model 14 which had clinical and all omics combined, had the highest prediction ability (ROC-AUC = 0.84, 95% CI 0.82, 0.86, $p < 0.001$). Figure 6 shows ROC-AUCs of the omics and combined models.

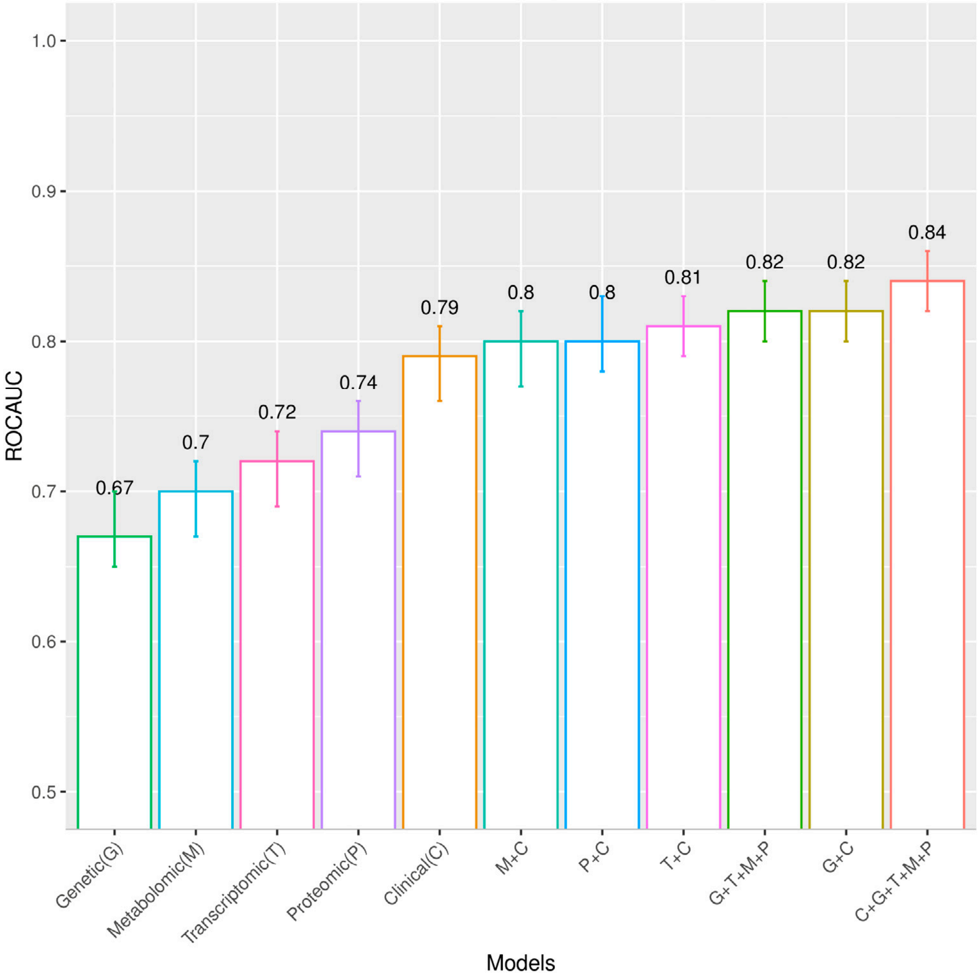


Figure 6
Predictive performance of (ROC-AUC and 95% CI) of clinical models and omics separately or in combination with the clinical model in the imi combined cohort. Clinical (C), model 4, with the 22 selected clinical variables. Genetic (G), model 5, with 23 SNPs. C+G, model 6, with clinical plus genetic variables. Transcriptomic (T), model 7, with 93 protein-coding genes. T+C, model 8, with transcriptomic plus clinical variables. Proteomic (P), model 9, with 22 proteins from exploratory proteomics. P+C, model 10, with proteomic plus clinical variables. Metabolomic (M), model 11, with 25 metabolites from targeted metabolomics. M+C, model 12, with metabolomic plus clinical variables. G+T+M+P, model 13, with all omics together. C+G+T+M+P, model 14, with all the omics combined with the clinical model.

Paper 3 discussion

In this study, we developed 18 prediction models for NAFLD and where the data allowed (model 1 and 2) validated the models in the UKB. Inclusion of HbA_{1c} or fasting glucose and fasting insulin improved the prediction ability of a model that had basic clinical variables and was also better than existing models in the combined cohort. In this study the FLI[141] which is a commonly used prediction model yielded similar performance in both cohorts in DIRECT (ROC ≈ 0.75) but was not superior to clinical model 3 (ROC = 82), and whis was no better than common anthropometric measures in predicting NAFLD in a comparative study[142]. The FLI, HIS and NAFLD-LFS use different parameters, with little overlap, but their discriminative abilities of NAFLD have been shown to be comparable[143]. In studies that deployed ML on omics data (different studies used different omics and different combinations), multi-domain models (models with both clinical and omics data) performed best in predicting NAFLD[144,145].

Given that imaging is not always present, is unable to distinguish who will progress beyond NALFD, and no screening recommendations are in place, prediction models offer a cheap and simple alternative that can be useful in many settings. As big data becomes more available with increase in sample size and computing power, there is opportunity for developing more enhanced prediction models, identifying new features that can be used as prediction or diagnostic biomarkers, or gene variants that can be further elucidated to identify intervention/therapeutic targets.

Paper 4

In this study, I set out to investigate causal associations between adiposity (assessed as BMI) and cardiometabolic outcomes, the nature of these causal relationships and any sex differences within the causal framework.

I used TSLS MR to estimate the causal relationship between BMI and each of the outcomes and stratified the analyses by sex and age to determine if there were any differences. To determine differences between men and women, I computed the Cochran's Q statistic for each outcome and also investigated whether BMI was causal of *any* cardiometabolic disease (T2D, CAD, hypertension, stroke, CKD) by combining results for each disease outcomes in a meta-analysis. To determine the nature of the causal relationships, I performed NLMR using both fractional polynomials and piecewise linear MR, in combined and sex-stratified analyses. Additional sensitivity analyses were performed as follows: excluding outliers of BMI; adjusting for lipid lowering medication and WHR; stratifying by menopause status in women; and applying other two different methods (G-estimator and TSLS with residual inclusion) to compute causal estimates.

Characteristics of participants used in this paper are shown in table 4 below. Figure 7 shows the prevalence of main cardiometabolic diseases across different ages in the cohort

Table 4
Participant characteristics, UKBB, N = 387,394

Characteristic	Men	Women
Count	45.9%	54.1%
Age (years)	57.1 (8.09)	56.7 (7.92)
BMI (kg/m^2)	27.8 (4.23)	27.0 (5.13)
TDI	-1.59 (2.89)	-1.53 (2.99)
Smoking status		
Never	41.2%	58.8%
Previous	51.1%	48.9%
Current	53.7%	46.3%
Alcohol intake status		
Never	24.7%	75.3%
Previous	43.0%	57.0%
Current	46.8%	53.2%
SBP (mmHg)	145 (19.4)	138 (21.2)
DBP (mmHg)	86.6 (11.0)	82.4 (11.1)
CAD	67.4%	32.6%
Type 2 diabetes	61.5%	38.5%
Stroke	61.3%	38.7%
CKD	55.2%	44.8%
Mortality	60%	40%
Glucose (mmol/L)	5.18 (1.37)	5.06 (1.04)
Hba$_{1c}$ (mmol/mol)	36.3 (7.29)	35.7 (5.70)
HbA$_{1c}$ (%)	6.08 (1.89)	6.04 (1.66)
Cholesterol (mmol/L)	5.49 (1.13)	5.90 (1.13)
HDL (mmol/L)	1.28 (0.311)	1.60 (0.377)
LDL (mmol/L)	3.48 (0.862)	3.64 (0.872)
Triglycerides (mmol/L)	1.98 (1.15)	1.57 (0.861)
Urea (mmol/L)	5.64 (1.44)	5.26 (1.32)

Continuous variables are presented as mean (SD) and categorical variables as percentage. TDI = Townsend Deprivation Index, CAD = Coronary artery disease, CKD = Chronic kidney disease.

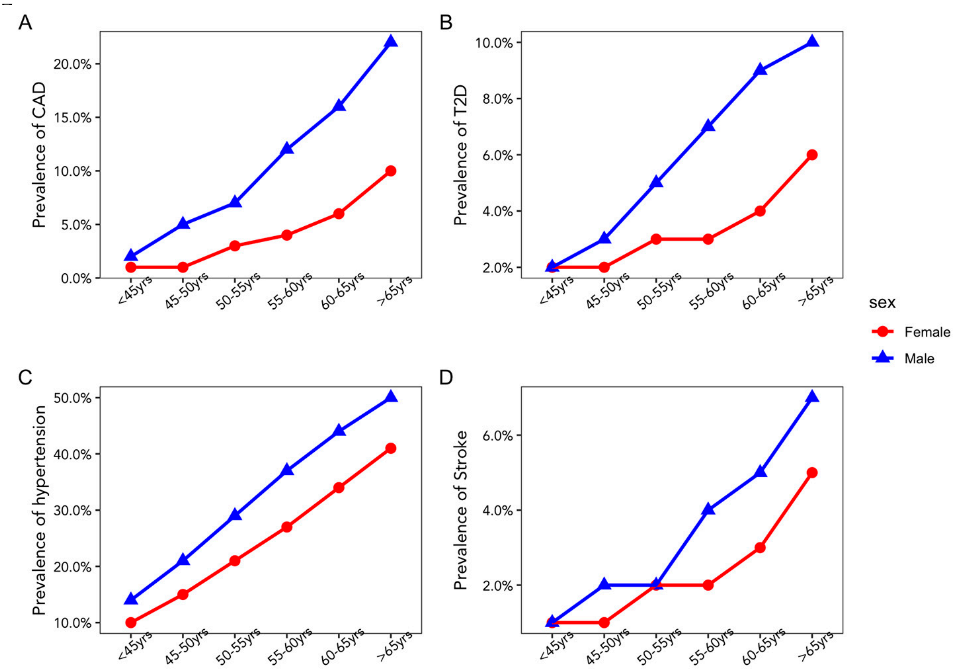


Figure 7
Prevalence of cardiometabolic disease in the UKB across different age groups

From analyses estimating causal effects, BMI was significantly associated with T2D, hypertension and CAD but not CKD or stroke. The largest effect was observed for T2D. In estimates of causal effect of BMI on cardiometabolic biomarkers and BP (coefficients expressed in SD units), only LPA levels failed to show a significantly association. BMI was causally associated with elevated levels of glycaemia biomarkers (glucose and HbA_{1c}) and triglycerides but decreased total cholesterol, LDL and HDL cholesterol. BMI was also causally associated with elevations in blood pressure with effect on diastolic blood pressure being almost double the effect on systolic blood pressure. These results remained unchanged when causal estimates were computed with other methods. Table 5 below shows details of the effect sizes in the main method and two other methods.

Table 5.
Estimates of causal relationships between BMI and cardiometabolic outcomes, comparing different methods, in the UKB

	METHOD		
	TSLS	TSLS-RI	G estimator
TRAIT	OR (95% CI)	OR (95% CI)	OR (95% CI)
CAD	1.20(1.08,1.33)	1.19(1.07,1.32)	1.19(1.07,1.32)
T2D	3.10(2.73,3.53)	3.05(2.67,3.48)	3.08(2.68,3.55)
Hypertension	1.53(1.44,1.62)	1.54(1.45,1.64)	1.55(1.45,1.64)
Stroke	1.08(0.92,1.28)	1.08(0.92,1.27)	1.08(0.92,1.27)
CKD	1.08(0.67,1.72)	1.07(0.67,1.71)	1.07(0.68,1.70)
	Beta (95% CI)	Beta (95% CI)	Beta (95% CI)
Glucose	0.16(0.13,0.20)	0.16(0.13,0.20)	0.18(0.12,0.25)
HBA1c	0.22(0.19,0.26)	0.22(0.19,0.26)	1.43(1.14,1.72)
HDL	-0.26(-0.3,-0.22)	-0.26(-0.3,-0.22)	-0.10(-0.12,-0.08)
LDL	-0.10(-0.14,-0.07)	-0.10(-0.14,-0.07)	-0.09(-0.13,-0.06)
Triglycerides	0.13(0.09,0.16)	0.13(0.09,0.16)	0.13(0.10,0.16)
LPA	0.02(-0.02,0.05)	0.02(-0.02,0.05)	0.75(-0.99,2.50)
DBP	0.15(0.12,0.19)	0.16(0.12,0.19)	1.60(0.82,2.39)
SBP	0.09(0.06,0.12)	0.09(0.06,0.12)	1.63(0.29,2.97)
Urea	0.05(0.01,0.08)	0.05(0.01,0.08)	0.07(0.01,0.12)
Cholesterol	-0.18(-0.21,-0.14)	-0.18(-0.21,-0.14)	-0.21(-0.25,-0.16)

In the combined meta-analysis, BMI was significantly causally associated with increased causal odds of *any* cardiometabolic disease. Figure shows a forest plot of the combined results.

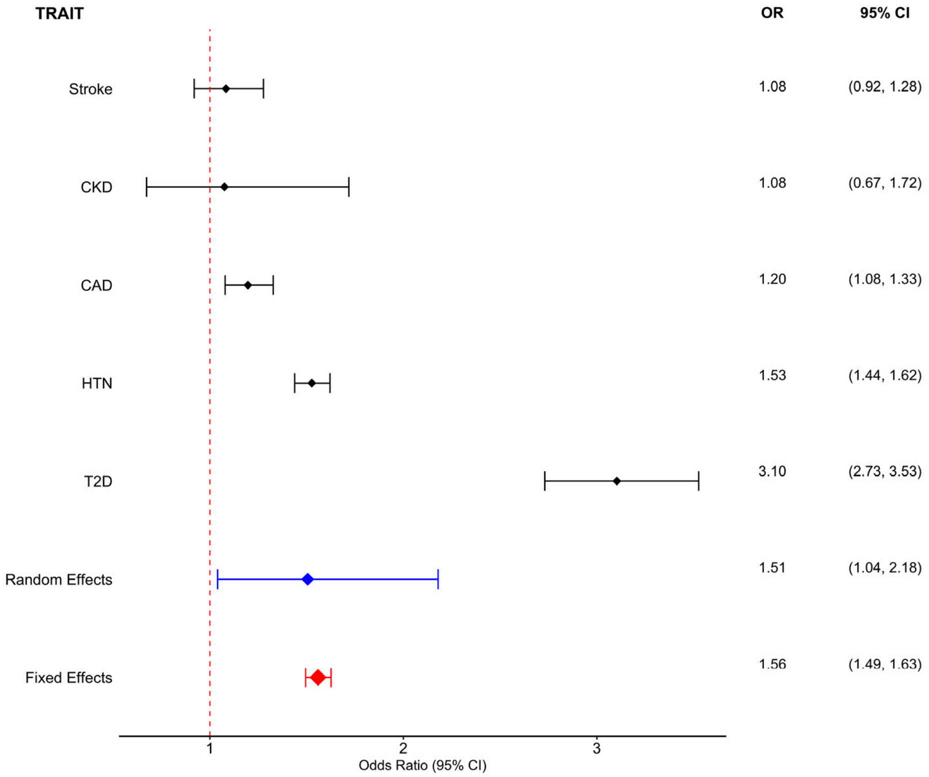


Figure 8
Forest plot of meta-analysis depicting causal risk of BMI on any cardiometabolic disease

When analyses were stratified by sex, BMI was not associated with CAD in women (OR = 0.97, 95% CI 0.82, 1.15, p = 0.69) but in men the association remained positive and significant (OR = 1.33, 95% CI 1.19, 1.49, p = 4.29 x10^{-7}). The P value for difference = 0.01, however this was not significant after accounting for multiple testing, p < 0.001. BMI was associated with CAD in pre-menopausal women (or younger women <55 years) but not older women (Figure 9). The causal effects of BMI on variation in LDL and total cholesterol levels were also different between men and women, with the LDL difference persisting after multiple testing.

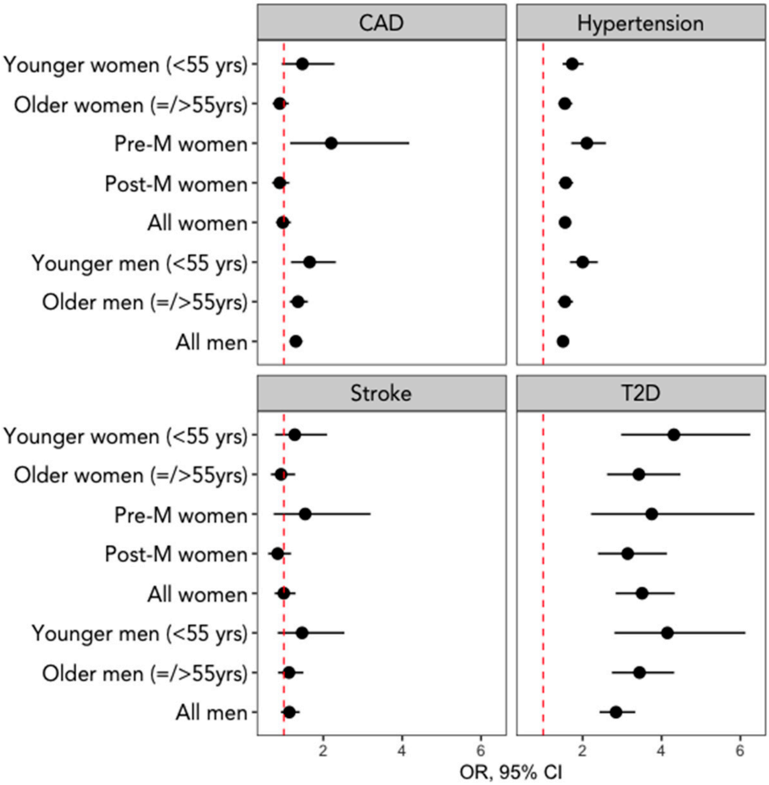


Figure 9
Causal risk of BMI on cardiometabolic diseases in different groups. Pre-M = premenopausal, Post-M = postmenopausal

From the NLMR, generally, the causal relationship between BMI and cardiometabolic outcomes was non-linear, for CAD and T2D and all risk factor biomarkers and BP when using fractional polynomials, and the results were supported by piecewise MR results. Results did not materially change when BMI outliers were excluded or between men and women, especially after triangulating the evidence for non-linearity.

See table 6 and 7, and figure 10-15. Figures 10-12 represent plots with BMI values truncated at 40 kg/m^2 and effect sizes restricted to 0 to 3 for disease outcomes, and -2 to 2 for biomarkers and BP. This is done to make the plots comparable. Figures 13–15 represent the untruncated plots.

Table 6
Tests for shapes of causal relationships between BMI and cardiometabolic phenotypes in the UKBB

	Non-linearity tests				Heterogeneity tests	
Trait	$P_{\text{FP degree}}$	$P_{\text{FP non-linearity}}$	$P_{\text{Quadratic}}$	P_{CochranQ}	P_{CochranQ}	P_{Trend}
CAD	0.18	0.10	0.02	0.36	0.03	0.53
T2D	0.28	0.01	$4.25x10^{-3}$	0.60	$3.07\ x10^{-3}$	0.30
HTN	0.15	1.00	0.83	0.73	0.05	0.41
Stroke	0.99	0.74	0.79	0.10	0.04	0.24
CKD	0.26	0.44	0.50	0.51	0.02	0.05
GLU	$1.96x10^{-2}$	$2.16x10^{-4}$	$2.17x10^{-5}$	0.24	0.13	0.46
HBA$_{1c}$	$3.27x10^{-3}$	$7.25x10^{-8}$	$9.54x10^{-10}$	$1.38x10^{-4}$	0.14	$1.36x10^{-2}$
HDL	$2.54x10^{-2}$	$1.82x10^{-6}$	$7.04x10^{-8}$	$2.19x10^{-2}$	0.10	0.11
LDLD	$9.00x10^{-9}$	$2.78x10^{-5}$	$2.56x10^{-13}$	$7.94x10^{-4}$	0.28	0.12
TG	$8.76x10^{-8}$	$4.07x10^{-5}$	$2.36x10^{-9}$	$2.61x10^{-5}$	0.15	0.73
LPA	0.97	0.47	$0.71x10^{-1}$	0.74	0.27	0.10
DBP	0.38	$9.12x10^{-3}$	$3.33x10^{-3}$	0.24	$4.53x10^{-2}$	$6.21x10^{-2}$
SBP	0.97	$3.62x10^{-2}$	$4.12x10^{-2}$	0.48	$5.61x10^{-2}$	$8.30x10^{-2}$
urea	0.50	$2.28x10^{-3}$	$3.33x10^{-3}$	0.48	0.26	0.62
cholesterol	$6.75x10^{-3}$	$3.42x10^{-8}$	$2.50x10^{-10}$	$5.68x10^{-5}$	0.77	0.84

HTN = hypertension, GLU = glucose, LDLD = LDL cholesterol, HTN = hypertension, LPA = lipoprotein(a), DBP = diastolic blood pressure, SBP = systolic blood pressure

Table 7

Tests for sex stratified shapes of causal relationships between BMI and cardiometabolic phenotypes in the UKBB

	Non-linearity tests								Heterogeneity tests			
	$P_{FP\ degree}$		$P_{FP\ non\text{-}linearity}$		$P_{Quadratic}$		$P_{CochranQ}$		$P_{CochranQ}$		P_{Trend}	
Trait	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women
CAD	0.83	0.44	0.10	0.43	0.08	0.28	0.44	0.00	0.08	0.32	0.26	0.03
T2D	0.21	0.10	0.05	0.17	0.02	0.08	0.73	0.02	0.87	0.01	0.40	0.78
HTN	0.27	0.20	0.23	0.08	0.10	0.08	0.21	0.03	0.04	0.64	0.56	0.04
Stroke	0.85	0.54	0.44	0.70	0.34	0.55	0.75	0.51	0.10	0.38	0.40	0.31
CKD	0.06	1.00	0.01	$4.68x10^{-03}$	0.06	0.01	0.00	0.00	0.49	0.19	0.06	0.56
GLU	0.01	0.82	$2.34x10^{-03}$	0.25	$1.32x10^{-03}$	0.22	0.39	0.67	0.38	0.41	0.07	0.63
HBA1c	0.01	0.33	$2.77x10^{-05}$	$9.13x10^{-04}$	$3.69x10^{-07}$	$3.11x10^{-04}$	0.04	0.00	0.06	0.07	0.72	0.40
HDL	0.62	0.01	$3.02x10^{-05}$	$2.31x10^{-03}$	$2.45x10^{-05}$	$1.62x10^{-04}$	0.02	0.19	0.74	0.18	0.69	0.79
LDLD	$4.52x10^{-03}$	$2.64x10^{-04}$	$2.85x10^{-03}$	$9.09x10^{-03}$	$4.41x10^{-06}$	$1.91x10^{-06}$	0.09	0.11	0.27	0.21	0.93	0.49
TG	$8.05x10^{-04}$	$1.24x10^{-03}$	$1.16x10^{-03}$	0.05	$3.34x10^{-07}$	$2.36x10^{-03}$	0.12	0.34	0.72	0.71	0.18	0.04
LPA	0.16	0.86	0.55	0.58	$7.37x10^{-02}$	0.51	0.75	0.74	0.84	0.42	0.92	0.85
DBP	0.61	0.34	$1.36x10^{-03}$	0.46	$1.12x10^{-03}$	0.31	0.24	0.40	0.04	$7.61x10^{-04}$	0.99	0.35
SBP	0.96	0.90	$9.36x10^{-03}$	0.60	$2.25x10^{-02}$	0.56	0.48	0.22	0.03	$1.39x10^{-03}$	0.91	0.48
urea	0.05	0.75	$4.23x10^{-02}$	0.23	$1.43x10^{-03}$	0.43	0.13	0.96	0.02	0.08	0.84	0.67
cholesterol	0.09	0.02	$1.86x10^{-04}$	$1.55x10^{-03}$	$4.95x10^{-05}$	$3.47x10^{-05}$	0.01	0.228	0.31	0.179	0.05	0.708

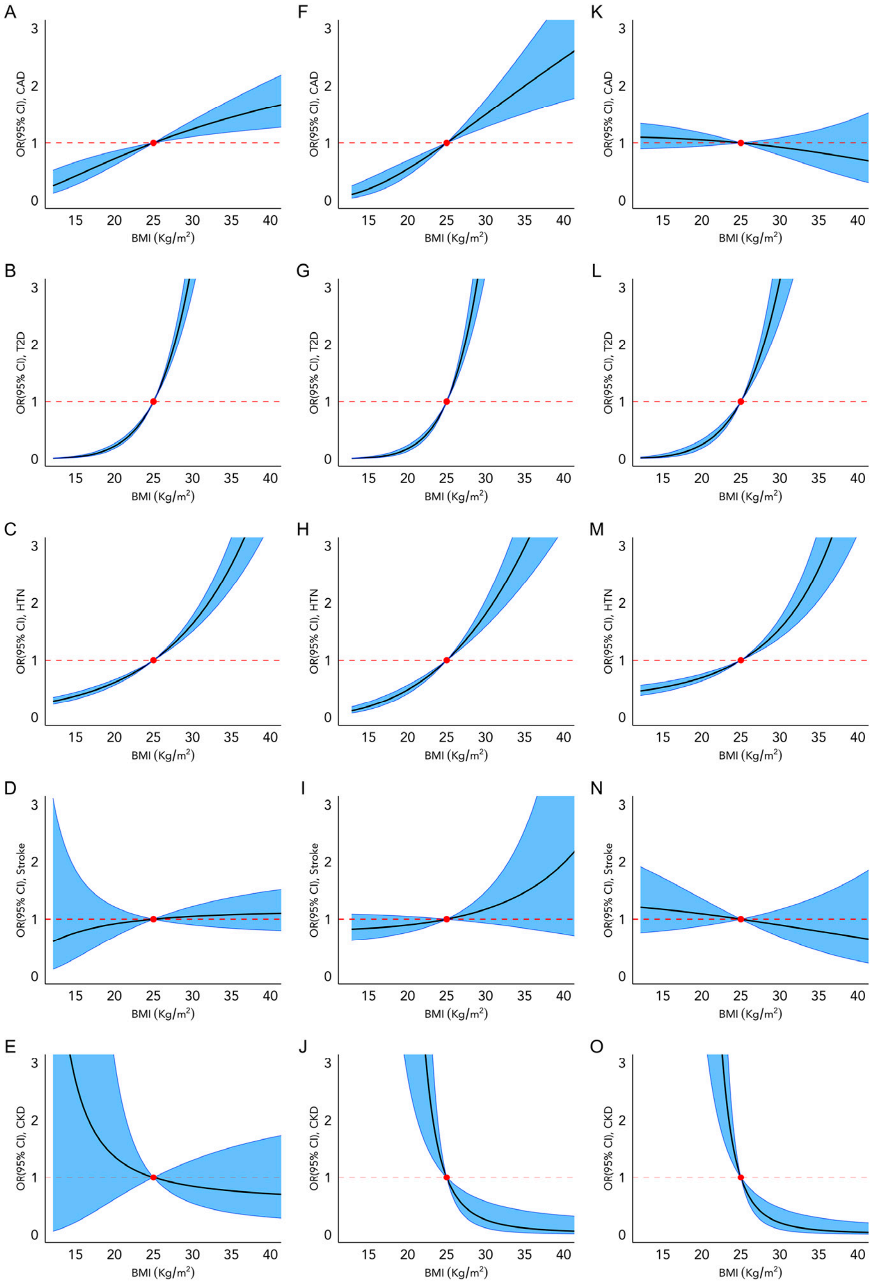


Figure 10
Plots showing estimated shape of the causal relationships between BMI and cardiometabolic diseases in combined analyses (panels A to E), men (panels F to J) and women (panels K to O). Shape estimates are derived from the function of fractional polynomials that best fits the data. Solid black line represents the function curve, blue band represents 95% CI, red dot represents reference BMI of 25kg/m2 and the dashed line represents the null effect size. The plots have been zoomed to depict estimated causal associations for BMI up to 40 Kg/m2 and OR up to 3.0 for ease of comparison. Figure 13 shows the untrancated version.

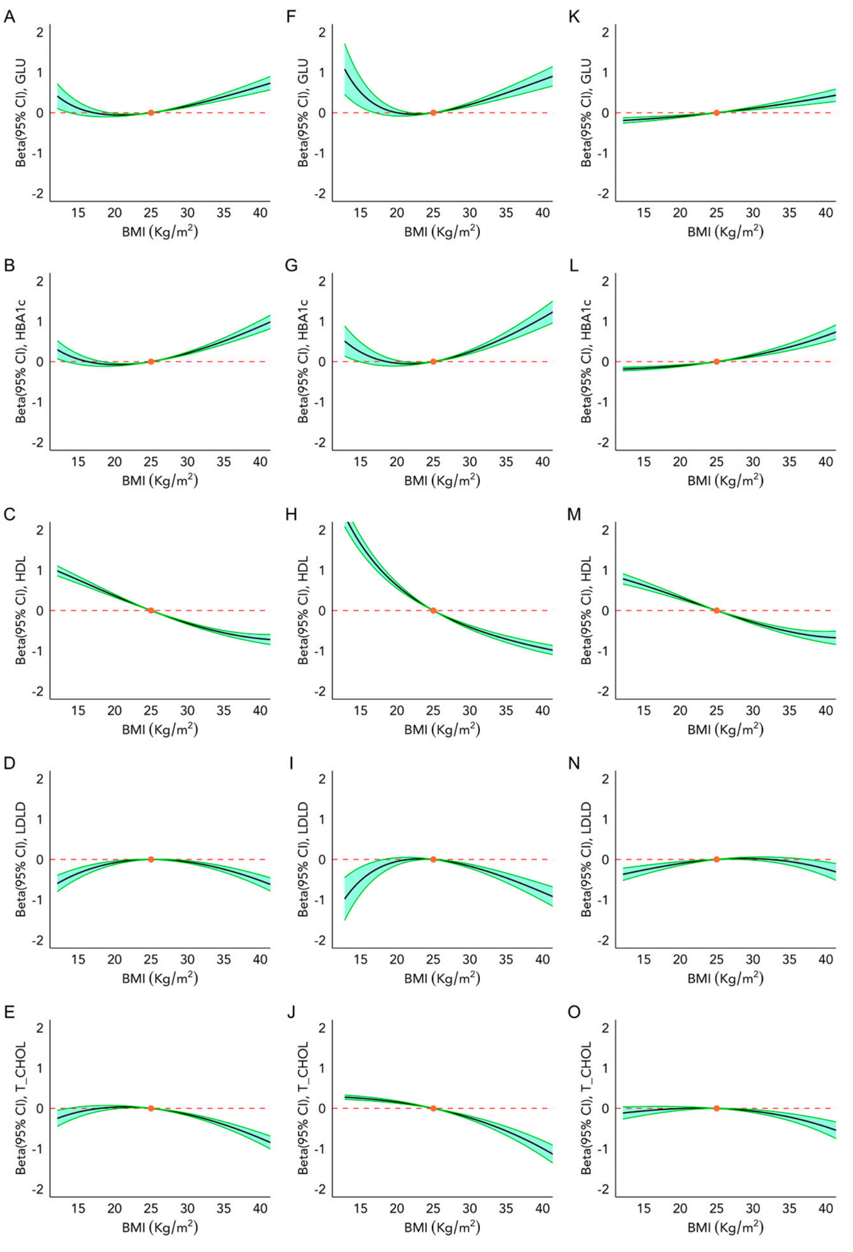


Figure 11
Plots showing estimated shape of the causal relationships between BMI and cardiometabolic biomarkers in combined analyses (panels A to E), men (panels F to J) and women (panels K to O). Shape estimates are derived from the function of fractional polynomials that best fits the data. Solid black line represents the function curve, green band represents 95% CI, red dot represents reference BMI of 25kg/m2 and the dashed line represents the null effect size. The plots have been zoomed to depict estimated causal associations for BMI up to 40 Kg/m2 and β between -2 and 2 for ease of comparison. Figure 14 shows the untruncated version. GLU = glucose, T_CHOL = total cholesterol.

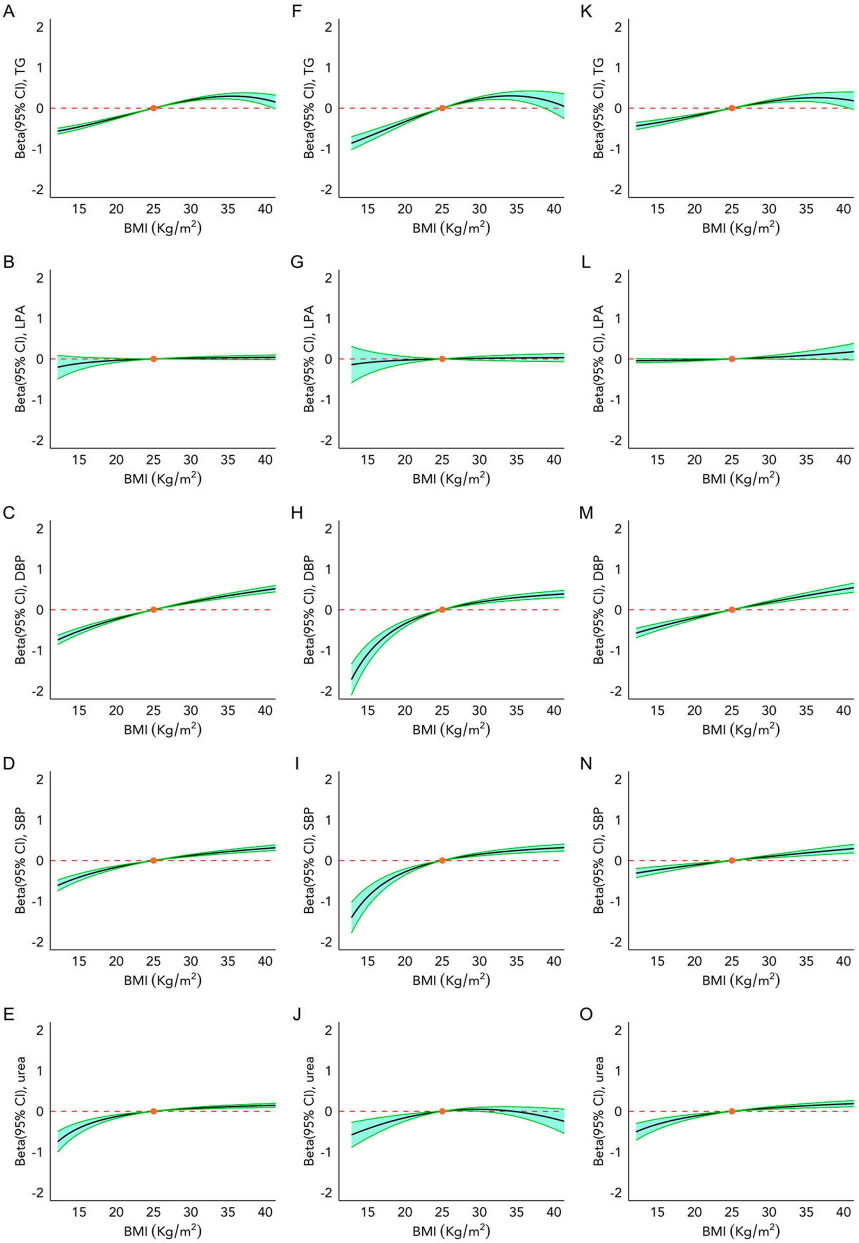


Figure 12

Plots showing estimated shape of the causal relationships between BMI and cardiometabolic biomarkers in combined analyses (panels A to E), men (panels F to J) and women (panels K to O). Shape estimates are derived from the function of fractional polynomials that best fits the data. Solid black line represents the function curve, green band represents 95% CI, red dot represents reference BMI of 25kg/m2 and the dashed line represents the null effect size. The plots have been zoomed to depict estimated causal associations for BMI up to 40 Kg/m2 and β between -2 and 2 for ease of comparison. Figure 15 shows the untruncated version. GLU = glucose, T_CHOL = total cholesterol.

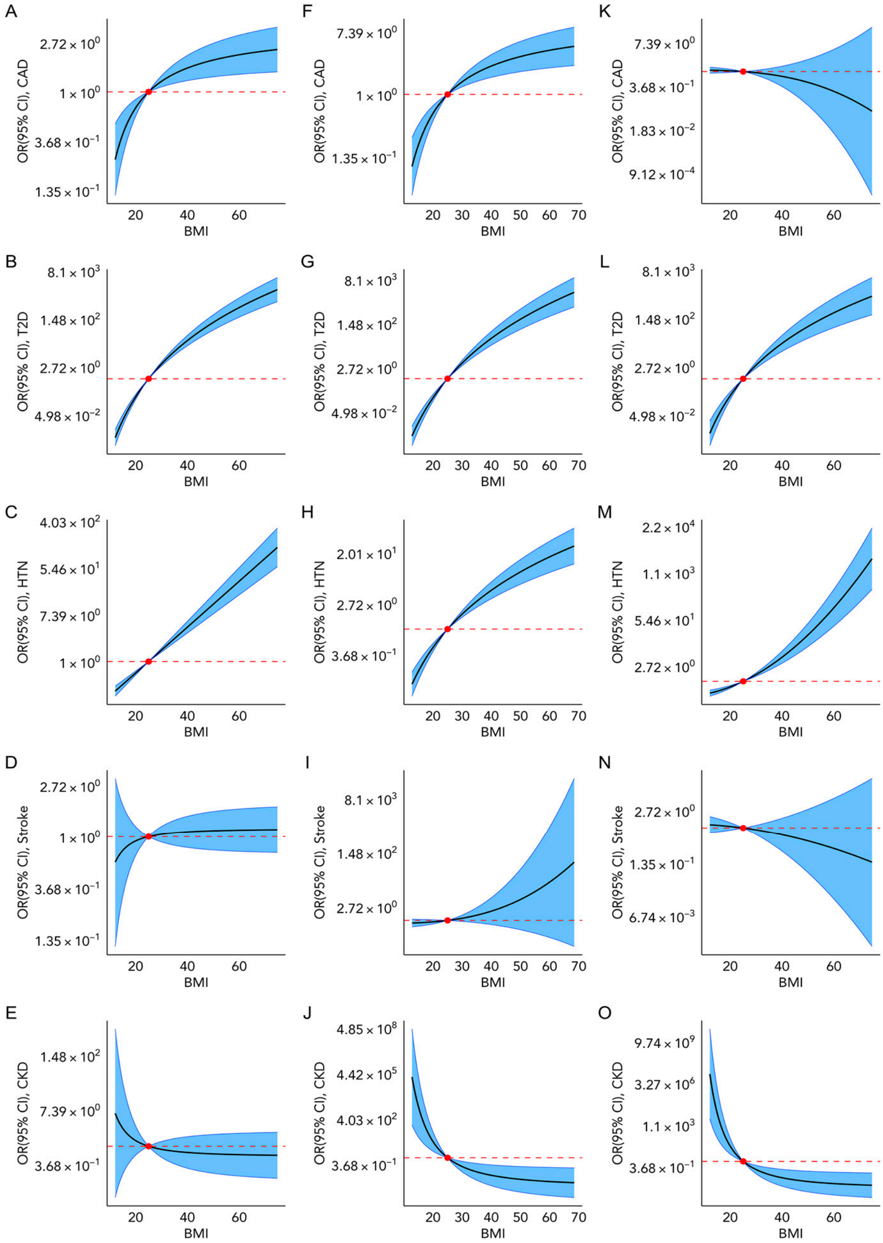


Figure 13

Untruncated plots showing estimated shape of the causal relationships between BMI and cardiometabolic diseases in combined analyses (panels A to E), men (panels F to J) and women (panels K to O). Shape estimates are derived from the function of fractional polynomials that best fits the data. Solid black line represents the function curve, blue band represents 95% CI, red dot represents reference BMI of 25kg/m2 and the dashed line represents the null effect size.

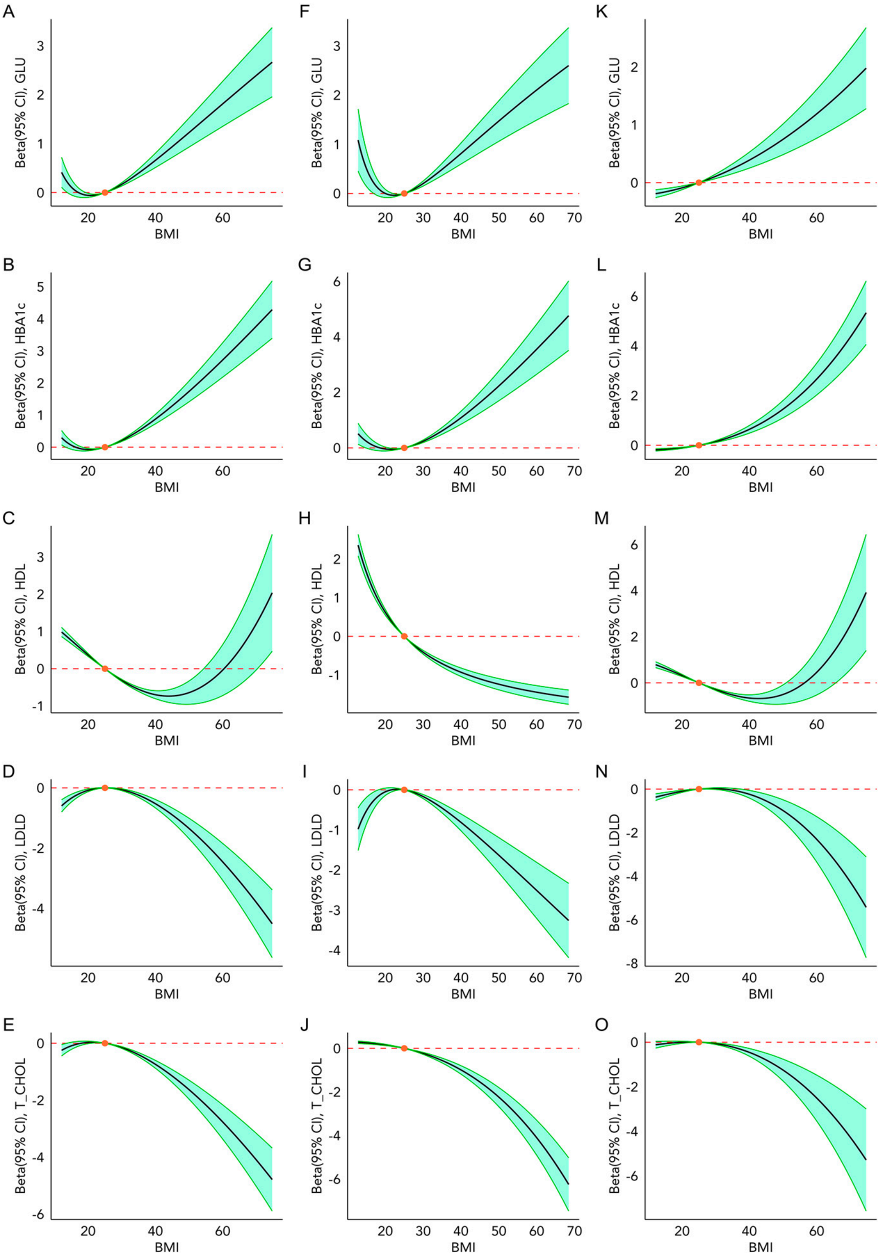


Figure 14

Untruncated plots showing estimated shape of the causal relationships between BMI and cardiometabolic biomarkers in combined analyses (panels A to E), men (panels F to J) and women (panels K to O). Shape estimates are derived from the function of fractional polynomials that best fits the data. Solid black line represents the function curve, green band represents 95% CI, red dot represents reference BMI of 25kg/m2 and the dashed line represents the null effect size. GLU = glucose, T_CHOL = total cholesterol.

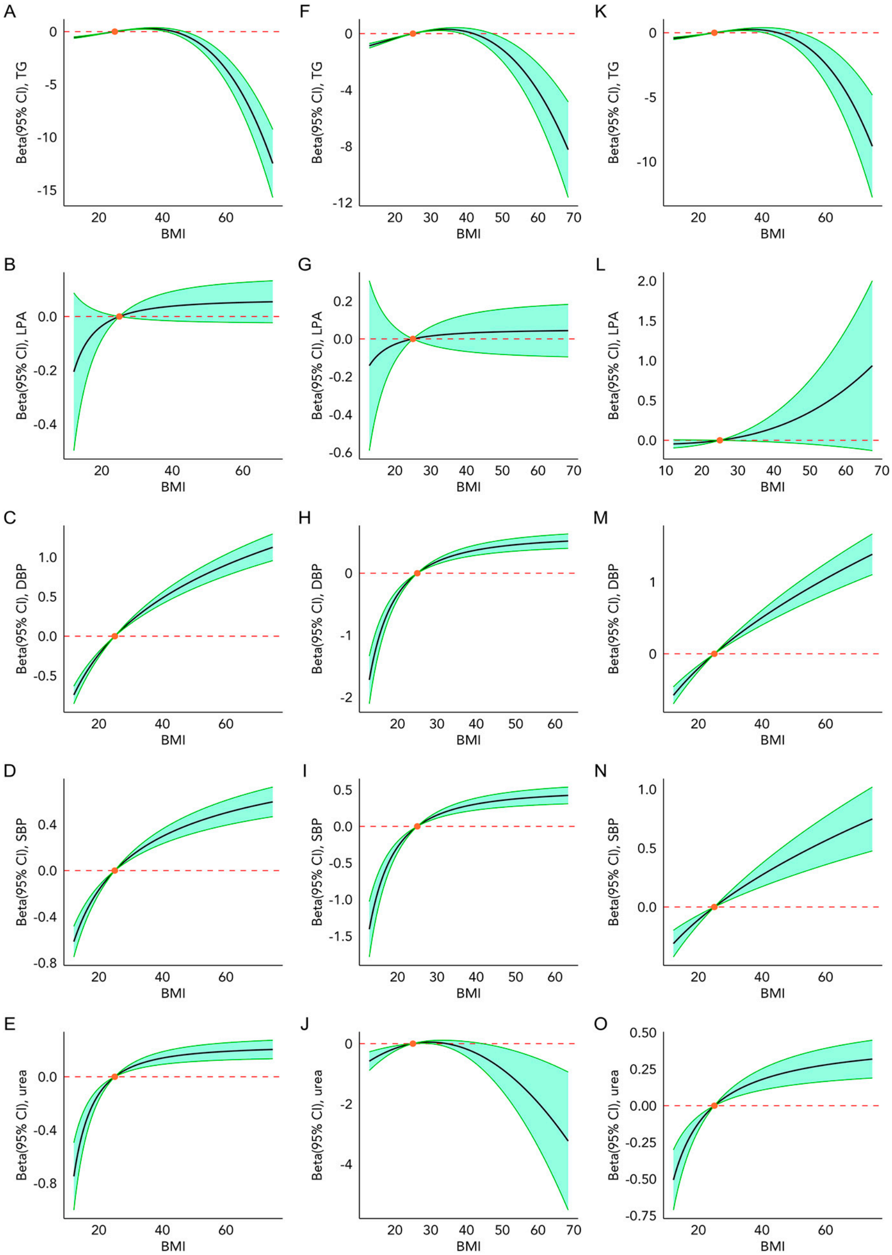


Figure 15
Untruncated plots showing estimated shape of the causal relationships between BMI and cardiometabolic biomarkers in combined analyses (panels A to E), men (panels F to J) and women (panels K to O). Shape estimates are derived from the function of fractional polynomials that best fits the data. Solid black line represents the function curve, green band represents 95% CI, red dot represents reference BMI of 25kg/m2 and the dashed line represents the null effect size. GLU = glucose, T_CHOL = total cholesterol.

Paper 4 discussion

In this study we demonstrate that the causal relationships between excess adiposity, assessed as BMI, and cardiometabolic diseases and risk factors are non-linear and that excess adiposity is causal of any cardiometabolic disease outcome. The largest causal effect of BMI was observed on T2D underscoring the fact that excess adiposity is the most important predictor of T2D. Significant sex differences were observed for causal effects of BMI on CAD risk and levels of total and LDL cholesterol but this significance only persisted for LDL after accounting for multiple testing. We further found evidence of non-linear causal associations between BMI and cardiometabolic biomarkers and disease events.

Despite caveats of multiple testing, the sex differences observed in causal effect of BMI on CAD, and the impact of menopause status, or younger age in women, on causal effects of obesity on CAD are important results. Excess adiposity seems invalidate the protections of premenopausal state which shows the deleterious effects obesity has on health.

Different approaches have been applied before to assess causal effects of adiposity various cardiometabolic outcomes. WHR adjusted for BMI was causally associated with T2D and CHD[146], and BMI was associated with T2D, CHD, lower HDL but not stroke[83]. At least one study investigated sex differences in causal effects of BMI on CAD but did not find a significant difference after multiple testing which may owe to different SNPs used to compute BMI PRS[93]. Findings on lipids may reflect dyslipidaemia of obesity characterized by high TGs and FFAs; normal or decreased LDL; and decreased HDL (accompanied by HDL dysfunction which is characterized by altered reverse cholesterol transport and tendency towards proinflammation), all attributed to altered lipid metabolism favouring hypertriglyceridaemia[147].

Owing to the conservative nature of Bonferroni correction, the sex differences identified within this causal framework are worth considering. Sexual dimorphism has been observed in peak age of obesity, differential distribution of body fat, and substrate preference for energy metabolism in different states of activity/rest, with women having better insulin sensitivity than men[8,89-92,148]. Despite these differences being partially attributed to the protective effects of oestrogen, obesity seems to attenuate this protection[149], which we also observed in our study. These findings are however contradicted by studies that showed that hormone replacement therapy increases the risk of CAD in postmenopausal women[150] or increased the risk of other diseases like breast cancer[151]. In the Women's Health Initiative (WHI) study, they did not report any important interaction with BMI nor were subgroup analyses based on BMI performed[150]. Further, there is conflicting evidence on the cardiometabolic risk profile in premenopausal women with some studies showing increased risk and others decreased[152]. Other studies also reported differential obesity-related risk for CAD by age though we only observed these differences in women only, and furthermore these studies were largely observational[153]. In older women, other factors influencing cardiovascular risk other than hormonal change have been suggested: psychosocial factors including stress and depression; sleep disorders; vitamin deficiencies, especially vitamin D; and inflammation related to rheumatoid arthritis[154].

We demonstrate sex differences in the causal effects of BMI on CAD and describe in detail the shapes of causal effects of BMI on cardiometabolic diseases and risk factors. We found evidence of nonlinearity in the causal effect of BMI on diseases and risk factor biomarkers except LPA. The adverse consequences of BMI in CAD risk are similar in men and women at younger ages. It is possible that as adults age, risk of CAD due to excess adiposity diminishes or other stronger or competing risk factors come into play. However, in men BMI continues to convey increased CAD risk, whereas in older women, BMI is no longer a risk factor. Thus, whilst weight loss may be a sensible preventive measure in younger women and men of all ages, it may not be beneficial in post-menopausal women.

The overall implication of this study is that we understand the nature of causal relationships between BMI and cardiometabolic and possibly can estimate benefits of intervening at various levels of BMI. It also highlights the role of sex in CAD, lipids and glucose homeostasis in the context of causal risk conferred by excess adiposity and underscores the need for sex consideration in the management of excess adiposity and cardiometabolic diseases. The study also emphasizes the importance of maintaining a healthy body weight even among women who are deemed to have natural protection by virtue of their sexual hormones.

Overall summary and conclusions

The advantage of establishing causality (and further elucidating the nature of the underlying causal relationships), is that we are able to institute interventions with a degree of certainty that we are targeting the right exposure and that the interventions will work. However, in order to develop therapeutic agents, we must discover the underlying biological mechanisms, through which the exposure causes disease, via functional studies. In this project, we established causal effects of prediabetes (hyperglycaemia independent of threshold-based T2D diagnosis) on micro- and macrovascular disease, and the causal effects of excess adiposity (BMI) on cardiometabolic outcomes and further described their nature of these relationships. Through functional studies, we described the role of *Lrig/sma-10* and *LRIG1* gene variants in lipid metabolism then investigated the effects these genetic variants on lipid/energy metabolism phenotypes in humans using population-level data. We also explored the utility of clinical and omics data in predicting NAFLD, an approach which can be used to identify important biomarkers for prognosis or diagnosis of NAFLD or other diseases. The sum of approaches in this project thus summarises the wish of research: establishing associations, confirming causality and unravelling the biological processes that form the basis of a phenotype; and identifying means of diagnosing or estimating prognosis of a disease. This project provides insights which can be pursued further to improve our understanding of relationships between energy homeostasis and health outcomes. I conclude this project as follows:

- Glucose metabolism and therefore glycaemic regulation is essential for energy homeostasis. Hyperglycaemia, which results from perturbations in glucose regulation progressively damages to the vasculature which ultimately manifest as overt vascular disease, especially CAD. These effects are unconfounded by the threshold-based diagnosis of T2D disease state but are rather driven by glucose levels that are above the tightly controlled normal range. The disproportionate burden of cardiovascular complications in T2D and the challenges of preventing these complications in established T2D may reflect prolonged exposure to hyperglycaemia, which leads to severe irreversible vascular damage.

  Given that cardiovascular diseases are the leading cause of death in T2D, it is necessary that recommendations consider simple and effective interventions for prediabetes, including therapeutic ones. However, therapeutic interventions are subject to risk-benefit and cost-benefit analyses in different population groups. Studies that have previously investigated use of medication for prediabetes did not necessarily enrol participants in their early stages of dysglycaemia leading to mixed results of pharmacological interventions with regards to key outcomes; CAD, mortality, CVD mortality and development of other vascular complications[155]. In addition, risk of both short-term and long-term side-effects of therapies may outweigh benefits of intervening in persons who generally are considered amenable to lifestyle modifications, though the ADA approves use of metformin for certain high-risk groups[156]. A more recent review of evidence found strong evidence for metformin but lifestyle interventions were superior to metformin in reducing T2D incidence which further divides opinion on whether to intervene pharmacologically or not[157]. While wading into this controversy was by no means the purpose of this thesis, precision medicine approaches with patient stratification may offer tailored solutions based on case-by-case merit. Further, screening for hyperglycaemia in those at risk (based on simple anthropometric measures, lifestyle and demographic factors) and instituting early interventions has the potential to significantly reduce T2D-related CAD and mortality.

- Functional studies are important in unravelling the underlying molecular biology of genetic variants associated with a particular phenotype. The *LRIG1* gene variants showed important discordant relationships with BMI and T2D, and a healthy lipid metabolic phenotype likely driven hyperplastic adiposity. In humans, functional studies of this gene have mainly been conducted in cancer cells making this one of the first studies to investigate associations with adipocyte morphology and metabolic traits. There are opportunities for further invitro functional studies in human cells and identification of potential therapeutic targets, and also population studies.

- Undetected and untreated NAFLD can progress to fibrosis or hepatocellular carcinoma, stages that are difficult to manage and likely fatal without liver transplantation. It is difficult to tell who would progress from NAFLD to cirrhosis and who would not. NAFLD however, can be managed with simple lifestyle changes if detected in time. However, given that definitive diagnosis with biopsy carries complications and imaging is not always available or reliable, methods for identifying NAFLD that are simple, safe and reliable would be helpful in bridging the detection gap. Prediction models have been used previously with different performance rates and using different variables. With proliferation of omics data and other big data in health and biomedical research, ML can be useful in developing prediction models based on more detailed data, and identifying biomarkers which can be further investigated and for diagnostic or prognostic purposes. While the eventual clinical use of identified biomarkers may take time, interventions with virtually no risk like lifestyle changes can be prescribed to high-risk groups detected using a prediction model.
- The deleterious effects of excess adiposity on cardiometabolic health have been studied extensively, and we show that adiposity causally drives cardiometabolic outcomes in a non-linear way. This is in line with studies that have shown non-linear causal relationship between adiposity and mortality. Visual representation of these relationships showed that the causal risk of disease increases exponentially as BMI increases. The implication of this is that early intervention in the trajectory of excess adiposity significantly reduces its detrimental effects. From these analyses, it is also possible to estimate the magnitude of causal risk attenuated by intervening at a particular level of exposure, BMI in this case. This offers a useful tool and incorporating such causal estimates would offer more valuable information to policy makers.
- Sexual dimorphism in cardiometabolic health remains an important factor in research and design of interventions. Differential risk profiles at same age and weight in men and women seem to favour women though some of these advantages disappear in obesity and with advanced age. It is also possible that on average, men are exposed to the detrimental effects of excess adiposity for a longer duration given that obesity peaks 10 years earlier compared to women. On the other hand, women tend to handle their naturally higher fat mass and circulating lipids more efficiently. The detrimental effects of obesity seem to attenuate the putative protective effects of sex hormones enjoyed by younger women. Surprisingly, in older women we observed no causal association between

BMI and CAD which could be due to changes in risk factor profile as women advance in age. Research in women's health, especially when considering menopause, faces criticism of medicalization of a natural process and misrepresentation of facts[158,159]. There is thus need for more evidence to further our understanding of women's cardiometabolic health and overall sexual dimorphism in cardiometabolic disease. Our finds thus warrant further investigations.

Future perspectives

Paper 1 can be extended further as follows:

- With individual data, more detailed analyses that incorporate covariate adjustment and stratification by sex, age and other factors like menopause status. In this case, we'd be investigate any sex or age group differences and the effects of these factors on causal associations between dysglycaemia and vascular diseases. Using individual data, we can also perform time-to-event causal inference analyses to determine if time influences these causal associations. We can identify genetic subgroups who develop and who do not develop prediabetes and investigate their underlying differences. One useful approach would be to stratify individuals by quantiles of a prediabetes PRS and then compare via the aforementioned analyses. Finally, for a more general study, we can compute the casual population attributable fractions (PAF) of CAD due to prediabetes.

Paper two offers opportunity for both cellular and population level studies.

- In population studies we can investigate effects of *LRIG1* genetic variants on the following:
    - Longitudinal weight change, with and without interventions
    - Modification of cardiometabolic risk due to obesity (including extreme obesity)
    - Interaction with environmental factors in phenotype development
    - Other measures of adiposity and fat distribution: SAT, WHR, waist circumference, visceral obesity and body fat levels
    - Prognosis of T2D, including risk of complications and response to treatment.

- In cellular studies, CRISPR-cas9 can be used to edit cell-lines to investigate the effects of different *LRIG1* alleles on lipid metabolism. Further, impact of lifestyle can be tested in vitro using lifestyle mimetic agents. Thereafter, for alleles with confirmed functions, gene-based recall studies can be set up to investigate population level effects.

In paper 3 we did not assess prediction of NAFLD in men and women separately mainly due to sample size issues. If that constraint could be overcome, we could determine whether our models perform equally in men and women or if there are differences. It would also be useful to determine subgroups of susceptibility especially defined by an obesity or T2D PRS and compare these groups, in terms of prediction yield, to subgroups defined by a NAFLD PRS.

Finally, paper 4 being my last one towards the end of my PhD is still work in progress. One of the things I realized while working through the project was that it was quite broad and therefore it would be prudent to have smaller more focused extensions in the future. These extensions could cover the following:

- Computing causal population attributable fraction of specific cardiometabolic outcomes due to BMI
- Perform time-to-event causal effect analyses
- Investigate effects of hormone replacement therapy on causal associations between BMI and cardiometabolic outcomes in women.
- Use different measures of adiposity e.g., WHR, body fat percentage, waist circumference, and compare the causal effects on cardiometabolic outcomes to those of BMI on the same outcomes.
- Lastly, we could compare causal effects generated from analyses that use sex-specific instruments (SNPs) to generate PRSs to those that use SNPs from combined GWAS.

References

1 Galgani, J. & Ravussin, E. Energy metabolism, fuel selection and body weight regulation. *International journal of obesity (2005)* **32 Suppl 7**, S109-S119, doi:10.1038/ijo.2008.246 (2008).

2 Hill, J. O., Wyatt, H. R. & Peters, J. C. Energy balance and obesity. *Circulation* **126**, 126-132, doi:10.1161/CIRCULATIONAHA.111.087213 (2012).

3 Hill, J. O., Wyatt, H. R. & Peters, J. C. Energy balance and obesity. *Circulation* **126**, 126-132, doi:10.1161/circulationaha.111.087213 (2012).

4 Franks, P. W. & Atabaki-Pasdar, N. Causal inference in obesity research. *J Intern Med* **281**, 222-232, doi:10.1111/joim.12577 (2017).

5 Bhupathiraju, S. N. & Hu, F. B. Epidemiology of Obesity and Diabetes and Their Cardiovascular Complications. *Circulation Research* **118**, 1723-1735, doi:doi:10.1161/CIRCRESAHA.115.306825 (2016).

6 Organization, W. H. *Obesity and overweight*, <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight> (2021).

7 Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19·2 million participants. *Lancet* **387**, 1377-1396, doi:10.1016/s0140-6736(16)30054-x (2016).

8 Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *New England Journal of Medicine* **377**, 13-27, doi:10.1056/NEJMoa1614362 (2017).

9 Kelly, T., Yang, W., Chen, C. S., Reynolds, K. & He, J. Global burden of obesity in 2005 and projections to 2030. *Int J Obes (Lond)* **32**, 1431-1437, doi:10.1038/ijo.2008.102 (2008).

10 Forouhi, N. G. & Wareham, N. J. Epidemiology of diabetes. *Medicine* **47**, 22-27, doi:https://doi.org/10.1016/j.mpmed.2018.10.004 (2019).

11 IDF Diabetes Atlas. (International Diabetes Federation, Brussels, Belgium, 2021).

12 Organization, W. H. Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation. (2006).

13 Diagnosis and classification of diabetes mellitus. *Diabetes Care* **34 Suppl 1**, S62-69, doi:10.2337/dc11-S062 (2011).

14 Yip, W. C. Y., Sequeira, I. R., Plank, L. D. & Poppitt, S. D. Prevalence of Pre-Diabetes across Ethnicities: A Review of Impaired Fasting Glucose (IFG) and Impaired Glucose Tolerance (IGT) for Classification of Dysglycaemia. *Nutrients* **9**, doi:10.3390/nu9111273 (2017).

15 Heianza, Y. *et al.* HbA1c 5·7-6·4% and impaired fasting plasma glucose for diagnosis of prediabetes and risk of progression to diabetes in Japan (TOPICS 3): a longitudinal cohort study. *Lancet* **378**, 147-155, doi:10.1016/s0140-6736(11)60472-8 (2011).

16 Knowler, W. C. *et al.* 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. *Lancet* **374**, 1677-1686, doi:10.1016/s0140-6736(09)61457-4 (2009).

17 Yeboah, J., Bertoni, A. G., Herrington, D. M., Post, W. S. & Burke, G. L. Impaired fasting glucose and the risk of incident diabetes mellitus and cardiovascular events in an adult population: MESA (Multi-Ethnic Study of Atherosclerosis). *J Am Coll Cardiol* **58**, 140-146, doi:10.1016/j.jacc.2011.03.025 (2011).

18 Li, G. *et al.* The long-term effect of lifestyle interventions to prevent diabetes in the China Da Qing Diabetes Prevention Study: a 20-year follow-up study. *Lancet* **371**, 1783-1789, doi:10.1016/s0140-6736(08)60766-7 (2008).

19 Lindström, J. *et al.* The Finnish Diabetes Prevention Study (DPS): Lifestyle intervention and 3-year results on diet and physical activity. *Diabetes Care* **26**, 3230-3236, doi:10.2337/diacare.26.12.3230 (2003).

20 Swinburn, B. A. *et al.* The global obesity pandemic: shaped by global drivers and local environments. *Lancet* **378**, 804-814, doi:10.1016/S0140-6736(11)60813-1 (2011).

21 Cooper, C. B., Neufeld, E. V., Dolezal, B. A. & Martin, J. L. Sleep deprivation and obesity in adults: a brief narrative review. *BMJ Open Sport Exerc Med* **4**, e000392-e000392, doi:10.1136/bmjsem-2018-000392 (2018).

22 Blüher, M. Obesity: global epidemiology and pathogenesis. *Nature Reviews Endocrinology* **15**, 288-298, doi:10.1038/s41574-019-0176-8 (2019).

23 Ludwig, J. *et al.* Neighborhoods, Obesity, and Diabetes — A Randomized Social Experiment. *New England Journal of Medicine* **365**, 1509-1519, doi:10.1056/NEJMsa1103216 (2011).

24 Loos, R. J. F. & Yeo, G. S. H. The genetics of obesity: from discovery to biology. *Nature Reviews Genetics* **23**, 120-133, doi:10.1038/s41576-021-00414-z (2022).

25 Nieuwdorp, M., Gilijamse, P. W., Pai, N. & Kaplan, L. M. Role of the microbiome in energy regulation and metabolism. *Gastroenterology* **146**, 1525-1533, doi:10.1053/j.gastro.2014.02.008 (2014).

26 DeGruttola, A. K., Low, D., Mizoguchi, A. & Mizoguchi, E. Current Understanding of Dysbiosis in Disease in Human and Animal Models. *Inflamm Bowel Dis* **22**, 1137-1150, doi:10.1097/MIB.0000000000000750 (2016).

27 El-Sayed Moustafa, J. S. & Froguel, P. From obesity genetics to the future of personalized obesity therapy. *Nat Rev Endocrinol* **9**, 402-413, doi:10.1038/nrendo.2013.57 (2013).

28 Bray, M. S. *et al.* NIH working group report—using genomic information to guide weight management: From universal to precision treatment. *Obesity* **24**, 14-22, doi:https://doi.org/10.1002/oby.21381 (2016).

29 Bouchard, C. Genetics of Obesity: What We Have Learned Over Decades of Research. *Obesity* **29**, 802-820, doi:https://doi.org/10.1002/oby.23116 (2021).

30 Farooqi, S. & O'Rahilly, S. Genetics of obesity in humans. *Endocr Rev* **27**, 710-718, doi:10.1210/er.2006-0040 (2006).

31 Scuteri, A. *et al.* Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet* **3**, e115, doi:10.1371/journal.pgen.0030115 (2007).

32 Hinney, A. *et al.* Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants. *PLoS One* **2**, e1361, doi:10.1371/journal.pone.0001361 (2007).

33 Peters, U. *et al.* A systematic mapping approach of 16q12.2/FTO and BMI in more than 20,000 African Americans narrows in on the underlying functional variation: results from the Population Architecture using Genomics and Epidemiology (PAGE) study. *PLoS Genet* **9**, e1003171, doi:10.1371/journal.pgen.1003171 (2013).

34 Li, H. *et al.* Association of genetic variation in FTO with risk of obesity and type 2 diabetes with data from 96,551 East and South Asians. *Diabetologia* **55**, 981-995, doi:10.1007/s00125-011-2370-7 (2012).

35 Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641-3649, doi:10.1093/hmg/ddy271 (2018).

36 Akbari, P. *et al.* Sequencing of 640,000 exomes identifies GPR75 variants associated with protection from obesity. *Science* **373**, doi:10.1126/science.abf8683 (2021).

37 Panzeri, I. & Pospisilik, J. A. Epigenetic control of variation and stochasticity in metabolic disease. *Mol Metab* **14**, 26-38, doi:10.1016/j.molmet.2018.05.010 (2018).

38 van der Klaauw, Agatha A. & Farooqi, I. S. The Hunger Genes: Pathways to Obesity. *Cell* **161**, 119-132, doi:https://doi.org/10.1016/j.cell.2015.03.008 (2015).

39 Richard, D. Cognitive and autonomic determinants of energy homeostasis in obesity. *Nature Reviews Endocrinology* **11**, 489-501, doi:10.1038/nrendo.2015.103 (2015).

40 Schwartz, M. W., Woods, S. C., Porte, D., Jr., Seeley, R. J. & Baskin, D. G. Central nervous system control of food intake. *Nature* **404**, 661-671, doi:10.1038/35007534 (2000).

41 Oussaada, S. M. *et al.* The pathogenesis of obesity. *Metabolism* **92**, 26-36, doi:https://doi.org/10.1016/j.metabol.2018.12.012 (2019).

42 Lutter, M. & Nestler, E. J. Homeostatic and Hedonic Signals Interact in the Regulation of Food Intake. *The Journal of Nutrition* **139**, 629-632, doi:10.3945/jn.108.097618 (2009).

43 Chapelot, D. & Charlot, K. Physiology of energy homeostasis: Models, actors, challenges and the glucoadipostatic loop. *Metabolism* **92**, 11-25, doi:https://doi.org/10.1016/j.metabol.2018.11.012 (2019).

44 Tabák, A. G. *et al.* Trajectories of glycaemia, insulin sensitivity, and insulin secretion before diagnosis of type 2 diabetes: an analysis from the Whitehall II study. *Lancet* **373**, 2215-2221, doi:10.1016/s0140-6736(09)60619-x (2009).

45 Ferrannini, E. *et al.* Mode of onset of type 2 diabetes from normal or impaired glucose tolerance. *Diabetes* **53**, 160-165, doi:10.2337/diabetes.53.1.160 (2004).

46 DeFronzo, R. A. From the triumvirate to the ominous octet: a new paradigm for the treatment of type 2 diabetes mellitus. *Diabetes* **58**, 773-795 (2009).

47 Morigny, P., Houssier, M., Mouisel, E. & Langin, D. Adipocyte lipolysis and insulin resistance. *Biochimie* **125**, 259-266, doi:10.1016/j.biochi.2015.10.024 (2016).

48 DeFronzo, R. A. *et al.* Type 2 diabetes mellitus. *Nat Rev Dis Primers* **1**, 15019, doi:10.1038/nrdp.2015.19 (2015).

49 Matsuda, M. *et al.* Glucagon dose-response curve for hepatic glucose production and glucose disposal in type 2 diabetic patients and normal individuals. *Metabolism-Clinical and Experimental* **51**, 1111-1119 (2002).

50 Gastaldelli, A. *et al.* Influence of obesity and type 2 diabetes on gluconeogenesis and glucose output in humans: a quantitative study. *Diabetes* **49**, 1367-1373 (2000).

51 Clore, J. N., Stillman, J. & Sugerman, H. Glucose-6-phosphatase flux in vitro is increased in type 2 diabetes. *Diabetes* **49**, 969-974 (2000).

52 Ferrannini, E. & Mari, A. β-Cell function in type 2 diabetes. *Metabolism* **63**, 1217-1227 (2014).

53 Kahn, S. E., Cooper, M. E. & Del Prato, S. Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future. *Lancet* **383**, 1068-1083, doi:10.1016/s0140-6736(13)62154-6 (2014).

54 Heymsfield, S. B. & Wadden, T. A. Mechanisms, Pathophysiology, and Management of Obesity. *New England Journal of Medicine* **376**, 254-266, doi:10.1056/NEJMra1514009 (2017).

55 Berrington de Gonzalez, A. *et al.* Body-Mass Index and Mortality among 1.46 Million White Adults. *New England Journal of Medicine* **363**, 2211-2219, doi:10.1056/NEJMoa1000367 (2010).

56 Kivimäki, M. *et al.* Overweight, obesity, and risk of cardiometabolic multimorbidity: pooled analysis of individual-level data for 120 813 adults from 16 cohort studies from the USA and Europe. *The Lancet Public Health* **2**, e277-e285, doi:https://doi.org/10.1016/S2468-2667(17)30074-9 (2017).

57 Bray, G. A., Kim, K. K., Wilding, J. P. H. & Federation, o. b. o. t. W. O. Obesity: a chronic relapsing progressive disease process. A position statement of the World Obesity Federation. *Obesity Reviews* **18**, 715-723, doi:https://doi.org/10.1111/obr.12551 (2017).

58 Grant, R. W. & Dixit, V. D. Adipose tissue as an immunological organ. *Obesity (Silver Spring)* **23**, 512-518, doi:10.1002/oby.21003 (2015).

59 Arita, Y. *et al.* Paradoxical decrease of an adipose-specific protein, adiponectin, in obesity. 1999. *Biochem Biophys Res Commun* **425**, 560-564, doi:10.1016/j.bbrc.2012.08.024 (2012).

60 Ellulu, M. S., Patimah, I., Khaza'ai, H., Rahmat, A. & Abed, Y. Obesity and inflammation: the linking mechanism and the complications. *Arch Med Sci* **13**, 851-863, doi:10.5114/aoms.2016.58928 (2017).

61 Sattar, N., Forrest, E. & Preiss, D. Non-alcoholic fatty liver disease. *Bmj* **349**, g4596, doi:10.1136/bmj.g4596 (2014).

62 Younossi, Z. M. *et al.* Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology* **64**, 73-84, doi:10.1002/hep.28431 (2016).

63 Chalasani, N. *et al.* The diagnosis and management of non-alcoholic fatty liver disease: practice Guideline by the American Association for the Study of Liver Diseases, American College of Gastroenterology, and the American Gastroenterological Association. *Hepatology* **55**, 2005-2023, doi:10.1002/hep.25762 (2012).

64 Fazel, Y., Koenig, A. B., Sayiner, M., Goodman, Z. D. & Younossi, Z. M. Epidemiology and natural history of non-alcoholic fatty liver disease. *Metabolism* **65**, 1017-1025, doi:https://doi.org/10.1016/j.metabol.2016.01.012 (2016).

65 Thoma, C., Day, C. P. & Trenell, M. I. Lifestyle interventions for the treatment of non-alcoholic fatty liver disease in adults: a systematic review. *J Hepatol* **56**, 255-266, doi:10.1016/j.jhep.2011.06.010 (2012).

66 Huang, Y., Cai, X., Mai, W., Li, M. & Hu, Y. Association between prediabetes and risk of cardiovascular disease and all cause mortality: systematic review and meta-analysis. *Bmj* **355**, i5953, doi:10.1136/bmj.i5953 (2016).

67 Palladino, R. *et al.* Association between pre-diabetes and microvascular and macrovascular disease in newly diagnosed type 2 diabetes. *BMJ Open Diabetes Research &amp; Care* **8**, e001061, doi:10.1136/bmjdrc-2019-001061 (2020).

68 Barr, E. L. *et al.* Risk of cardiovascular and all-cause mortality in individuals with diabetes mellitus, impaired fasting glucose, and impaired glucose tolerance: the Australian Diabetes, Obesity, and Lifestyle Study (AusDiab). *Circulation* **116**, 151-157 (2007).

69 Sourij, H. *et al.* Post-challenge hyperglycaemia is strongly associated with future macrovascular events and total mortality in angiographied coronary patients. *European Heart Journal* **31**, 1583-1590, doi:10.1093/eurheartj/ehq099 (2010).

70 Stratton, I. M. *et al.* Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. *BMJ* **321**, 405-412, doi:10.1136/bmj.321.7258.405 (2000).

71 Selvin, E. *et al.* Glycated hemoglobin, diabetes, and cardiovascular risk in nondiabetic adults. *N Engl J Med* **362**, 800-811, doi:10.1056/NEJMoa0908359 (2010).

72 Organization, W. H. *Obesity and overweight*, <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight> (2021).

73 Aune, D., Schlesinger, S., Norat, T. & Riboli, E. Body mass index, abdominal fatness, and the risk of sudden cardiac death: a systematic review and dose-response meta-analysis of prospective studies. *Eur J Epidemiol* **33**, 711-722, doi:10.1007/s10654-017-0353-9 (2018).

74 Clifton, P. M. & Keogh, J. B. Effects of Different Weight Loss Approaches on CVD Risk. *Curr Atheroscler Rep* **20**, 27, doi:10.1007/s11883-018-0728-8 (2018).

75 Haase, C. L. *et al.* Weight loss and risk reduction of obesity-related outcomes in 0.5 million people: evidence from a UK primary care database. *International Journal of Obesity* **45**, 1249-1258, doi:10.1038/s41366-021-00788-4 (2021).

76 Chiasson, J. L. *et al.* Acarbose treatment and the risk of cardiovascular disease and hypertension in patients with impaired glucose tolerance: the STOP-NIDDM trial. *JAMA* **290**, 486-494, doi:10.1001/jama.290.4.486 (2003).

77 Ma, C. *et al.* Effects of weight loss interventions for adults who are obese on mortality, cardiovascular disease, and cancer: systematic review and meta-analysis. *BMJ* **359**, j4849, doi:10.1136/bmj.j4849 (2017).

78 Thomas, D. C. & Conti, D. V. Commentary: the concept of 'Mendelian Randomization'. *Int J Epidemiol* **33**, 21-25, doi:10.1093/ije/dyh048 (2004).

79 Angrist, J. D., Imbens, G. W. & Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American statistical Association* **91**, 444-455 (1996).

80 Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med* **27**, 1133-1163, doi:10.1002/sim.3034 (2008).

81 Merino, J. *et al.* Genetically Driven Hyperglycemia Increases Risk of Coronary Artery Disease Separately From Type 2 Diabetes. *Diabetes care* **40**, 687-693, doi:10.2337/dc16-2625 (2017).

82 Au Yeung, S. L., Luo, S. & Schooling, C. M. The Impact of Glycated Hemoglobin (HbA1c) on Cardiovascular Disease Risk: A Mendelian Randomization Study Using UK Biobank. *Diabetes Care* **41**, 1991-1997, doi:10.2337/dc18-0289 (2018).

83 Dale, C. E. *et al.* Causal Associations of Adiposity and Body Fat Distribution With Coronary Heart Disease, Stroke Subtypes, and Type 2 Diabetes Mellitus: A Mendelian Randomization Analysis. *Circulation* **135**, 2373-2388, doi:10.1161/CIRCULATIONAHA.116.026560 (2017).

84 Larsson, S. C., Back, M., Rees, J. M. B., Mason, A. M. & Burgess, S. Body mass index and body composition in relation to 14 cardiovascular conditions in UK Biobank: a Mendelian randomization study. *Eur Heart J* **41**, 221-226, doi:10.1093/eurheartj/ehz388 (2020).

85 Fall, T. *et al.* The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis. *PLoS Med* **10**, e1001474, doi:10.1371/journal.pmed.1001474 (2013).

86 Larsson, S. C. & Burgess, S. Causal role of high body mass index in multiple chronic diseases: a systematic review and meta-analysis of Mendelian randomization studies. *BMC Medicine* **19**, 320, doi:10.1186/s12916-021-02188-x (2021).

87 Sun, Y. Q. *et al.* Body mass index and all cause mortality in HUNT and UK Biobank studies: linear and non-linear mendelian randomisation analyses. *BMJ* **364**, l1042, doi:10.1136/bmj.l1042 (2019).

88 Zheng, J. *et al.* Trans-ethnic Mendelian-randomization study reveals causal relationships between cardiometabolic factors and chronic kidney disease. *International Journal of Epidemiology*, doi:10.1093/ije/dyab203 (2021).

89 Goossens, G. H., Jocken, J. W. E. & Blaak, E. E. Sexual dimorphism in cardiometabolic health: the role of adipose tissue, muscle and liver. *Nature Reviews Endocrinology* **17**, 47-66, doi:10.1038/s41574-020-00431-8 (2021).

90 Lumish, H. S., O'Reilly, M. & Reilly, M. P. Sex Differences in Genomic Drivers of Adipose Distribution and Related Cardiometabolic Disorders: Opportunities for Precision Medicine. *Arterioscler Thromb Vasc Biol* **40**, 45-60, doi:10.1161/atvbaha.119.313154 (2020).

91 Mauvais-Jarvis, F. Sex differences in metabolic homeostasis, diabetes, and obesity. *Biol Sex Differ* **6**, 14, doi:10.1186/s13293-015-0033-y (2015).

92 Koutsari, C. *et al.* Nonoxidative free fatty acid disposal is greater in young women than men. *J Clin Endocrinol Metab* **96**, 541-547, doi:10.1210/jc.2010-1651 (2011).

93 Censin, J. C. *et al.* Causal relationships between obesity and the leading causes of death in women and men. *PLOS Genetics* **15**, e1008405, doi:10.1371/journal.pgen.1008405 (2019).

94 Gumienny, T. L. *et al.* Caenorhabditis elegans SMA-10/LRIG is a conserved transmembrane protein that enhances bone morphogenetic protein signaling. *PLoS Genet* **6**, e1000963, doi:10.1371/journal.pgen.1000963 (2010).

95 Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206, doi:10.1038/nature14177 (2015).

96 Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature genetics* **44**, 991-1005 (2012).

97 Wheeler, E. *et al.* Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS medicine* **14**, e1002383 (2017).

98 Harst, P. v. d. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circulation Research* **122**, 433-443, doi:doi:10.1161/CIRCRESAHA.117.312086 (2018).

99 Deloukas, P. *et al.* Large-scale association analysis identifies new risk loci for coronary artery disease. *Nature genetics* **45**, 25-33 (2013).

100 Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics* **43**, 333-338 (2011).

101 Malik, R. *et al.* Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nature genetics* **50**, 524-537 (2018).

102 Wuttke, M. *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature genetics* **51**, 957-972 (2019).

103 Mahajan, A. *et al.* Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nature Genetics* **50**, 559-571, doi:10.1038/s41588-018-0084-1 (2018).

104 Kulyté, A. *et al.* Genome-wide association study of adipocyte lipolysis in the GENetics of adipocyte lipolysis (GENiAL) cohort. *Molecular Metabolism* **34**, 85-96, doi:https://doi.org/10.1016/j.molmet.2020.01.009 (2020).

105 Löfgren, P., Hoffstedt, J., Näslund, E., Wiren, M. & Arner, P. Prospective and controlled studies of the actions of insulin and catecholamine in fat cells of obese women following weight reduction. *Diabetologia* **48**, 2334-2342 (2005).

106 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209, doi:10.1038/s41586-018-0579-z (2018).

107 Doherty, A. *et al.* Large scale population assessment of physical activity using wrist worn accelerometers: the UK biobank study. *PloS one* **12**, e0169649 (2017).

108 Biobank, U. *UK Biobank research ethics approval* <https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics> (2021).

109 Koivula, R. W. *et al.* Discovery of biomarkers for glycaemic deterioration before and after the onset of type 2 diabetes: rationale and design of the epidemiological studies within the IMI DIRECT Consortium. *Diabetologia* **57**, 1132-1142, doi:10.1007/s00125-014-3216-x (2014).

110 Koivula, R. W. *et al.* Discovery of biomarkers for glycaemic deterioration before and after the onset of type 2 diabetes: descriptive characteristics of the epidemiological studies within the IMI DIRECT Consortium. *Diabetologia* **62**, 1601-1615, doi:10.1007/s00125-019-4906-1 (2019).

111 Sebastião, Y. V. & St. Peter, S. D. An overview of commonly used statistical methods in clinical research. *Seminars in Pediatric Surgery* **27**, 367-374, doi:https://doi.org/10.1053/j.sempedsurg.2018.10.008 (2018).

112 Baird, R. Systematic reviews and meta-analytic techniques. *Seminars in Pediatric Surgery* **27**, 338-344, doi:https://doi.org/10.1053/j.sempedsurg.2018.10.009 (2018).

113 ROBINS, J., GREENLAND, S. & BRESLOW, N. E. A GENERAL ESTIMATOR FOR THE VARIANCE OF THE MANTEL HAENSZEL ODDS RATIO. *American Journal of Epidemiology* **124**, 719-723, doi:10.1093/oxfordjournals.aje.a114447 (1986).

114 Higgins, J. P. *et al. Cochrane handbook for systematic reviews of interventions*. (John Wiley & Sons, 2019).

115 Egger, M., Davey Smith, G., Schneider, M. & Minder, C. Bias in meta-analysis detected by a simple, graphical test. *Bmj* **315**, 629-634, doi:10.1136/bmj.315.7109.629 (1997).

116 Greenland, S. & Morgenstern, H. Confounding in health research. *Annu Rev Public Health* **22**, 189-212, doi:10.1146/annurev.publhealth.22.1.189 (2001).

117 Burgess, S. *et al.* Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *European Journal of Epidemiology* **30**, 543-552, doi:10.1007/s10654-015-0011-z (2015).

118 Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* **37**, 658-665, doi:10.1002/gepi.21758 (2013).

119 Pierce, B. L., Ahsan, H. & VanderWeele, T. J. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *International Journal of Epidemiology* **40**, 740-752, doi:10.1093/ije/dyq151 (2010).

120 Burgess, S. & Thompson, S. G. Use of allele scores as instrumental variables for Mendelian randomization. *Int J Epidemiol* **42**, 1134-1144, doi:10.1093/ije/dyt093 (2013).

121 Burgess, S., Bowden, J., Fall, T., Ingelsson, E. & Thompson, S. G. Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. *Epidemiology* **28**, 30-42, doi:10.1097/ede.0000000000000559 (2017).

122 Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512-525, doi:10.1093/ije/dyv080 (2015).

123 Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic Epidemiology* **40**, 304-314, doi:https://doi.org/10.1002/gepi.21965 (2016).

124 Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics* **50**, 693-698, doi:10.1038/s41588-018-0099-7 (2018).

125 Didelez, V. & Sheehan, N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res* **16**, 309-330, doi:10.1177/0962280206077743 (2007).

126 Burgess, S., Davies, N. M. & Thompson, S. G. Instrumental variable analysis with a nonlinear exposure-outcome relationship. *Epidemiology* **25**, 877-885, doi:10.1097/ede.0000000000000161 (2014).

127 Staley, J. R. & Burgess, S. Semiparametric methods for estimation of a nonlinear exposure-outcome relationship using instrumental variables with application to Mendelian randomization. *Genet Epidemiol* **41**, 341-352, doi:10.1002/gepi.22041 (2017).

128 Royston, P. & Altman, D. G. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **43**, 429-453, doi:https://doi.org/10.2307/2986270 (1994).

129 Badillo, S. *et al.* An Introduction to Machine Learning. *Clin Pharmacol Ther* **107**, 871-885, doi:10.1002/cpt.1796 (2020).

130 Altman, N. & Krzywinski, M. The curse(s) of dimensionality. *Nature Methods* **15**, 399-400, doi:10.1038/s41592-018-0019-x (2018).

131 Oskolkov, N. in *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications* (ed Roman Egger) 151-167 (Springer International Publishing, 2022).

132 Stein-O'Brien, G. L. *et al.* Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends in Genetics* **34**, 790-805, doi:https://doi.org/10.1016/j.tig.2018.07.003 (2018).

133 Meng, C. *et al.* Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics* **17**, 628-641, doi:10.1093/bib/bbv108 (2016).

134 Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9** (2008).

135 Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* **67**, 301-320 (2005).

136 Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267-288 (1996).

137 Fonti, V. & Belitser, E. Feature selection using lasso. *VU Amsterdam research paper in business analytics* **30**, 1-25 (2017).

138 Fendler, W. *et al.* Less but better: cardioprotective lipid profile of patients with GCK-MODY despite lower HDL cholesterol level. *Acta diabetologica* **51**, 625-632 (2014).

139 Steele, A. M. *et al.* Prevalence of vascular complications among patients with glucokinase mutations and prolonged, mild hyperglycemia. *JAMA* **311**, 279-286 (2014).

140 Cardiovascular Effects of Intensive Lifestyle Intervention in Type 2 Diabetes. *New England Journal of Medicine* **369**, 145-154, doi:10.1056/NEJMoa1212914 (2013).

141 Bedogni, G. *et al.* The Fatty Liver Index: a simple and accurate predictor of hepatic steatosis in the general population. *BMC Gastroenterol* **6**, 33, doi:10.1186/1471-230x-6-33 (2006).

142 Motamed, N. *et al.* Fatty liver index vs waist circumference for predicting non-alcoholic fatty liver disease. *World J Gastroenterol* **22**, 3023-3030, doi:10.3748/wjg.v22.i10.3023 (2016).

143 Lee, J. H. *et al.* Hepatic steatosis index: a simple screening tool reflecting nonalcoholic fatty liver disease. *Dig Liver Dis* **42**, 503-508, doi:10.1016/j.dld.2009.08.002 (2010).

144 Wood, G. C. *et al.* A multi-component classifier for nonalcoholic fatty liver disease (NAFLD) based on genomic, proteomic, and phenomic data domains. *Sci Rep* **7**, 43238, doi:10.1038/srep43238 (2017).

145 Perakakis, N. *et al.* Non-invasive diagnosis of non-alcoholic steatohepatitis and fibrosis with the use of omics and supervised learning: A proof of concept study. *Metabolism* **101**, 154005, doi:10.1016/j.metabol.2019.154005 (2019).

146 Emdin, C. A. *et al.* Genetic Association of Waist-to-Hip Ratio With Cardiometabolic Traits, Type 2 Diabetes, and Coronary Heart Disease. *JAMA* **317**, 626-634, doi:10.1001/jama.2016.21042 (2017).

147 Klop, B., Elte, J. W. F. & Cabezas, M. C. Dyslipidemia in obesity: mechanisms and potential targets. *Nutrients* **5**, 1218-1240, doi:10.3390/nu5041218 (2013).

148 Schorr, M. *et al.* Sex differences in body composition and association with cardiometabolic risk. *Biology of Sex Differences* **9**, 28, doi:10.1186/s13293-018-0189-3 (2018).

149 Manrique-Acevedo, C., Chinnakotla, B., Padilla, J., Martinez-Lemus, L. A. & Gozal, D. Obesity and cardiovascular disease in women. *International Journal of Obesity* **44**, 1210-1226, doi:10.1038/s41366-020-0548-0 (2020).

150 Investigators, W. G. f. t. W. s. H. I. Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal WomenPrincipal Results From the Women's Health Initiative Randomized Controlled Trial. *JAMA* **288**, 321-333, doi:10.1001/jama.288.3.321 (2002).

151 Beral, V. Breast cancer and hormone-replacement therapy in the Million Women Study. *Lancet* **362**, 419-427, doi:10.1016/s0140-6736(03)14065-2 (2003).

152 Roa-Díaz, Z. M. *et al.* Menopause and cardiometabolic diseases: What we (don't) know and why it matters. *Maturitas* **152**, 48-56, doi:https://doi.org/10.1016/j.maturitas.2021.06.013 (2021).

153 The Emerging Risk Factors, C. Separate and combined associations of body-mass index and abdominal adiposity with cardiovascular disease: collaborative analysis of 58 prospective studies. *The Lancet* **377**, 1085-1095, doi:https://doi.org/10.1016/S0140-6736(11)60105-0 (2011).

154 Pérez-López, F. R., Chedraui, P., Gilbert, J. J. & Pérez-Roncero, G. Cardiovascular risk in menopausal women and prevalent related co-morbid conditions: facing the post-Women's Health Initiative era. *Fertil Steril* **92**, 1171-1186, doi:10.1016/j.fertnstert.2009.06.032 (2009).

155 Bansal, N. Prediabetes diagnosis and treatment: A review. *World J Diabetes* **6**, 296-303, doi:10.4239/wjd.v6.i2.296 (2015).

156 Standards of medical care in diabetes--2014. *Diabetes Care* **37 Suppl 1**, S14-80, doi:10.2337/dc14-S014 (2014).

157 Jonas, D. E. *et al.* Screening for Prediabetes and Type 2 Diabetes: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA* **326**, 744-760, doi:10.1001/jama.2021.10403 (2021).

158 Meyer, V. F. The medicalization of menopause: critique and consequences. *Int J Health Serv* **31**, 769-792, doi:10.2190/m77d-yv2y-d5nu-fxnw (2001).

159 Tunstall-Pedoe, H. Myth and paradox of coronary risk and the menopause. *The Lancet* **351**, 1425-1427, doi:https://doi.org/10.1016/S0140-6736(97)11321-6 (1998).

Paper I

nature COMMUNICATIONS

ARTICLE

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

An investigation of causal relationships between prediabetes and vascular complications

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

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

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

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

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

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

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

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

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

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

Results

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

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

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

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

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

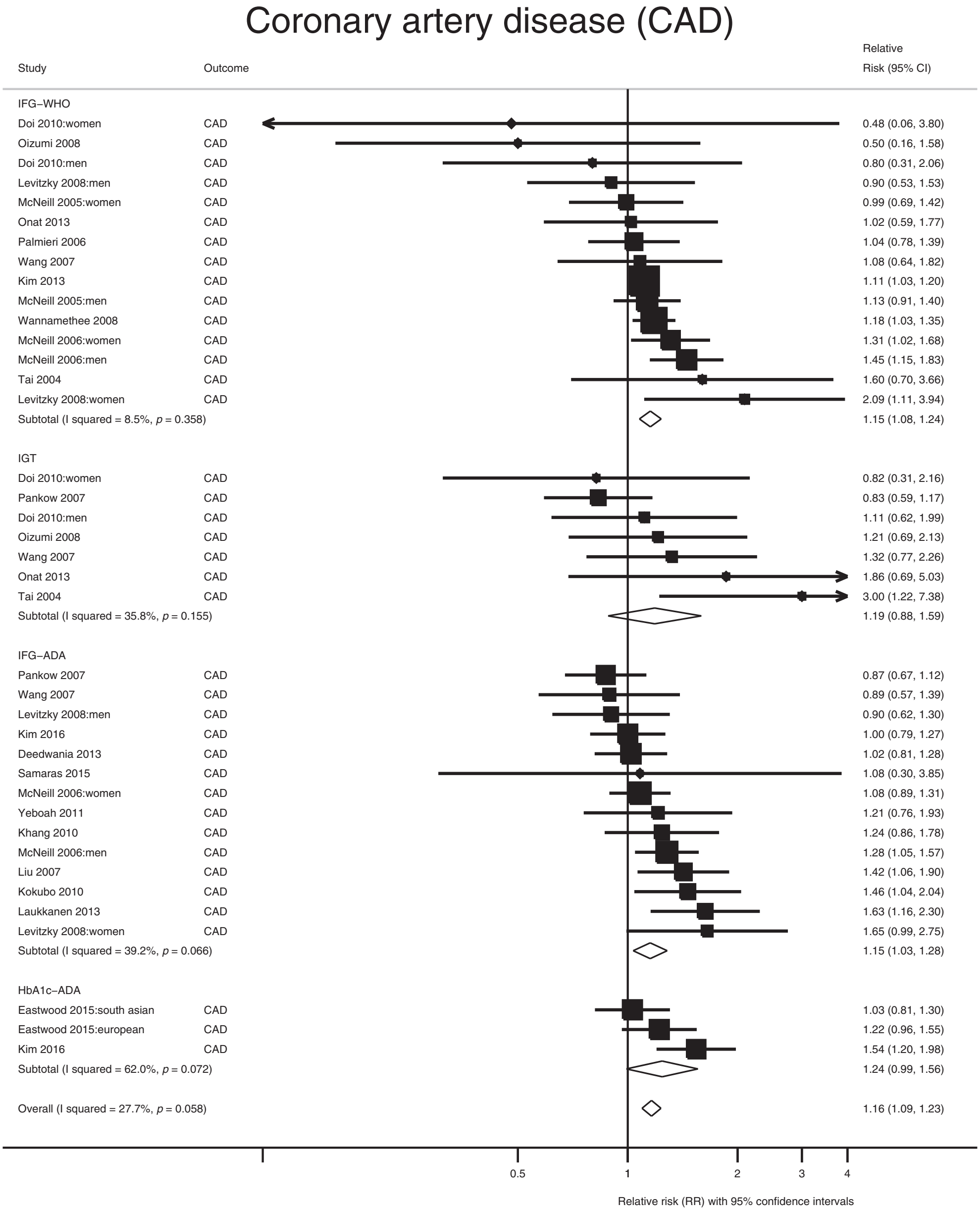


Fig. 1 Meta-analysis of the association between prediabetes and CAD. The square and diamond shapes represent effect size (relative risk estimates), while the horizontal bars represent the 95% confidence intervals. A total of 21 studies are included. All *P* values are two-sided. Source data are provided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

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

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

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

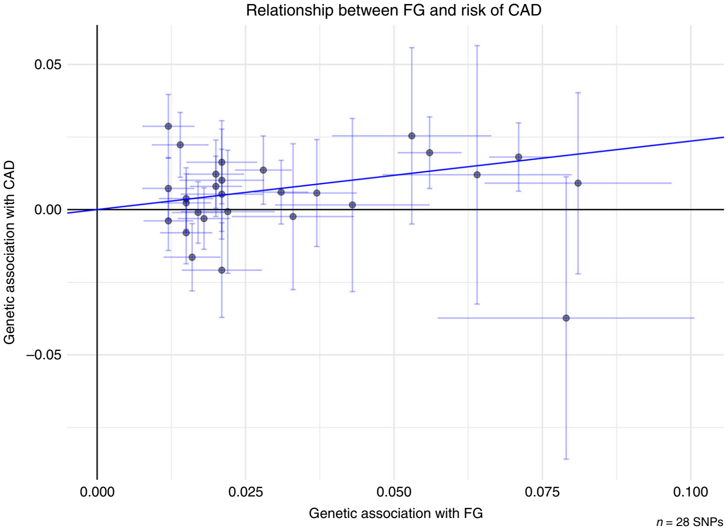


Fig. 2 Relationship between genetic effects of prediabetes only and CAD. Data are represented as log-odds and 95% confidence intervals for each trait. Slope of the line represents an estimate of the causal effect of fasting glucose on risk of CAD. The points represent effect sizes for each individual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

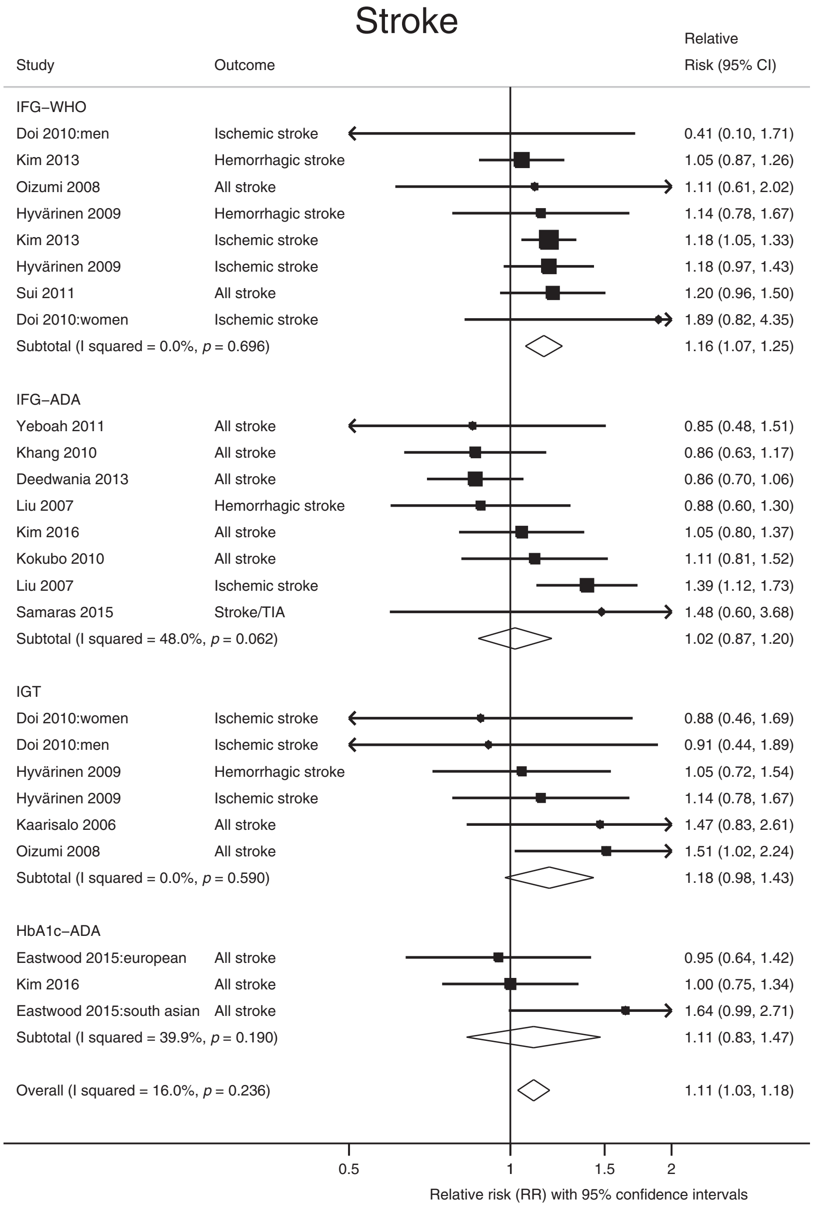
prediabetes instruments and other cardiovascular risk factors (Total, LDL, and HDL cholesterol levels; tryglyceride levels; and body mass index) did not show any significant association (Supplementary Note 2 and Supplemetary Tables 2–6).

Discussion

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

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

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



Study	Outcome	Relative Risk (95% CI)
IFG–WHO		
Doi 2010:men	Ischemic stroke	0.41 (0.10, 1.71)
Kim 2013	Hemorrhagic stroke	1.05 (0.87, 1.26)
Oizumi 2008	All stroke	1.11 (0.61, 2.02)
Hyvärinen 2009	Hemorrhagic stroke	1.14 (0.78, 1.67)
Kim 2013	Ischemic stroke	1.18 (1.05, 1.33)
Hyvärinen 2009	Ischemic stroke	1.18 (0.97, 1.43)
Sui 2011	All stroke	1.20 (0.96, 1.50)
Doi 2010:women	Ischemic stroke	1.89 (0.82, 4.35)
Subtotal (I squared = 0.0%, *p* = 0.696)		1.16 (1.07, 1.25)
IFG–ADA		
Yeboah 2011	All stroke	0.85 (0.48, 1.51)
Khang 2010	All stroke	0.86 (0.63, 1.17)
Deedwania 2013	All stroke	0.86 (0.70, 1.06)
Liu 2007	Hemorrhagic stroke	0.88 (0.60, 1.30)
Kim 2016	All stroke	1.05 (0.80, 1.37)
Kokubo 2010	All stroke	1.11 (0.81, 1.52)
Liu 2007	Ischemic stroke	1.39 (1.12, 1.73)
Samaras 2015	Stroke/TIA	1.48 (0.60, 3.68)
Subtotal (I squared = 48.0%, *p* = 0.062)		1.02 (0.87, 1.20)
IGT		
Doi 2010:women	Ischemic stroke	0.88 (0.46, 1.69)
Doi 2010:men	Ischemic stroke	0.91 (0.44, 1.89)
Hyvärinen 2009	Hemorrhagic stroke	1.05 (0.72, 1.54)
Hyvärinen 2009	Ischemic stroke	1.14 (0.78, 1.67)
Kaarisalo 2006	All stroke	1.47 (0.83, 2.61)
Oizumi 2008	All stroke	1.51 (1.02, 2.24)
Subtotal (I squared = 0.0%, *p* = 0.590)		1.18 (0.98, 1.43)
HbA1c–ADA		
Eastwood 2015:european	All stroke	0.95 (0.64, 1.42)
Kim 2016	All stroke	1.00 (0.75, 1.34)
Eastwood 2015:south asian	All stroke	1.64 (0.99, 2.71)
Subtotal (I squared = 39.9%, *p* = 0.190)		1.11 (0.83, 1.47)
Overall (I squared = 16.0%, *p* = 0.236)		1.11 (1.03, 1.18)

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

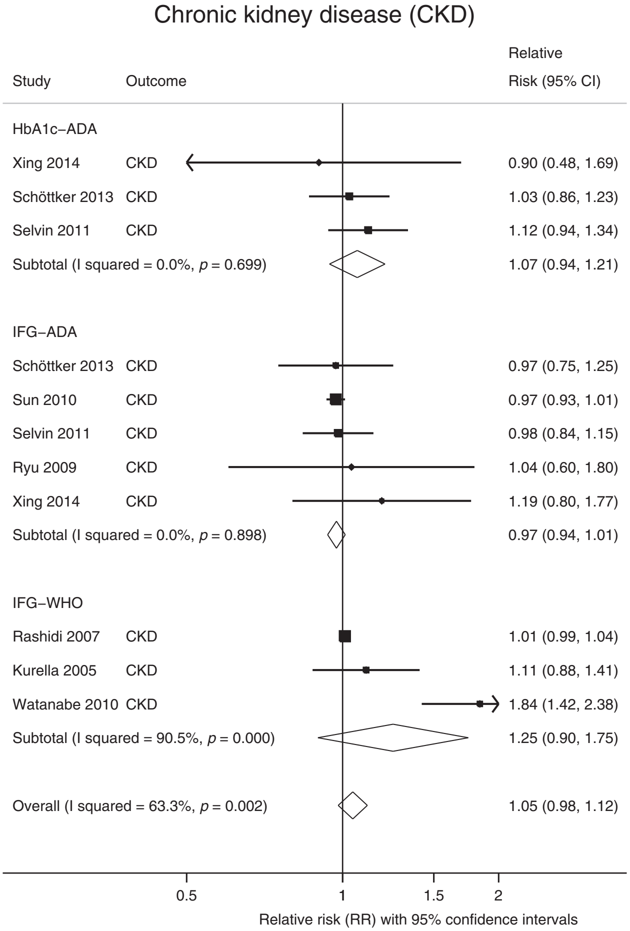


Fig. 4 Meta-analysis of the association between prediabetes and CKD. The square and diamond shapes represent effect size (relative risk estimates), while the horizontal bars represent the 95% confidence intervals. In total, eight studies are included. All *P* values are two-sided. Source data are provided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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

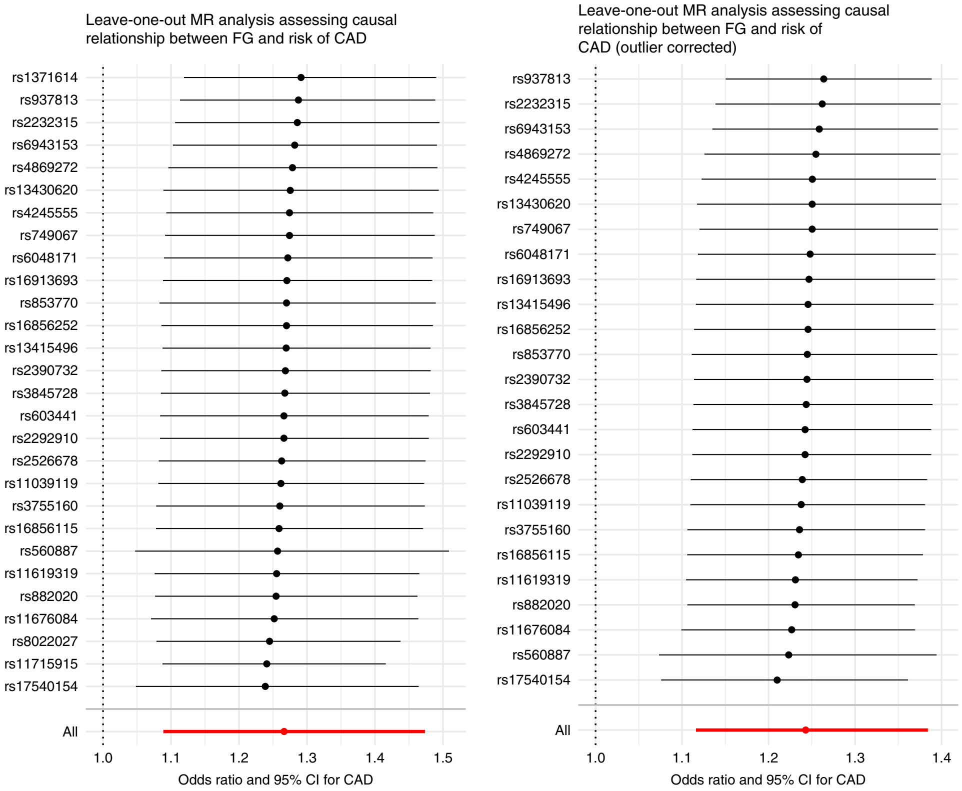


Fig. 5 Leave-one-out analysis plots of causal relationship between fasting glucose and CAD. Data are presented as odds (OR) ratio and 95% confidence interval (95% CI) of the exposure-outcome relationship for each SNP. Center points represent the causal effect estimate and the horizontal bars represent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

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

Methods

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

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

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

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

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

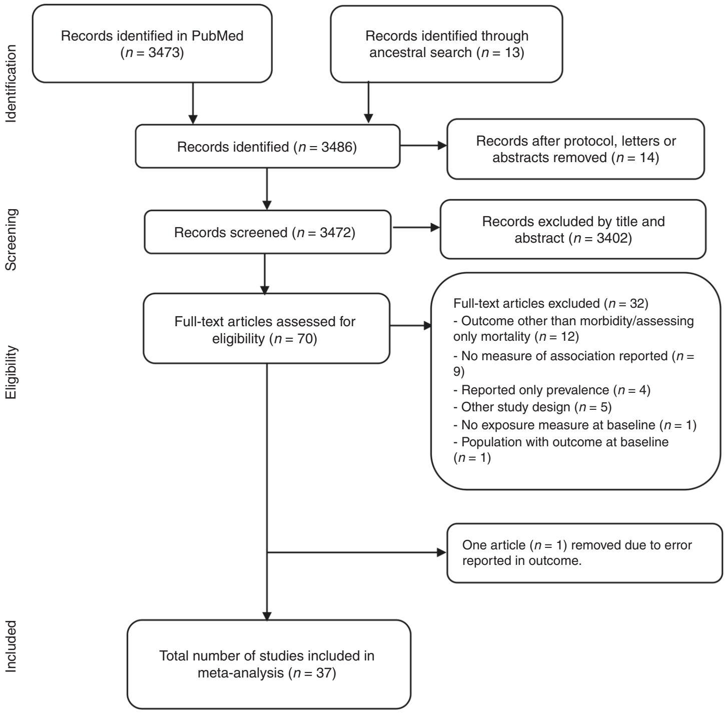

Identification
Records identified in PubMed (n = 3473)
Records identified through ancestral search (n = 13)
Records identified (n = 3486)
Records after protocol, letters or abstracts removed (n = 14)
Screening
Records screened (n = 3472)
Records excluded by title and abstract (n = 3402)
Eligibility
Full-text articles assessed for eligibility (n = 70)
Full-text articles excluded (n = 32)
- Outcome other than morbidity/assessing only mortality (n = 12)
- No measure of association reported (n = 9)
- Reported only prevalence (n = 4)
- Other study design (n = 5)
- No exposure measure at baseline (n = 1)
- Population with outcome at baseline (n = 1)
One article (n = 1) removed due to error reported in outcome.
Included
Total number of studies included in meta-analysis (n = 37)

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

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

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

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

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

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

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

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

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

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

MR-Egger is used to test for directional horizontal pleiotropy, a violation of the instrumental variable assumption where the effect of the instrumental variable on the outcome is mediated via another pathway other than the exposure of interest. MR-Egger tests for violation of IV assumptions and bias in the inverse variance-weighted (IVW) methods and includes the intercept as part of the regression (unlike IVW, where the intercept is forced to zero)[34]. The resulting coefficient, therefore, provides an asymptotically consistent estimate of the causal effect, even if all variants are pleiotropic with the outcome[35]. This holds when the Instrument Strength Independent of Direct Effect assumption is true, i.e., the instrument strength is independent of its pleiotropic effect. When this criterion is met, MR-Egger provides an unbiased assessment of the association between the exposure and outcome, providing the intercept, which provides the average pleiotropic effect, does not significantly differ from the null. When the intercept is significantly different from the null, it represents an estimate of the directional horizontal pleiotropic effect of the genetic variants[35]. The median-weighted method provides a reliable estimate of the causal association between exposure and outcome when at least half of the instrumental variables are valid[36].

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

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

Data availability

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

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

References

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

Acknowledgements

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

Author contributions

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

Funding

Open Access funding provided by Lund University.

Competing interests

The authors declare no competing interests.

Ethics approval

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

Additional information

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

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

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

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

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

Paper II

COMMUNICATIONS BIOLOGY

ARTICLE

https://doi.org/10.1038/s42003-020-01613-w OPEN

LRIG proteins regulate lipid metabolism via BMP signaling and affect the risk of type 2 diabetes

Carl Herdenberg[1], Pascal M. Mutie[2], Ola Billing[3], Ahmad Abdullah[1], Rona J. Strawbridge[4,5,6], Ingrid Dahlman[7], Simon Tuck[8], Camilla Holmlund[1], Peter Arner[7], Roger Henriksson[1], Paul W. Franks[2] & Håkan Hedman[1✉]

Leucine-rich repeats and immunoglobulin-like domains (LRIG) proteins have been implicated as regulators of growth factor signaling; however, the possible redundancy among mammalian LRIG1, LRIG2, and LRIG3 has hindered detailed elucidation of their physiological functions. Here, we show that *Lrig*-null mouse embryonic fibroblasts (MEFs) are deficient in adipogenesis and bone morphogenetic protein (BMP) signaling. In contrast, transforming growth factor-beta (TGF-β) and receptor tyrosine kinase (RTK) signaling appeared unaltered in *Lrig*-null cells. The BMP signaling defect was rescued by ectopic expression of LRIG1 or LRIG3 but not by expression of LRIG2. *Caenorhabditis elegans* with mutant *LRIG/sma-10* variants also exhibited a lipid storage defect. Human *LRIG1* variants were strongly associated with increased body mass index (BMI) yet protected against type 2 diabetes; these effects were likely mediated by altered adipocyte morphology. These results demonstrate that LRIG proteins function as evolutionarily conserved regulators of lipid metabolism and BMP signaling and have implications for human disease.

[1] Department of Radiation Sciences, Oncology, Umeå University, SE-90187 Umeå, Sweden. [2] Genetic and Molecular Epidemiology Unit, Department of Clinical Sciences, Lund University, Skane University Hospital, Malmo, Sweden. [3] Department of Surgical and Perioperative Sciences, Surgery, Umeå University, SE-90187 Umeå, Sweden. [4] Institute of Health and Wellbeing, University of Glasgow, Glasgow, UK. [5] Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden. [6] Health Data Research UK, University of Glasgow, Glasgow, UK. [7] Department of Medicine Huddinge, Karolinska Institutet, Stockholm, Sweden. [8] Umeå Center for Molecular Medicine, Umeå University, SE-90187 Umeå, Sweden. ✉email: hakan.hedman@umu.se

The mammalian leucine-rich repeats and immunoglobulin-like domains (LRIG/Lrig) protein family consists of three transmembrane proteins, LRIG1, LRIG2, and LRIG3[1]. The three mammalian LRIG paralogs appear to have both distinct and redundant functions, which is evident during mouse development[2]. Because *Lrig*-null (i.e., *Lrig1-/-;Lrig2-/-;Lrig3-/-*) mice are not viable, molecular investigations regarding the functions of the mammalian LRIG proteins have been hampered. Nevertheless, numerous reports have indicated that LRIG proteins are important etiological and prognostic factors in cancer[3,4]. In most cases, these roles have been attributed to the ability of LRIG1 to negatively regulate various receptor tyrosine kinases (RTKs)[5,6,7,8,9,10,11]. However, in the nematode *Caenorhabditis elegans* (*C. elegans*), the sole LRIG homolog, SMA-10, regulates body size by promoting bone morphogenetic protein (BMP) signaling[12,13]. Whether the mammalian LRIG proteins also regulate BMP signaling remains to be investigated.

The BMP signaling system is evolutionarily conserved and regulates major developmental and homeostatic processes[14–16]. The BMP families of cytokines and their receptors belong to the transforming growth factor-beta (TGF-β) and TGF-β receptor superfamilies, respectively. The human genome encodes at least 20 BMP ligands, three BMP type 1 receptors, and three type 2 receptors[15,17–19]. Upon ligand binding, the constitutively active type 2 receptor phosphorylates and activates an associated type 1 receptor, which, in turn, phosphorylates the downstream signaling mediators SMAD1, SMAD5, and SMAD8[15]. The phosphorylated SMAD1/5/8 complex then recruits the co-SMAD SMAD4 and is translocated into the nucleus, where it regulates expression of BMP-responsive genes[20]. In addition to the SMAD-mediated signaling pathway, BMPs may also initiate SMAD-independent signaling, including the activation of the MAP kinases ERK1/2, p38, and JNK[21,22]. Furthermore, BMP signaling is fine-tuned by regulatory proteins that either enhance or suppress signaling[15,20,23]. Although the BMP system has been extensively studied for decades, novel regulators and key signaling proteins may still await discovery.

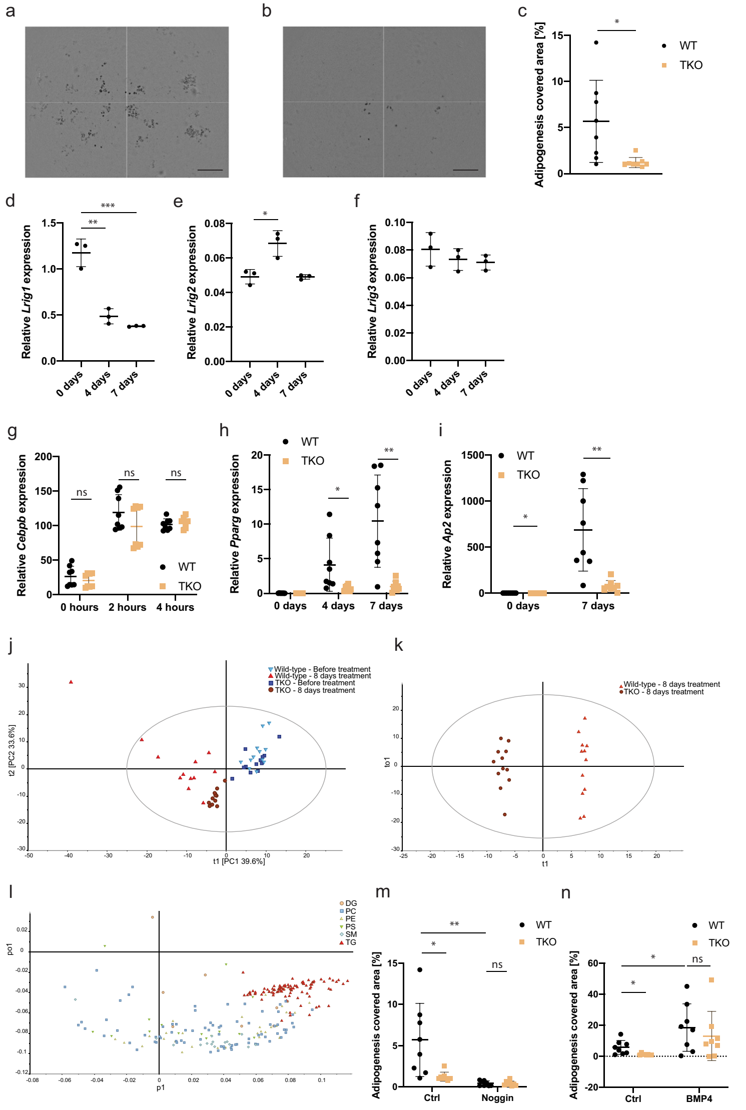
Obesity constitutes a global epidemic and is a major risk factor for several conditions, including insulin resistance, type 2 diabetes, heart disease, and several forms of cancer[24–26]. Adipose tissue serves as a key regulator of energy homeostasis in humans[27]. Adipose tissue can expand in volume either by enlarging adipocyte size (hypertrophy) or by adipocyte proliferation (hyperplasia). Of these two processes, adipocyte hypertrophy is associated with an unfavorable metabolic profile, whereas hyperplasia may improve metabolic homeostasis due to the increased number of insulin-sensitive cells[27,28]. Adipocyte differentiation involves the sequential commitment of mesenchymal stem cells to preadipocytes followed by their numerical expansion and terminal differentiation into adipocytes[29,30]. In this process, BMP signaling is involved in the commitment of mesenchymal stem cells into preadipocytes[30,31], as well as in the choice between white or brown/beige adipocyte differentiation and adipocyte size[32]. *C. elegans*, on the other hand, lacks dedicated adipocytes;[33] however, evidence suggests that BMP signaling may also regulate lipid accumulation in the lipid-storing intestinal cells of *C. elegans*[34,35]. In mice, *Lrig3*-deficient animals display altered plasma lipid levels[36]. However, a direct link between LRIG or SMA-10 proteins and adipogenesis, lipid metabolism, or type 2 diabetes has not yet been studied.

In the present study, we generated *Lrig*-null mouse embryonic fibroblasts (MEFs) to analyze the physiological and molecular functions of LRIG proteins in isogenic cells, without the possibly confounding expression of endogenous LRIG proteins. By exploiting these cells, we demonstrated that mammalian LRIG proteins regulate adipogenesis and sensitize cells to low concentrations of BMPs. We also analyzed the *sma-10/LRIG* mutant *C. elegans* and showed that LRIG also regulates fat accumulation in the worm. Finally, we investigated possible associations between *LRIG1* single nucleotide polymorphisms (SNPs) and human metabolic traits and revealed a striking discordant association between common *LRIG1* variants, a reduced risk of type 2 diabetes, and an increased body mass index (BMI), which we showed was likely mediated by adipocyte morphology.

Results

Generation and characterization of *Lrig*-null MEFs. To investigate the molecular functions of Lrig proteins, we generated *Lrig*-null cell lines by immortalizing MEFs that carried floxed *Lrig1*, *Lrig2*, and *Lrig3* alleles, followed by gene ablation via cell transduction with Cre recombinase-expressing adenoviruses. Thereby, we created four *Lrig*-null (herein also called *Lrig* triple knockout; TKO) MEF lines (TKO1-4) together with four corresponding wild-type control MEF lines (WT1-4). The stable ablation of *Lrig1*, *Lrig2*, and *Lrig3* was confirmed by polymerase chain reaction (PCR) genotyping (Supplementary Fig. 1a), Western blotting (Supplementary Fig. 1b), and a sensitive duplex droplet digital PCR (ddPCR) assay (Supplementary Table 1). With the ddPCR assay, the fraction of contaminating wild-type MEFs in the *Lrig*-null populations was analyzed by quantifying the relative gene copy numbers of the targeted *Lrig3* exon 1 versus a reference locus. This analysis showed that, in all the analyzed TKO MEF populations, more than 99.6% of the MEFs were *Lrig3*-negative (Supplementary Table 1). Importantly, long-term culturing of the TKO MEF lines for 60 days did not enrich for contaminating wild-type MEFs as assessed with any of the three genotyping methods (Supplementary Fig. 1; Supplementary Table 1). We then compared wild-type and *Lrig*-null MEF lines with regard to their proliferation and migration rates, morphology, and basic metabolic functions. The proliferation rates were similar between the wild-type and *Lrig*-null MEFs, both under standard cell culture conditions in 10% fetal bovine serum (FBS) (Supplementary Fig. 1c) and under proliferation-limiting FBS concentrations, although the *Lrig*-null MEFs showed a higher apparent proliferation rate than the wild-type MEFs specifically with 5% FBS (Supplementary Fig. 1d). The migratory rates of wild-type and *Lrig*-null MEFs were also similar, in both 10% FBS and in 0% FBS (Supplementary Fig. 1e). We were also unable to detect any apparent difference in cell morphology between the wild-type and *Lrig*-null MEF lines by light microscopy. Accordingly, flow cytometry analysis did not reveal any significant differences in forward or side scatter profiles between the wild-type and *Lrig*-null MEF lines (Supplementary Table 2). We then analyzed basic metabolic functions on a Seahorse XF analysis platform. These analyses did not reveal any significant difference between the wild-type and *Lrig*-null MEF lines with regard to their aerobic or anaerobic responses, as measured by the oxygen consumption rate (OCR) and extracellular acidification rate (ECAR), respectively (Supplementary Fig. 1f, g).

Lrig proteins promote adipogenesis in vitro. To investigate the role of Lrig proteins in adipogenesis in vitro, wild-type and *Lrig*-null MEF lines were treated with an adipogenic cocktail consisting of the glucocorticoid dexamethasone, the cAMP diesterase inhibitor 3-isobutyl-1-methylxanthine, insulin, and the PPARγ activator rosiglitazone. After nine days of treatment, adipocytic transformation was assessed by Oil Red O staining. Three out of four wild-type MEF lines were clearly able to transform into adipocytes in response to the adipogenic cocktail, whereas all the *Lrig*-null MEF lines studied showed impaired adipogenesis. However, because the adipogenic potential was highly variable



among the wild-type MEF lines, we included four additional biological replicates of both wild-type and *Lrig*-null lines in the analysis. Taken together, the analysis of the eight wild-type and eight *Lrig*-null MEF lines clearly showed that compared to wild-type MEFs, *Lrig*-null MEFs had deficient adipocyte differentiation (Fig. 1a–c). During the adipogenic process of wild-type MEFs, *Lrig1* was downregulated and *Lrig2* transiently upregulated, whereas *Lrig3* did not show any significant changes (Fig. 1d–f, Supplementary Fig. 2a–f). It has been reported that the induction of adipogenesis results in the rapid induction of *Cebpb* expression, followed by the induction of *Pparg* expression, and finally, in differentiated adipocytes, *Ap2* expression[37–40]. Gene expression analyses via quantitative reverse transcription-PCR (qRT-PCR) revealed that the induction of *Cebpb* was not diminished in *Lrig*-null MEFs compared to in wild-type MEFs (Fig. 1g). However, the induction of *Pparg* and *Ap2* was severely impaired in *Lrig*-null MEFs compared to in wild-type MEFs (Fig. 1h, i). Although Oil Red O staining is a well-established method to visualize triglyceride-containing adipocytes, we wanted to investigate the biochemical changes in lipid composition that were associated with adipogenesis in our experimental system. To this end, we performed lipid profiling of the MEFs by liquid chromatography coupled with tandem mass spectrometry prior to the adipogenic treatment as well as eight days after the induction of adipogenesis.

Fig. 1 Lrig proteins regulate adipogenesis of MEFs in vitro. Wild-type (WT) and *Lrig*-null (TKO) MEFs were treated with an adipogenic cocktail as described in the Methods section for the indicated times. **a**–**c** Adipogenesis of wild-type and *Lrig*-null MEF lines. Wild-type and *Lrig*-null MEFs were treated with the adipogenic cocktail for nine days followed by the quantification of adipocytes via Oil Red O staining. Shown are representative images of wild-type cells with 6% covered area (**a**) and *Lrig*-null cells with 1.7% covered area (**b**) (scale bar, 0.6 mm) and the quantifications of percentage area coverage for the Oil Red O stained biological replicates ($n = 8$ per genotype) (**c**). **d**–**i** Relative mRNA expression levels of *Lrig1* (**d**), *Lrig2* (**e**), *Lrig3* (**f**), *Cebpb* (**g**), *Pparg* (**h**), and *Ap2* (**i**). Cells were treated as in a-c. At the indicated time points after induction, the cells were lysed and analyzed by quantitative RT-PCR. Expression was normalized to the reference gene *Rn18s*. Shown in d-f are wild-type cells only, whereas both wild-type and *Lrig*-null cells are shown in g–i, as indicated, for eight biological replicates per genotype. **j**–**l** Lipidomic analyses. Lipids were extracted from wild-type and *Lrig*-null MEFs before or after eight days of treatment with the adipogenic cocktail. Lipid analysis was then performed by liquid chromatography coupled with tandem mass spectrometry. Each symbol represents one experimental replicate; shown are the results of three biological replicates per genotype with four experimental replicates each. **j** PCA score plots of all samples, labeled according to sample category. The variation explained by PC1 and PC2 was 39.6% and 33.6%, respectively. **k** The score plot of the OPLS-DA model built from the lipid profiles of the 8-day samples to determine the maximal variance between the wild-type and *Lrig*-null sample groups. **l** The corresponding loading plot explaining the contributions of different lipid species to the OPLS-DA model, indicating that triglycerides (TGs) (red triangles) in the wild-type samples were highly enriched compared to in the *Lrig*-null samples. The lipids are labeled according to lipid class (DG: diacylglycerol, PC: phosphatidylcholine, PE: phosphatidylethanolamine, PS: phosphatidylserine, SM: sphingomyelin, TG: triacylglycerol). **m**, **n** Role of BMP for adipogenesis in vitro. Wild-type and *Lrig*-null MEFs were treated as in a, without (Ctrl) or with the addition of 100 ng/ml of the BMP inhibitor noggin (Noggin) (**m**), or without (Ctrl) or with the addition of 50 ng/ml of BMP4 (BMP4) (**n**). Adipogenesis was scored through Oil Red O staining as described under **a**. In **c**–**i**, **m** and **n** the means of the biological replicates (**c**, **g**–**i**, **m** and **n**, $n = 8$ per genotype; **d**–**f**, $n = 3$) are shown by horizontal lines, and the means of the individual biological replicates analyzed by three experimental repeats are shown by dots and squares. Error bars represent the standard deviations of the means of the biological replicates. $^{ns}P > 0.05$, $^{*}P < 0.05$, $^{**}P < 0.01$ (Student's *t*-test).

In total, 244 putative lipids were quantified. A principal component analysis (PCA) of all samples did not reveal any apparent difference between the wild-type and *Lrig*-null MEFs prior to adipogenesis induction (Fig. 1j). Eight days after induction of adipogenesis, both the wild-type and the *Lrig*-null MEF lines separated distinctly from the untreated control MEFs. At this time point, separation was also evident between wild-type and *Lrig*-null MEFs. Thus, all the MEF lines responded to the adipogenic stimulus by altering their lipid composition; however, the wild-type and *Lrig*-null MEFs did so in different ways. In fact, a supervised orthogonal projections to latent structures discriminant analysis (OPLS-DA) plot completely separated the treated wild-type samples from the treated *Lrig*-null samples (Fig. 1k). The corresponding loadings plot, using the lipid classes, showed that the main contributors to the separation between the treated wild-type and treated *Lrig*-null samples were the triacylglycerides, of which the majority showed higher levels in wild-type MEFs than in *Lrig*-null MEFs (Fig. 1l).

TGF-β and BMP signaling pathways have been reported to play important roles in adipogenesis in vitro and in vivo[41]. Accordingly, the BMP inhibitor noggin was able to inhibit the adipogenesis cocktail-induced adipogenesis of wild-type MEFs (Fig. 1m). Conversely, a high dose of BMP4 (50 ng/ml) was able to greatly enhance the cocktail-induced adipogenesis rate of wild-type MEFs and, intriguingly, rescued the adipogenesis deficiency of the *Lrig*-null MEFs (Fig. 1n).

***Lrig*-null MEFs show impaired BMP signaling**. To investigate the role of the LRIG proteins in BMP signaling, we used a BMP-responsive element-driven luciferase reporter gene assay and analyzed the phosphorylation levels of Smad1/5 by fluorescent immunocytochemistry and Western blotting. First, we transiently transfected the wild-type and *Lrig*-null MEFs with the BMP reporter *pGL3-BRE-luciferase* and then stimulated them with different concentrations of BMP4. In this assay, the *Lrig*-null MEF lines showed a lower sensitivity to BMP4 than the wild-type MEF lines (Fig. 2a). Similarly, the pSmad1/5 analysis showed that the *Lrig*-null MEFs had a lower BMP4 sensitivity than the wild-type MEFs (Fig. 2b); however, the maximal pSmad1/5 response did not appear to differ between the wild-type and *Lrig*-null MEFs in this assay. Additionally, the *Lrig*-null MEFs showed a reduced sensitivity for BMP6 (Fig. 2c), whereas the sensitivity for BMP9/GDF2 was similar between the wild-type and the *Lrig*-null MEF lines (Fig. 2d). Compared with wild-type MEFs, *Lrig1* and *Lrig3* single-knockout MEF lines (Supplementary Fig. 2g, h) showed an apparently unaltered BMP4 sensitivity (Fig. 2e, f). BMPs are also able to activate noncanonical BMP signaling, which includes the activation of MAPK signaling cascades[21,22]. Intriguingly, the *Lrig*-null MEFs showed a reduced sensitivity for BMP4 when noncanonical phosphorylation of p38 was analyzed (Fig. 2g–i); however, the detection of increased p38 phosphorylation levels required higher BMP4 concentrations than the detection of increased pSmad1/5.

To investigate whether the BMP signaling deficiency of the *Lrig*-null MEF lines was the result of reduced expression of one or several of the BMP receptors, the BMP receptor levels were analyzed at the transcript level by RNA sequencing (RNAseq) and at the protein level by Western blotting. The RNAseq analysis revealed no significant difference in the levels of the different BMP receptor transcripts between the wild-type and *Lrig*-null MEF lines (Fig. 2j). Accordingly, Acvr1 and Bmpr2 showed similar protein expression levels in wild-type and *Lrig*-null MEFs when analyzed through Western blotting (Supplementary Fig. 4a–c). In addition, there were no significant differences in the transcript levels of Bmp ligands, signaling mediators, responsive genes, or Tgf-β receptors (Fig. 2j).

LRIG1 and LRIG3 rescue BMP signaling in *Lrig*-null MEFs. To investigate whether individual *LRIG* alleles could rescue the *Lrig*-null phenotype, an *Lrig*-null MEF line was transduced with the inducible human alleles *LRIG1*, *LRIG2*, or *LRIG3*, with an empty vector serving as a control (Supplementary Fig. 2i–k). As assessed by flow cytometry, a majority of the transduced cells expressed LRIG1 or LRIG3 after induction, whereas the lower expression level of LRIG2 made it difficult to determine the fraction of LRIG2-positive cells with this method (Supplementary Fig. 2l–o). Intriguingly, the induction of *LRIG1* or *LRIG3* expression rescued the canonical BMP sensitivity phenotype of the *Lrig*-null MEFs, whereas the induction of *LRIG2* expression, or vector control, did not (Fig. 3a–d). Interestingly, noncanonical BMP signaling through p38 and Jnk phosphorylation was only rescued by LRIG1 and not by LRIG3 (Fig. 3e, f, g–j). Increased phosphorylation of Erk was not observed under the BMP stimulation-protocol used (Fig. 3e, f, k, l). To investigate whether LRIG1 and LRIG3 could also rescue the adipogenesis deficiency of *Lrig*-null MEFs, *LRIG1*- and *LRIG3*-inducible MEFs were analyzed. Clearly, the induced

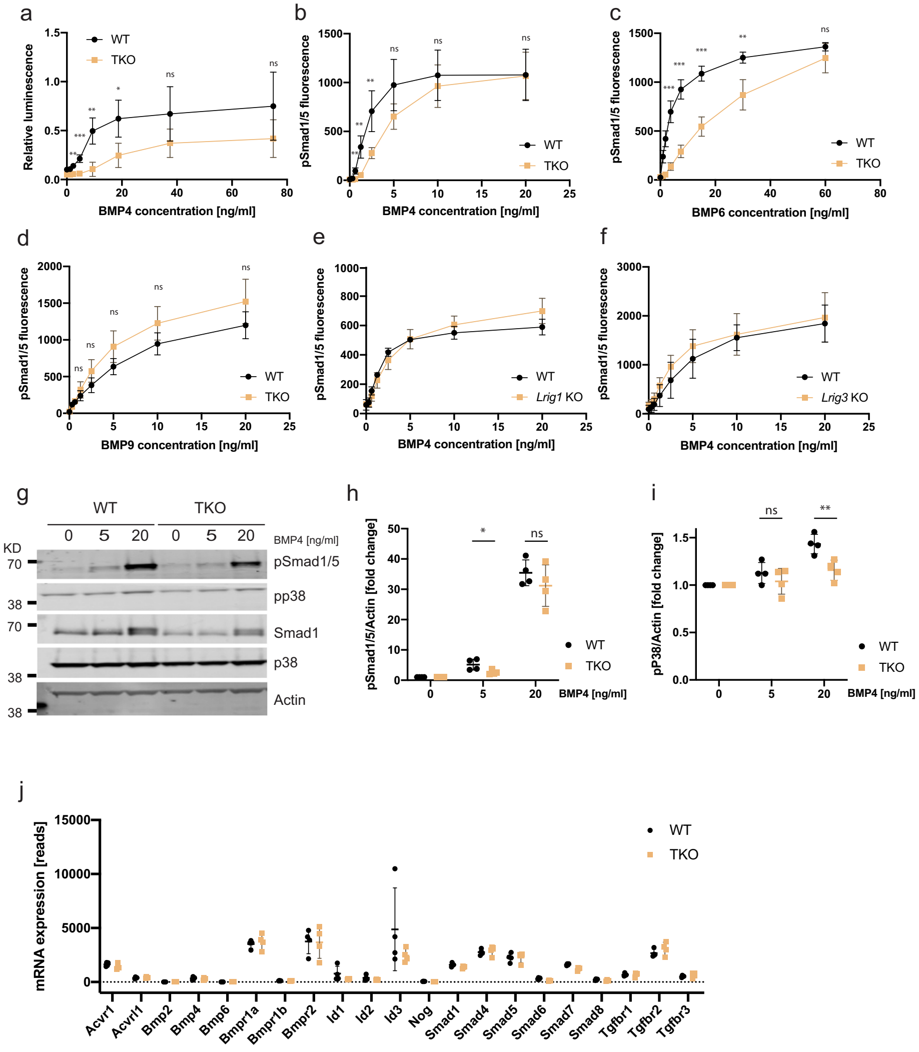
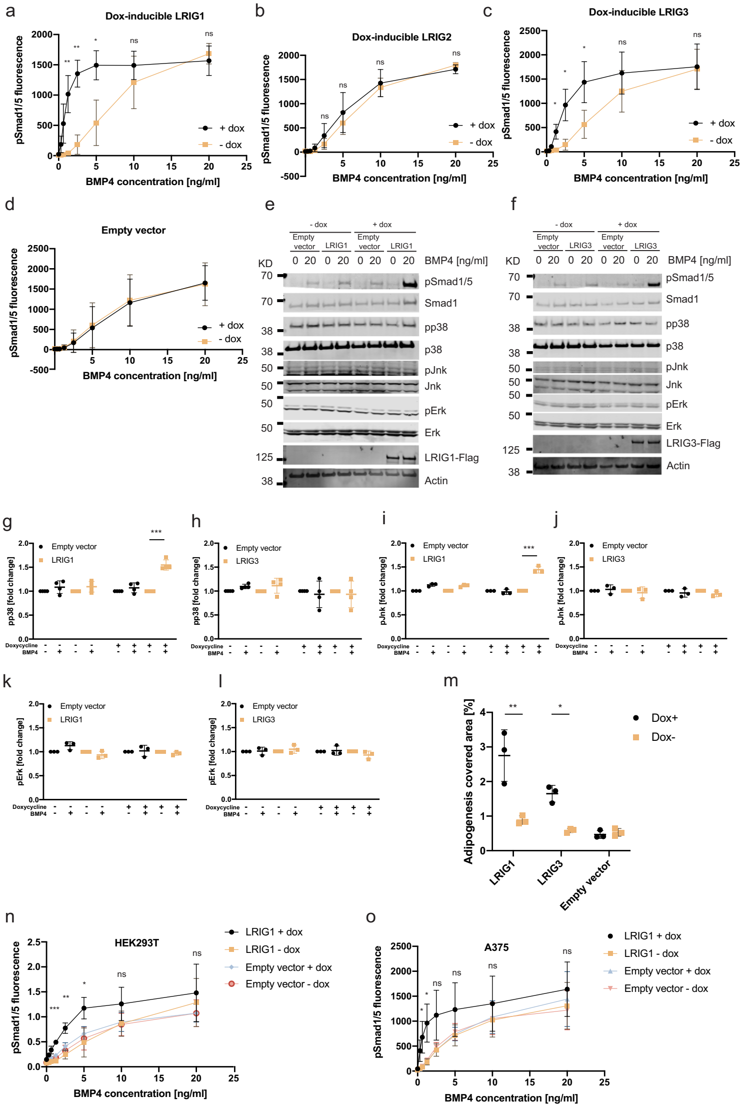


Fig. 2 *Lrig*-null MEFs show impaired BMP signaling without any apparent changes in the expression of receptors or signaling mediators. a Wild-type (WT) and *Lrig*-null (TKO) MEFs expressing the BMP reporter plasmid *pGL3-BRE-luciferase* were treated with the indicated concentrations of BMP4 for three hours. Thereafter, the cells were lysed, and the luciferase activity was analyzed and normalized to the control. **b**–**d** Wild-type and *Lrig*-null MEFs were stimulated with various concentrations of BMP4 (**b**), BMP6 (**c**), or BMP9 (**d**) for one hour followed by immunocytofluorescence analysis of nuclear phospho-Smad1/5 (pSmad1/5). **e**, **f** Wild-type and *Lrig1*-null MEFs (**e**) or wild-type and *Lrig3*-null MEFs (**f**) were stimulated with various concentrations of BMP4 for one hour followed by nuclear pSmad1/5 analysis. **g**–**i** Western blot analyses of canonical BMP4 signaling through pSmad1/5 and noncanonical BMP signaling through phosphorylated p38 (pp38). Wild-type and *Lrig*-null MEFs were stimulated with the indicated concentrations of BMP4 for one hour followed by cell lysis and Western blot analysis. Uncropped blots are shown in Supplementary Fig. 3. **g** Representative Western blots showing pSmad1/5, pp38, total Smad1, total p38, and the loading control actin. **h** Quantification of the pSmad1/5/actin ratios. **i** Quantification of the pp38/actin ratios. **j** Gene expression levels were analyzed in wild-type (WT) and *Lrig*-null (TKO) MEFs via RNA sequencing (RNAseq). The apparent number of RNAseq reads for respective gene is indicated. All the values in **a**–**f**, **h** and **i** represent the means of four biological replicates that were analyzed by three experimental repeats each. **j** The values represent the means of four biological replicates that were analyzed once. Error bars represent the standard deviations of means from four biological replicates. $^{ns}P > 0.05$, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$ (Student's *t*-test).



expression of LRIG1 or LRIG3 enhanced the adipogenesis rate of the MEF lines (Fig. 3m). To investigate whether the observed sensitizing effects of LRIG proteins on BMP signaling were restricted to MEFs, or if it was a more general phenomenon, the *LRIG1*-inducible human HEK293T and A375 cell lines were analyzed. Clearly, when *LRIG1* expression was induced in HEK293T or A375 cells, both cell lines showed a greatly enhanced sensitivity to low concentrations of BMP4 (Fig. 3n, o). To compare the BMP4-sensitizing potencies of the human LRIG proteins, different levels of LRIG1, LRIG2, or LRIG3 were induced by titrating the specific transcription-inducer doxycycline. Then, the correlations between the LRIG protein expression levels and the BMP4-induced pSmad1/5 levels were determined. This correlation analysis revealed that LRIG1 and LRIG3 were approximately equally potent in sensitizing the *Lrig*-null MEFs to BMP4; as expected, LRIG2 showed a negligible effect on BMP4 signaling (Fig. 4a). Next, we performed a structure-function analysis of the relationships between different protein domains and the BMP-sensitizing function of LRIG1. To this end, *Lrig*-null MEFs were transiently transfected with different amounts of expression vectors encoding a green fluorescent protein (GFP) -tagged full-length LRIG1, a GFP-tagged LRIG1 variant that lacked the cytosolic tail (LRIG1-Δcyto), or GFP only (as transfection control). Because of the poor transfection

Fig. 3 Ectopic LRIG1 or LRIG3 resensitize *Lrig*-null cells to BMP4. a–m *Lrig*-null MEFs were transduced with doxycycline-inducible *LRIG1*, *LRIG2*, or *LRIG3* constructs or with empty vector as a noninducible control. LRIG protein expression was not induced or induced by treatment of the cells with 100 ng/ml (**a**–**d**, **m**–**o**) or 500 ng/ml (**e**–**l**) doxycycline for 24 h followed by stimulation of the cells with different concentrations of BMP4 for one hour (**a**–**l**, **n**, **o**) or with adipogenic cocktail for ten days (**m**). **a**–**d** Immunofluorescence analyses of nuclear pSmad1/5 in cells not induced (−dox) or induced (+dox) to express *LRIG1* (**a**), *LRIG2* (**b**), or *LRIG3* (**c**). The empty vector served as a negative control for doxycycline treatment (**d**). **e**–**l** Western blot analyses of canonical (pSmad1/5) and noncanonical (pp38, pJnk, and pErk) BMP4 signaling. *LRIG1*- or *LRIG3*-inducible MEFs were induced, or not induced, with doxycycline followed by stimulation with 0 or 20 ng/ml of BMP4 for one hour. Thereafter, the cells were lysed, and the lysates were analyzed by Western blotting. **e, f** Representative Western blots showing pSmad1/5, total Smad1, pp38, total p38, pJnk, total Jnk, pErk, total Erk, LRIG1-FLAG (**e**), LRIG3-FLAG (**f**), and the loading control actin. Uncropped blots are shown in Supplementary Fig. 5. **g**–**l** Quantification of pp38 (**g**, **h**), pJnk (**I**, **j**), and pErk (**k**, **l**) normalized to actin in *LRIG1*-inducible (**g**, **I**, **k**) or *LRIG3*-inducible (**h**, **j**, **l**) MEFs. Plotted values in **a**–**d** represent means from three biological replicates, each with three experimental repeats. Error bars represent the standard deviations of means from three biological replicates. **g**, **h** Shown are four experimental repeats using an *LRIG1*- or *LRIG3*-inducible MEF line. Error bars show the standard deviations of the four means. **i**–**l** Shown are three experimental repeats using an *LRIG1*- or *LRIG3*-inducible MEF line. Error bars show standard deviations of the four means. **m** *LRIG1* or *LRIG3* expression was induced or not in *Lrig*-null MEFs with doxycycline followed by treatment of the cells with the adipogenic cocktail for ten days and quantification of adipocytes via Oil Red O staining. Shown are quantifications of the percentage area coverage for the Oil Red O stained biological replicates ($n = 3$ per genotype and treatment). **n**, **o** *LRIG1* expression was induced or not in *LRIG1*-inducible HEK293T cells (**n**) and A375 cells (**o**) via the treatment of cells with doxycycline for 24 h, followed by stimulation with different concentrations of BMP4 for one hour. Immunofluorescence analyses of nuclear pSmad1/5 in HEK293T cells (**n**) or A375 cells (**o**) that were not induced (-dox) or induced (+dox) to express *LRIG1*, with empty vector serving as a noninducible control. Plotted values in **n** and **o** represent the means from three independent experiments, performed as triplicates using one biological replicate of each cell line. Error bars represent standard deviations from three means. $^{ns}P > 0.05$, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ (Student's *t*-test).

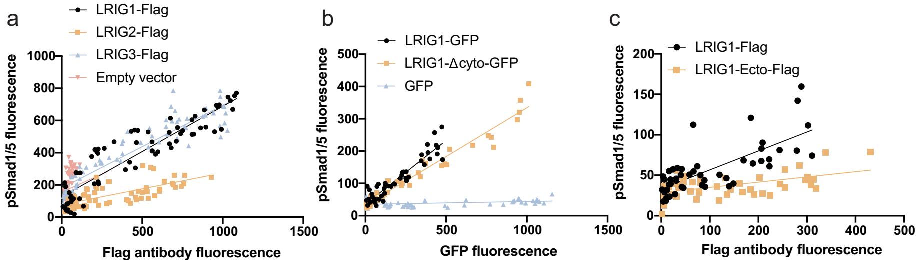


Fig. 4 LRIG1 and LRIG3 promote BMP signaling. a Different LRIG protein expression levels were induced in *LRIG*-inducible MEFs by treating the cells with different concentrations of the inducer doxycycline. Thereafter, the cells were stimulated with 2.5 ng/ml BMP4 for one hour followed by coimmunocytofluorescence analyses of nuclear pSmad1/5 and the FLAG-epitope present as a tag on the induced LRIG proteins. Correlation plots of phosphorylated Smad1/5 versus the FLAG-LRIG protein expression levels are shown. Fitted lines indicate the linear relationship with pSmad1/5 for each LRIG protein. Pearson's correlation coefficients for the respective genes were LRIG1, 0.9162; LRIG2, 0.6939; LRIG3, 0.9251; and empty vector, 0.1659. Shown are three experimental repeats using three biological replicates. **b** *Lrig*-null MEFs were transfected with a full-length LRIG1-GFP fusion protein (LRIG1-GFP), an LRIG1-GFP fusion protein variant lacking the cytosolic domain of LRIG1 (LRIG1-Δcyto-GFP), or empty vector (GFP) as a transfection control. Thereafter, the cells were stimulated with 20 ng/ml of BMP4 for 20 min followed by coimmunocytofluorescence analyses of nuclear pSmad1/5 and the green fluorescence from GFP fusion proteins or control GFP. The correlation plots between pSmad1/5 and GFP fluorescence are shown. The fitted lines indicate the linear relationship to pSmad1/5 for the respective construct. Pearson's correlation coefficients for the respective constructs were as follows: full length LRIG1, 0.8943; LRIG1-Δcyto, 0.9663; GFP control, 0.2564. Shown are two experimental repeats using three biological replicates. **c** *Lrig*-null MEFs were transfected with different amounts of expression vectors encoding FLAG-tagged full-length LRIG1 (LRIG1) or FLAG-tagged LRIG1 ectodomains (LRIG1-ecto). Thereafter, the cells were stimulated with 20 ng/ml of BMP4 for 20 min followed by coimmunocytofluorescence analyses of nuclear pSmad1/5 and the FLAG-epitope. Shown are the correlation plots between pSmad1/5 and FLAG-LRIG expression levels. Fitted lines indicate the linear relationship between pSmad1/5 and the respective FLAG-LRIG construct. Pearson's correlation coefficients for the respective constructs were as follows: full-length LRIG1, 0.7393; and LRIG1-ecto, 0.5287. Shown are two experimental repeats using three biological replicates.

efficiencies of our MEF lines, the background signals from the majority of nontransformed *Lrig*-null cells imposed an analytical problem. To resolve this problem, we changed the stimulation protocol from 2.5 ng/ml BMP4 for 60 min to 20 ng/ml BMP4 for 20 min. This modified protocol enabled us to monitor the pSmad1/5 signals among the minority of *LRIG1*-transformed MEFs while keeping the background signals from the majority of nontransformed *Lrig*-null MEFs to a minimum. By correlating the BMP4-induced pSmad1/5 responses with the expression levels of the transfected LRIG1 or LRIG1-Δcyto proteins, it was revealed that full-length LRIG1 and LRIG1-Δcyto were approximately equally potent in promoting BMP4 signaling in the *Lrig*-null MEFs (Fig. 4b). The BMP4-sensitizing function of full-length LRIG1 was also compared with the isolated ectodomain of LRIG1, that is, LRIG1 lacking its transmembrane and cytosolic domains. Apparently, the LRIG1 variant lacking the transmembrane and cytosolic domains lost its BMP4-sensitizing function (Fig. 4c). Thus, the cytosolic tail, but not the transmembrane domain, was dispensable for the BMP-sensitizing function of LRIG1 in the context studied.

TGF-β and RTK-MAPK signaling pathways appear to be Lrig-independent in MEFs. Because the TGF-β and BMP signaling pathways share many common features and because RTK signaling has been reported to be regulated by LRIG proteins, we

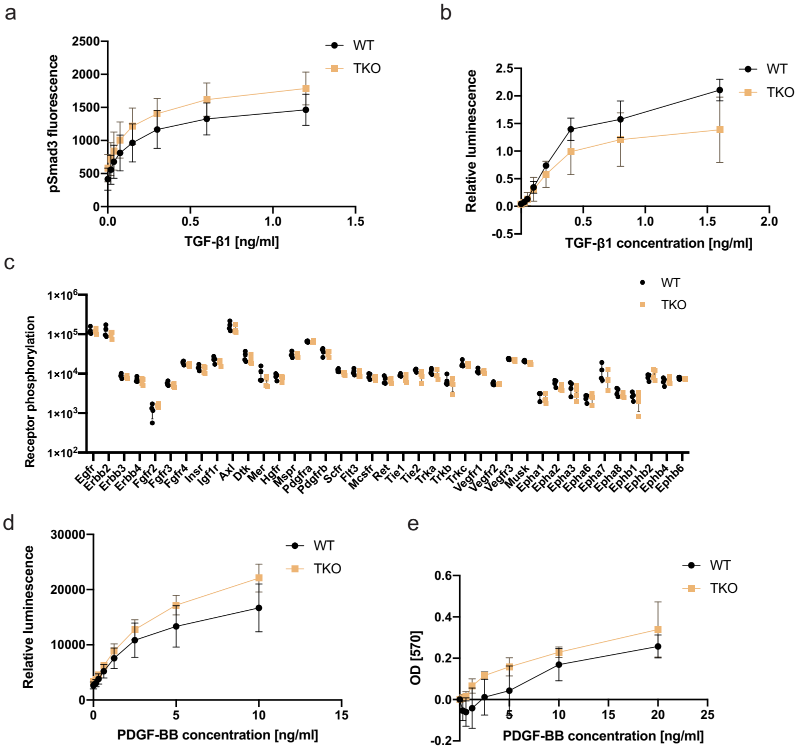


Fig. 5 TGF-β and RTK signaling appear unaltered in *Lrig*-null MEFs. a Wild-type (WT) and *Lrig*-null (TKO) MEFs were stimulated with various concentrations of TGF-β1 for one hour and then analyzed for nuclear phospho-Smad3 (pSmad3) by an immunocytofluorescence assay. **b** Wild-type and *Lrig*-null MEFs were transiently transfected with the TGF-β reporter plasmid *p(CAGA)*$_{12}$*MLP-Luc* followed by treatment of the cells with the indicated concentrations of TGF-β1 for three hours. Cell lysates were then analyzed for luciferase activity. Shown is the relative luminescence on an arbitrary scale. **c** Wild-type and *Lrig*-null MEFs, cultivated under standard cell culture conditions in 10% FBS, were lysed and analyzed for RTK phosphorylation levels with a phospho-RTK array kit (R&D Systems). **d** Wild-type and *Lrig*-null MEFs stably expressing the MAPK reporter gene *ELK1/SRF-luc* were treated with different concentrations of PDGF-BB for four hours and then lysed and analyzed for luciferase activity. **e** Wild-type and *Lrig*-null MEFs were cultivated in FBS-free medium containing different amounts of PDGF-BB for 48 h. Thereafter, the relative cell numbers were determined using an MTT assay. Shown are the OD values of treated MEFs normalized to those of untreated MEFs. **a**, **b**, **d** and **e** The plotted values represent the means of four biological replicates, each with three experimental repeats. **c** The plotted values represent the means of four biological replicates, each with one experimental repeat. Error bars represent the standard deviations of the means from four biological replicates.

also compared TGF-β and RTK signaling between wild-type and *Lrig*-null MEFs. TGF-β signaling was assessed using a TGF-β-responsive element-driven luciferase reporter, *p(CAGA)*$_{12}$*MLP-Luc*, assay[42] or by analyzing the phosphorylation levels of Smad3 by fluorescent immunocytochemistry. Neither of these analyses revealed any significant difference between wild-type and *Lrig*-null MEFs with regard to their TGF-β1 responses (Fig. 5a, b). To investigate the role of Lrig proteins in steady-state RTK signaling levels in MEFs under standard cell culture conditions with 10% FBS, a phospho-RTK array was used. Surprisingly, there was no apparent difference in the specific RTK phosphorylation levels between the wild-type and *Lrig*-null MEFs under the standard cell culture conditions employed (Fig. 5c). To further analyze the role of LRIG proteins in RTK signaling, a wild-type MEF line carrying floxed *Lrig* alleles was stably transduced with the MAPK reporter gene *ELK1/SRF-luc*. Thereafter, the stably transduced MAPK reporter MEFs were transduced with Cre recombinase or a control vector to generate four independent MAPK reporter *Lrig*-null MEF lines together with four MAPK reporter wild-type MEF lines (Supplementary Fig. 4d–f). When these MEF lines were treated with different concentrations of platelet-derived growth factor (PDGF) -BB, luciferase expression was induced; however, there was no apparent difference in the sensitivity to PDGF-BB between the wild-type and *Lrig*-null MEFs (Fig. 5d). Additionally, the dose-response of PDGF-BB-induced proliferation was similar between the wild-type and *Lrig*-null MEF lines (Fig. 5e).

The *LRIG* homolog *sma-10* regulates lipid metabolism in *C. elegans*. In *C. elegans*, LRIG/SMA-10 is reported to promote normal body size through BMP signaling[12]. Given that several BMP mutants have been reported to be defective in lipid homeostasis[34,35], we hypothesized that defective lipid homeostasis could be a general trait of BMP mutants, including

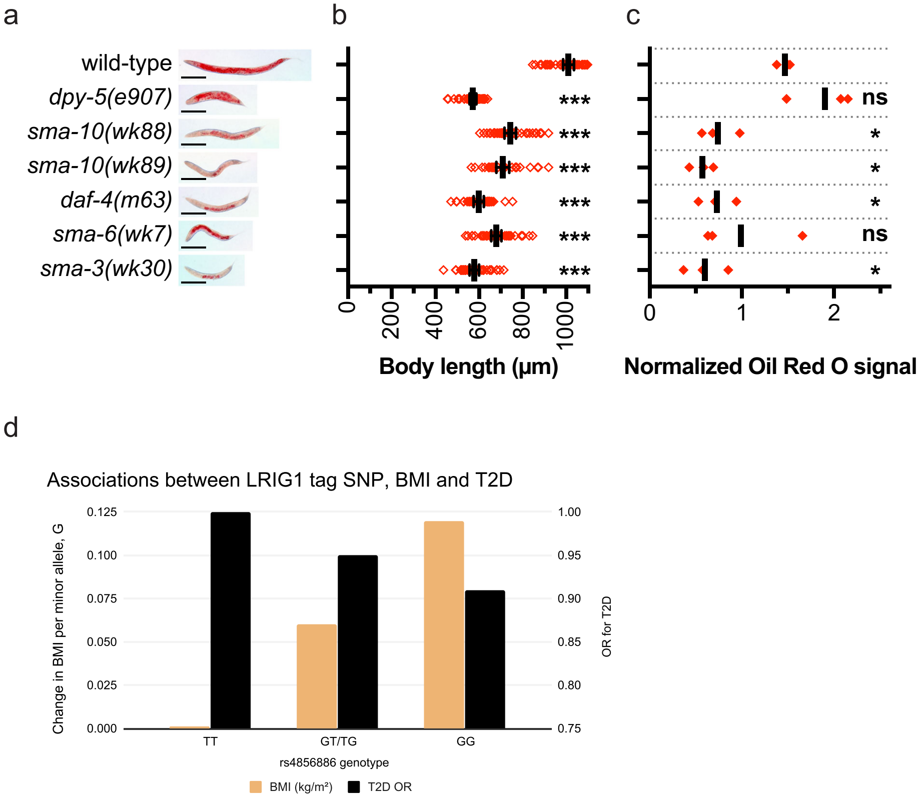


Fig. 6 *LRIG/sma-10* and other BMP pathway genes promote lipid accumulation in lipid-storing cells in *C. elegans*, and *LRIG1* SNPs predict BMI and type 2 diabetes in humans. a Representative whole-body images of Oil Red O-stained adult hermaphrodite worms. Scale bars, 200 µm. **b** Adult body lengths of wild-type animals ($n = 46$), *dpy-5(e907)* ($n = 37$), *sma-10(wk88)* ($n = 41$), *sma-10(wk89)* ($n = 32$), *daf-4(m63)* ($n = 32$), *sma-6(wk7)* ($n = 41$), and *sma-3 (wk30)* ($n = 37$). The body length of each individual animal is plotted as a red square. Solid lines and error bars indicate the means and 95% confidence intervals, respectively. The order of the dot plots, from top-to-bottom, is the same as that for the images in **a**. **c** Oil Red O signal intensities from three independent experiments. Each experiment was normalized to its combined mean signal intensity across all genotypes. For each genotype, solid, red squares indicate the mean normalized signal in each independent experiment. Solid lines indicate the combined means from three experiments. The order of the dot plots, from top-to-bottom, is the same as those for the images in a. Statistical significance versus wild-type was determined with multiplicity-adjusted *P*-values, calculated using Holm-Sidak multiple comparisons tests. $^*P < 0.05$, $^{***}P < 0.001$. **d** Plot illustrating the difference in predicted BMI (least square means, LSMs) across the genotypes of rs4856886 (minor allele = G, major allele = T) and odds ratio for type 2 diabetes (T2D) across genotypes. The *x*-axis represents the rs4856886 genotypes compared. The *y*-axis on the left represents the difference in BMI LSMs per single minor allele across the genotypes, while the y-axis on the right represents the odds ratios for T2D. In this study, the minor allele was associated with an increase in BMI and a lower odds of T2D risk.

mutants for *LRIG/sma-10*. As a BMP pathway-independent control for short body length, we included the cuticle collagen mutant *dpy-5(e907)* in our analysis[43]. Both body length and fat accumulation were assessed in wild-type and mutant worms (Fig. 6a–c). As expected, compared to wild-type worms, all of the *dpy-5*, *sma-10*, *daf-4*, *sma-6*, and *sma-3* mutant worms showed a reduced body length (Fig. 6a, b). Intriguingly, compared to lipid accumulation in wild-type worms, lipid accumulation was reduced in both the *LRIG/sma-10* mutants *wk88* and *wk89*, in the *ACVR2A* and *ACVR2B* homolog mutant *daf-4(m63)* and in the *SMAD1* homolog mutant *sma-3(wk30)* (Fig. 6a, c). Compared to the wild-type, also the *BMPR1A/sma-6(wk7)* mutants showed an apparent, although nonsignificant ($p = 0.14$), reduction in lipid accumulation. For unknown reasons, the *sma-6(wk7)* series showed a higher variance than the other BMP mutants, which might have contributed to the lack of significance for this series. Despite being short, the *dpy-5(e907)* mutants had normal levels of lipid deposits in lipid-storing intestinal cells. Short body length per se thereby appears uncoupled from fat accumulation in somatic tissue of *C. elegans*. Hence, we conclude that *LRIG/sma-10* promotes lipid accumulation in postmitotic tissue of *C. elegans*, likely through BMP signaling and independent of body size regulation.

Human *LRIG1* variants are associated with an altered risk of type 2 diabetes and with BMI and adipocyte morphology. To investigate possible associations between *LRIG* gene variants and human metabolism and metabolic disease, data from the UK Biobank were analyzed ($n = 398{,}810$, Supplementary Table 3 for participant characteristics). Here, we identified nine variants at *LRIG1* that were strongly associated with BMI ($P < 5 \times 10^{-8}$) (Supplementary Table 4); these signals have also been reported by others as a part of a larger meta-analysis conducted while this study was in progress[44]. In our analyses, each copy of the minor allele of *rs4856886* (G) (tag SNP with the strongest effect) increased the BMI by ~0.05 kg/m^2 (Fig. 6d); this phenomenon was largely attributed to genetic variation (Supplementary Table 5). These variants were also associated with a decreased risk of diabetes and adjusting for BMI strengthened these associations (Fig. 6d; Supplementary Table 6). Secondary signals for many other metabolic traits, including plasma triglyceride levels (Supplementary Table 7), were also observed, suggesting a metabolically favorable phenotype in people carrying the BMI-associated *LRIG1* alleles. The results of association with liver fat percentage were not significant (Supplementary Table 8), which might be attributable to the low statistical power given the relatively small sample size ($n = 3{,}192$, compared to $n = 398{,}810$ for other analyses). We hypothesized that the observed relationships may be attributed to a metabolically favorable adipose tissue phenotype, which is typically observed in adipose tissue comprised of many small adipocytes, i.e., hyperplastic adipose morphology[45]. Thus, for the two strongest *LRIG1* signals from UK Biobank (*rs4856886* and *rs9840088*), we tested associations with adipocyte size in adults from the GENiAL cohort ($n =$ 948, Supplementary Table 9). In these analyses, the BMI increasing (type 2 diabetes risk decreasing) alleles at *rs4856886* ($P = 0.039$) and *rs9840088* ($P = 0.014$) were associated with adipose hyperplasia. For *rs9840088*, the mean adipose morphology values were +7.5 picolitres for the common A allele and −9.2 picolitres for the minor C allele ($P = 0.026$ by analysis of covariance, adjusted for age).

Discussion

Lipid metabolism is central to energy homeostasis at both the cellular and organismal levels. Here, we found that the LRIG proteins function as regulators of lipid metabolism by regulating BMP signaling in several different biological systems, including adipocyte differentiation of mouse fibroblasts and in lipid accumulation in *C. elegans*; in addition, we found that human *LRIG1* gene variants were associated with a decreased risk of type 2 diabetes, an increased BMI, and altered adipocyte morphology. Collectively, these observations show that LRIG proteins function as evolutionarily conserved regulators of BMP signaling and lipid metabolism and have important implications for human metabolic health.

LRIG proteins regulated lipid accumulation and adipogenesis of MEFs in response to adipogenic stimuli. The former was shown by a reduced triglyceride accumulation, whereas the latter was suggested by a reduced number of equally sized Oil Red O-positive cells and impaired induction of the adipocyte markers *Pparg* and *Ap2* among the *Lrig*-null MEFs. Furthermore, the associations between *LRIG1* gene variants and adipocyte numbers in humans suggests that LRIG proteins might also regulate adipogenesis in vivo in humans. However, although the lipid profiles of undifferentiated MEFs were indistinguishable between the wild-type and *Lrig*-null cells, our experiments did not address whether LRIG proteins could regulate metabolism in differentiated adipocytes. In this regard, it was intriguing that the *LRIG/sma-10* mutant nematodes also showed a lipid accumulation defect, although *C. elegans* lack dedicated adipocytes. Thus, further investigations will reveal whether mammalian LRIG proteins, in addition to regulating adipogenesis, also regulate lipid metabolism in differentiated energy-regulating cells such as adipocytes, hepatocytes, or skeletal muscle cells.

The sole *C. elegans* LRIG homolog, SMA-10, regulates body size by regulating BMP signaling. Here, we showed that the BMP-promoting function of SMA-10 is conserved in human LRIG1 and LRIG3, which promoted BMP signaling by increasing signaling strength at low BMP4 and BMP6 ligand concentrations. These results show that the BMP signal-regulating function of the hypothesized common protein ancestor of nematode SMA-10 and mammalian LRIG proteins seems to be retained in human LRIG1 and LRIG3, but not in LRIG2. This finding is consistent with a recent whole genome CRISPR-Cas9 phenotypic screen for regulators of BMP signaling in HEK293 cells[46]. In this dataset, LRIG1 and LRIG3 were among the 170 significant activators of BMP2-induced signaling, whereas LRIG2 was not (https://orcs.thebiogrid.org/Screen/172). Our demonstration that the human LRIG1 cytosolic domain was dispensable for its BMP signal-promoting function is consistent with the fact that in contrast to the mammalian LRIG proteins, *C. elegans* SMA-10 lacks a prominent cytosolic domain but still promotes BMP signaling. Intriguingly, the mammalian LRIG proteins showed a striking specificity with regard to the BMP pathways that they regulated. Thus, BMP4 and BMP6 signaling was strongly dependent on LRIG proteins, whereas BMP9 signaling was not. This discrimination between the BMP4/6 and BMP9 pathways may indicate possible molecular targets for the mammalian LRIG proteins. BMP4, BMP6, and BMP9 share the same type 2 receptors; however, BMP4 and BMP6 specifically interact with the type 1 receptors BMPR1A (ALK-3) and BMPR1B (ALK-6), whereas BMP9 specifically interacts with ACTRL1 (ALK-1). Accordingly, it can be speculated that LRIG1 and LRIG3 may regulate the BMP type 1 receptors BMPR1A and/or BMPR1B but not the type 2 receptors or the type 1 receptor ACTRL1. Mammalian LRIG proteins specifically regulating type 1 BMP receptors would be in line with the demonstration that *C. elegans* SMA-10 is required for the proper trafficking of SMA-6, a type 1 BMP receptor, but not for the trafficking of DAF-4, the type 2 receptor regulating body size[13]. Curiously, we found that LRIG1, but not LRIG3, could rescue noncanonical BMP signaling through p38 and Jnk MAPKs. The molecular basis for these specific functions of LRIG proteins remains to be elucidated.

We further examined whether genetic variation in *LRIG1* was associated with risk of type 2 diabetes, BMI, and adipocyte morphology among humans. Intriguingly, we identified diabetes-preventing *LRIG1* alleles that were strongly associated with an increased BMI and hyperplastic adipose morphology, e.g. many small adipocytes given total body fat. All the identified SNPs were intronic, and their functional exploration in humans remains to be undertaken. However, because we also found that LRIG1 enhanced both BMP signaling and adipogenesis, it seems reasonable to speculate that the diabetes-preventing *LRIG1* alleles may stimulate adipogenesis by enhancing BMP signaling. The diabetes-protecting effect could, thus, result from a more efficient use of excess energy by the LRIG1-mediated increased number of adipocytes[47]. Further analyses of the associations between *LRIG1* gene variants and human metabolic traits may unravel further insights about the physiological function of LRIG1. Nevertheless, LRIG1, LRIG3, or other functionally associated proteins may provide novel targets for the prevention or treatment of type 2 diabetes or other metabolic diseases. However, the molecular mechanisms involved need to be elucidated before potential clinical applications can be explored.

We propose that LRIG proteins play important roles as BMP sensitizers in the context of lipid metabolism, during development, in tissue homeostasis, and in diseases such as cancer. In this regard, it is intriguing that *Lrig3*-deficient mice show both craniofacial and inner ear defects[48], which are consistent with the central role for balanced BMP signaling during the development of these anatomical structures[49,50]. Similarly, we suggest revisiting the *Lrig1*-deficient phenotypes, including the cutaneous[51,52] and intestinal[9,53] cell hyperproliferation phenotypes, to investigate the role of dysregulated BMP signaling in these processes. In cancer, the roles of LRIG1 and LRIG3 should be re-evaluated in light of their functions as BMP sensitizers. In glioma, for example, LRIG1 functions as a tumor suppressor[11], as does BMP signaling[54–56]. Thus, it can be hypothesized that LRIG1 may suppress glioma growth by enhancing BMP signaling, in addition to its previously proposed regulation of RTK signaling.

Future in vitro cell culture experiments concerning LRIG proteins need to take into consideration the BMP content in FBS. Commercial FBS, which is commonly used for in vitro cell cultivation, contains BMP ligands in concentrations ranging from 6 to 14 ng/ml[57,58], i.e., concentrations where the sensitizing functions of LRIG1 and LRIG3 were highly relevant. Thus, endogenously expressed LRIG proteins are likely to affect BMP signal transduction in cells cultured under standard conditions with FBS.

Lrig-null mice are not viable[2]. Nevertheless, the only clear phenotype that we could establish for the *Lrig*-null MEFs was impaired adipogenesis and a reduced sensitivity for BMP4 and BMP6. The *Lrig*-null MEFs did not show any obvious phenotype regarding their viability, morphology, proliferation, migration, energy metabolism, or signaling through TGF-β or RTKs. The lack of a detectable RTK phenotype was particularly intriguing given the substantial body of evidence showing that LRIG proteins regulate RTK signaling. Thus, compared to the wild-type MEFs, the *Lrig*-null MEFs showed no apparent alterations in their steady-state levels of phosphorylated RTKs, PDGF-driven cell proliferation, or PDGF-induced reporter gene activation. The reason for the apparent lack of an RTK phenotype in the *Lrig*-null cells remains enigmatic. Hypothetically, one possible explanation could be a canceling-out effect that occurs when knocking out of genes with opposing functions. For example, LRIG1 and LRIG3 have been shown to have opposing functions with regard to the regulation of ERBB RTKs[59]. It is also possible that the RTK-regulating functions of LRIG proteins become apparent only under specific conditions, such as those observed when signaling proteins are ectopically overexpressed, as has been the case in many of the previous studies[e.g.,5,6,7,8,10,11]. It may also be relevant to consider possible cross-talk between the BMP and RTK signaling pathways, i.e., a primary effect on one of the pathways may indirectly affect the other pathway and vice versa. In this regard, the timing of the different events will be important to resolve to shed light on their causal relationships. Nevertheless, the *Lrig*-null MEFs revealed that the LRIG proteins are not required for basal RTK signaling, at least not in MEFs under standard cell culture conditions.

In summary, we showed that mammalian LRIG proteins function as cellular BMP sensitizers and regulators of adipogenesis. Furthermore, we showed that the *C. elegans LRIG* homolog, *sma-10*, also regulates lipid accumulation in the worm. Importantly, specific human *LRIG1* gene variants were associated with a decreased risk of type 2 diabetes, increased BMI, and altered adipocyte morphology, suggesting that LRIG proteins play important physiological roles in the regulation of lipid homeostasis in humans. It will be important to further investigate the detailed molecular mechanisms involved, which could unravel new molecular players and treatment targets for common human metabolic diseases.

Methods

Cell lines and cell culture. MEFs were isolated from 12-day-old mouse embryos with floxed *Lrig* genes ($Lrig1^{flox/flox}$;$Lrig2^{flox/flox}$;$Lrig3^{flox/flox}$) that had been generated through interbreeding of the previously described mouse strains B6.129-Lrig1$^{\text{tm1Hhed}}$ [11], B6.129-Lrig2$^{\text{tm1Hhed}}$[60], and B6.129-Lrig3$^{\text{tm1Hhed}}$[36] or from embryos with wild-type or deficient *Lrig1* or *Lrig3* genes that had been obtained from inter crosses of B6.129-Lrig1$^{\text{tm1.1Hhed}}$[11] or B6.129-Lrig3$^{\text{tm1.1Hhed}}$ mice[36], respectively, in a C57BL/6 J genetic background. All mice were housed and maintained and all experiments performed in accordance with the European Communities Council Directive (86/609/EEC). Experimental protocols were approved by the Regional Ethics Committee of Umeå University, Umeå, Sweden (registration nos. A5-2010, A193-12, and A1-16). The cells were immortalized according to the 3T3 protocol described by Todaro and Green[61]. The MEFs were cultured in Dulbecco's modified Eagle's medium (DMEM) (Sigma-Aldrich Sweden AB, Stockholm, Sweden) supplemented with 10% FBS (Fisher Scientific GTF AB, Gothenburg, Sweden), MEM-nonessential amino acids (Fisher Scientific GTF AB), 50 μM 2-mercaptoethanol (Sigma-Aldrich Sweden AB), and 50 μg/ml gentamicin (Invitrogen, Fisher Scientific GTF AB). In experiments where MEFs were subjected to multiple washes, the cell culture plates were coated with 0.1% bovine gelatin (Sigma-Aldrich Sweden AB, catalog # G9391) for 30 min at 37 °C prior to cell seeding. To generate *Lrig*-null ($Lrig1^{-/-}$;$Lrig2^{-/-}$;$Lrig3^{-/-}$) cells, herein also referred to as TKO cells, the triple-floxed cells were transduced with adenovirus Ad(RGD)-GFP-iCre or its control adenovirus Ad(RGD)-GFP (Vector Biolabs, Malvern, PA, USA) at a multiplicity of infection of 100 and a cell seeding density of 10,300 cells/cm^2. Twenty-four hours after adenovirus transduction, the cells were washed with phosphate-buffered saline (PBS), and after an additional 24 h, the top 20% of cells that showed the highest green fluorescence intensity were isolated using a FACSAria III cell sorter (BD Biosciences, San Jose, CA, USA). The transduction-selection procedure was repeated independently four times, thereby producing four different cell line pairs composed of an *Lrig*-null and a wild-type MEF line named TKO1-4 and WT1-4, respectively. *LRIG*-inducible MEF lines were generated by stably transducing an *Lrig*-null MEF line with a doxycycline-inducible *LRIG1*, *LRIG2*, *LRIG3*, or empty control vector according to a previously described protocol[11]. The *LRIG1*-inducible human embryonic kidney cell line HEK293T, clone 32:3:10, has been described previously[62], and the human melanoma cell line A375 was obtained from Dr. Oskar Hemmingsson of Umeå University. A375 cells were profiled for short tandem repeats (STRs) by American Type Culture Collection (ATCC) and were confirmed to have a 100% match with the ATCC cell line CRL-1619 (A375). HEK293T cells and A375 cells were cultured in DMEM supplemented with 10% FBS and 50 μg/ml gentamicin. The cell culture plates used for HEK293T and A375 cells were coated with 10 μg/ml poly-D-lysine (Sigma-Aldrich Sweden AB, catalog # P0899) for 30 min at 37 °C followed by washes with PBS before the cells were plated. The *LRIG1*-deficient A375 subclone, clone Pc1-5-4, was generated via CRISPR-Cas9-mediated mutagenesis. To this end, two sgRNAs targeting both strands of *LRIG1* exon 11, were cloned into the *pD1401-AD* plasmid (Atum, Newark, CA, USA), which contains a Cas9(D10A)-GFP-nickase sequence under the CMV promoter. A375 cells were transfected with the resulting plasmid using Lipofectamine 2000 (Fisher Scientific GTF AB) according to the manufacturers protocol. GFP-positive cells were then single-sorted into 96-well plates containing DMEM supplemented with 10% FBS and 200 U/ml penicillin-streptomycin (Fisher Scientific GTF AB) using a BD FACSAria™ III sorter. Single cells were expanded, split, and expanded as duplicates in 6-well plates. One well in each duplicate was lysed and screened for large indels and insertions using PCR (forward primer sequence: 5′-CATTCCATGGGCTTGTGTTG-3′, reverse primer sequence: 5′-CCACTACCATTAATCAGAC-3′). Genomic DNA from a clone that lacked the 278-bp wild-type band was then PCR amplified using primers flanking *LRIG1* exon 11 (forward primer sequence: 5′-GTTTGACTCTAACTCTGTTG-3′, reverse primer sequence: 5′-GCATAATGCAATTGCAGAAG-3′). Each of the three resulting bands were purified and cloned into a TOPO vector (Fisher Scientific GTF AB), sequenced, and found to represent three different mutant variants of *LRIG1*: one with a deletion and one with an insertion, both resulting in frameshifts; the third had a silent intronic insertion, and both PAM sequences were intact. By repeating the entire mutational process on this clone, we isolated the Pc1-5-4 subclone, which was found to contain an additional insertion close to the splice acceptor site at the intron 10/exon 11 boundary. The *LRIG1*-inducible A375 cell line was generated through the cotransduction of the A375 clone Pc1-5-4 with the vectors *pLVX-LRIG1-TRE3G* and *pLVX-Tet3G* as described previously for other cells[11]. Lentiviral particles with vectors for the *srf/elk-1 luciferase* reporter and *Renilla* control were obtained from Qiagen AB (Sollentuna, Sweden, catalog nos. CLS-010L and CLS-RCL, respectively) and were used to cotransduce triple-floxed MEFs with 10 infection units (IU) per cell for *srf/elk-1* and 3.2 IU per cell for *Renilla*. Stably transduced MEFs were selected with puromycin. Thereafter, *Lrig*-null (TKO) and control (WT) MEF lines were generated independently four times from the puromycin-resistant MEFs through transduction of the cells with Ad(RGD)-GFP-iCre or Ad(RGD)-GFP as described above.

PCR and ddPCR genotype analyses. The mouse *Lrig* genotypes were routinely monitored via allele-specific PCRs using primers 5′-CATCGCATTGTCTGAG TAGGTGTC-3′ and 5′-CTCCAGAATCACGCTCACCT-3′, yielding an 824 bp product for the floxed wild-type *Lrig1* allele and no product for the knockout allele, primers 5′-TGCACTAGGCAGTCTTAAACCA-3′ and 5′-TCAGGCAGTGACA GAAGGTGT-3′, yielding a 450 bp product for the floxed wild-type *Lrig2* allele and no product for the knockout allele, and primers 5′-CATCGCATTGTCTGAG TAGGTGT-3′ and 5′-CGAGGCTGATGGTCTGCTAAT-3′, yielding a 630 bp product for the floxed wild-type *Lrig3* allele and no product for the knockout allele. The targeted exon 1 of *Lrig3* together with an untargeted region of *Lrig3* (used as the reference locus) were quantitated using a duplex ddPCR assay. The primers and probes for ddPCR were purchased from Integrated DNA Technologies (Leuven, Belgium). The ddPCR primers used were for *Lrig3* exon 1: 5′-CGCCTTCCCGATC CTCTC-3′ and 5′-GTCTCCTTCACCCCACCG-3′ and for the untargeted *Lrig3* locus: 5′-AACCGTCACCAAGGGAGA-3′ and 5′-CCACCAAAGGGCTGTCATC-3′. The probes used were for *Lrig3* exon 1: FAM-conjugated 5′-ATACTGATACT CACAGCCGTGTGACCCAGG-3′ and for the untargeted *Lrig3* locus, HEX-conjugated 5'-CATTGCTGGAGGGAGCCCGCCC-3′. The final concentrations of forward and reverse primers were 400 nM, and the final concentrations of the probes were 200 nM. DNA, ddPCR Supermix (with no dUTP) (Bio-Rad Laboratories AB, Stockholm, Sweden, catalog # 1863024), Hind III restriction enzyme (Fisher Scientific GTF AB, FastDigest, catalog # FD0505), and nuclease-free water were mixed with primer/probe sets of the targeted and untargeted regions of *Lrig3*. Droplets were generated using a QX200 droplet generator followed by PCR using a T100 thermal cycler (Bio-Rad Laboratories AB) with PCR parameters of 37°C for 5 min, 95 °C for 5 min; 40 cycles of 30 s at 95 °C and 1 min at 58 °C, followed by 98 °C for 10 min. After PCR amplification, the plate was loaded into the QX200 droplet reader (Bio-Rad Laboratories AB) to acquire the data. The data were analyzed using QuantaSoft software (Bio-Rad Laboratories AB, version 1.7.4.0917). Investigators were blinded to the cell line identities at the time of performing ddPCR and data analysis.

Western blot analysis. Cells were lysed for 30 min on ice with cell extraction buffer (Invitrogen, Fisher Scientific GTF AB) supplemented with cOmplete, EDTA-free Protease Inhibitor (Roche Diagnostics Scandinavia AB, Bromma, Sweden) and, when analyzing phosphorylated proteins, phosphatase inhibitor PhosSTOP (Roche Diagnostics Scandinavia AB). The lysates were then centrifuged at 20,800 x g for 10 min at 4°C. The resulting pellets were discarded. The protein concentrations of the cleared lysates were determined using a Pierce BCA Protein Assay Kit (Fisher Scientific GTF AB). The protein samples were separated through polyacrylamide gel electrophoresis using 3-8% Tris-acetate gels or 10% Bis-Tris gels (Invitrogen, Fisher Scientific GTF AB, catalog # EA03752 and NP0302, respectively) and then electrotransferred onto polyvinylidene fluoride or nitrocellulose membranes (Bio-Rad Laboratories AB). The membranes were then blocked with Odyssey blocking buffer (LI-COR Biosciences GmbH, Bad Homburg, Germany) or 5% fat-free milk in Tris-buffered saline with 0.1% Tween 20 (TBS-T). The blocked membranes were incubated at 4°C overnight with the primary antibodies at the indicated concentrations (Supplementary Table 10). After three washes with TBS-T, the membranes were incubated with the appropriate secondary antibodies for an hour at room temperature, followed by washes in TBS-T. Thereafter, immune-reactive bands were visualized and analyzed using the Odyssey CLx imaging system (LI-COR Biosciences GmbH) or ECL-select (GE Healthcare, Uppsala, Sweden) together with the ChemiDoc Touch Imaging System (Bio-Rad Laboratories AB). The primary and secondary antibodies used for Western blotting are listed in Supplementary Table 10.

Cell proliferation assays. Cell proliferation rates were determined by direct cell counting and an MTT assay. For cell counting, cells were seeded at a density of 2,800 cells per cm^2 in TC 6-well standard plates (Sarstedt AB, Helsingborg, Sweden). The cells were trypsinized at different times after seeding and counted via the use of a Countess Automated Cell Counter (Invitrogen, Fisher Scientific GTF AB). For the MTT-assay, cells were seeded at the same density in TC 96-well standard plates (Sarstedt AB). Twenty-four hours after the seeding, the medium was changed to cell culture medium containing different FBS or PDGF-BB concentrations. Thereafter, the cells were incubated for an additional 48 h followed by quantification of relative cell numbers via an MTT proliferation kit (Sigma-Aldrich Sweden AB) according to the manufacturer's instructions.

Cell migration assay. Migration assays were performed using Corning Transwell cell culture plates with 6.5 mm inserts of 8 µm pore size (Fisher Scientific, GTF AB). Five thousand cells were plated in the upper chamber with medium containing, or not containing, 10% FBS in the bottom chamber. Twenty-four hours after the plating, the membranes were washed with PBS, fixed in ice-cold methanol for 20 min, and stained with 0.1% crystal violet in 20% methanol for 20 min. Five fields from each chamber were counted manually using an Axio Vert.A1 inverted microscope (Carl Zeiss AB, Stockholm, Sweden) equipped with a 5x objective.

Flow cytometry. For flow cytometry analyses, cells were dissociated using Accutase cell detachment solution (Sigma-Aldrich, Sweden AB) and then washed in PBS containing 5% FBS. For analysis of intracellular antigens, cells were fixated in 4% phosphate-buffered formaldehyde for 10 min and then permeabilized using 0.2% saponin from Quillaja bark (Sigma-Aldrich, Sweden AB) in PBS for 10 min. The cells were labeled with primary antibodies for 30 min on ice, washed and then incubated with secondary antibodies for 30 min on ice. The primary and secondary antibodies used for flow cytometry analysis are listed in Supplementary Table 10. The flow cytometry analyses were performed on a BD Accuri C6 instrument (BD Biosciences).

Cell metabolism analyses. Cell metabolism was analyzed with a Seahorse XFe cell analyzer (Agilent Technologies, Inc., Santa Clara, CA, USA) using the mito stress test and glyco stress test assays according to the manufacturer's instructions. In the mito stress assay, the cells were sequentially treated with 1 µM oligomycin, 1 µM FCCP, and 0.5 µM rotenone and antimycin A. After each treatment, the OCR was measured at three time points. In the glycolytic stress test, the cells were glucose-starved for 1 h followed by sequential treatments with 10 mM glucose, 1 µM oligomycin, and 50 mM 2-deoxy-glucose. After each treatment, the ECAR was measured at three time points. The measurements were normalized to the relative cell numbers, which were determined through the measurement of cell nuclei fluorescence after staining with Hoechst 34580 (Sigma-Aldrich Sweden AB), using a Synergy2 microplate reader (BioTek Instruments SAS, Colmar Cedex, France).

In vitro adipogenesis assay. For adipogenic transformation, MEFs were seeded at a density of 28,000 cells/cm^2 at day −1. At day 0, cells were subjected to an initial adipogenic cocktail containing 1 µM dexamethasone (Sigma-Aldrich Sweden AB), 0.5 mM 3-isobutyl-1-methylxanthine (Sigma-Aldrich, Sweden AB), 10 µg/ml bovine insulin in HEPES buffer (Sigma-Aldrich Sweden AB), and 16 µg/ml rosiglitazone (Sigma-Aldrich, Sweden AB). At day 2, the medium was thereafter changed to a cocktail containing 10 µg/ml insulin and 16 µg/ml rosiglitazone and was changed every two days until day 9 when the cells were fixed with 4% formaldehyde (Unimedic Pharma AB, Stockholm, Sweden) for 30 min and stained using a 60% isopropanol solution with 0.5% Oil Red O (Sigma-Aldrich, Sweden AB). In some experiments, 100 ng/ml recombinant murine noggin (PeproTech Nordic, Stockholm, Sweden, catalog # 250-38) was added at day −1 or 50 ng/ml recombinant human BMP4 (PeproTech Nordic, catalog # 120-05ET) was added at day 0. The Oil Red O stained cells were quantified in a Spectramax i3x plate reader (Molecular Devices, San Jose, CA, USA) using the Softmax Pro 7 software (Molecular Devices).

RNA-extraction and quantitative RT-PCR-analyses. For qRT-PCR, RNA was prepared using a PureLink RNA Mini Kit (Invitrogen, Fisher Scientific GTF AB) followed by treatment with PureLink DNase (Invitrogen, Fisher Scientific GTF AB) according to the manufacturer's instructions. The TaqMan gene expression assays for *Lrig1* (Mm00456116_m1), *Lrig3* (Mm00622766_m1), *Cebpb* (mm00843434_s1), *Pparg* (mm00440940_m1), and *Fabp4* (mm00445878_m1) were purchased from Fisher Scientific GTF AB. Primers and probes for *Lrig2* and *RN18S* have been previously described[1]. Data were acquired using a CFX96 system C1000 thermal cycler (Bio-Rad Laboratories AB) as previously described[63]. The specific gene expression levels were normalized to that of *RN18S* by transforming the ΔCT values from log2 to linear values.

Lipidomics. One million cells were trypsinized, washed with PBS, and then frozen at −80 °C until use. Lipid extraction and liquid chromatography-quadrupole time-of-flight mass spectrometry-based lipidomics analysis was performed at the Swedish Metabolomics Centre at the Swedish University of Agricultural Sciences (Umeå) as previously described[64].

Luciferase reporter assays. Canonical BMP and TGFβ signaling was assessed by transiently transfecting the indicated MEFs with *pGL3-BRE-Luciferase*[65] (Addgene) or *p (CAGA)$_{12}$MLP-Luc*[42] (kindly provided by Serhiy Souchelnytskyi, Ludwig Institute for Cancer Research, Uppsala, Sweden), respectively. Transfections were performed with Fugene 6 transfection reagent (Promega Biotech AB, Nacka, Sweden) according to the manufacturer's instructions using a 1:3 DNA:reagent ratio, with a DNA amount corresponding to 0.7 µg/cm^2 and a reporter plasmid;*Renilla* reference plasmid ratio of 1:10. Twenty-four hours after transfection, cells were starved for one hour and treated with BMP4 or recombinant human TGF-β1 (PeproTech Nordic, catalog # 100-21) for three hours. PDGF signaling was assessed using the stably transduced *srf/elk-1 luciferase* MEF lines. Here, the cells were serum-starved for one hour followed by treatment with different concentrations of PDGF-BB (PeproTech Nordic, catalog # 315-18) for four hours. After the treatments with growth factors, cells were lysed using a Dual-Glo Luciferase Assay System (Promega Biotech AB) with a 20-minute incubation time for both assay reagents. Plates were analyzed with a Glomax 96 microplate luminometer (Promega Biotech AB) using an exposure time of 1 s per well. When analyzing transient BMP and TGFβ reporters, the data were normalized by taking the ratio of luciferase/Renilla. For the stably transduced *srf/elk-1* reporter cells, only luciferase was used.

Phospho-Smad immunofluorescence and Western blot assays. To analyze BMP- or TGFβ-induced phosphorylation of Smad1/5 and Smad3, respectively, cells were seeded the day before stimulation in a 96-well cell culture microplate (Greiner Bio-One International GmbH, Monroe, NC, USA, catalog # 655090) at densities of 3,000 cells per well for wild-type or *Lrig*-null MEFs, 1,800 cells per well for *LRIG*-inducible MEFs, and 10,000 cells per well for HEK293T and A375 cells. LRIG expression was induced in *LRIG*-inducible cells by treatment of the cells with 100 ng/ml, unless otherwise indicated, of doxycycline (Clontech Laboratories, Bio-Nordika Sweden AB, Stockholm, Sweden) for 24 hours prior to starvation. Cells were starved in serum-free cell culture medium for one hour and then stimulated with BMP4, recombinant human BMP6 (PeproTech Nordic, catalog # 120-06), recombinant human GDF2/BMP9 (PeproTech Nordic, catalog # 120-07), or TGF-β1 for one hour. Thereafter, the cells were fixed with 4% formaldehyde for 10 min, permeabilized with 0.2% saponin for 10 min, and blocked in blocking buffer composed of PBS, 0.1% Tween 20, and 5% FBS. After blocking, cells were incubated overnight with the appropriate primary antibody followed by washes and an incubation for one hour with the corresponding secondary fluorescent antibodies. For cell number normalization, 1 μg/ml Hoechst 33342 was added before analysis. Plates were imaged using a whole-well imaging device Trophos Plate Runner HD (Trophos/Dioscure, Marseille, France) with exposure time set to detect stained area. Images were analyzed using the Tina analysis package (Trophos/Dioscure) adjusting the threshold to remove background and rolling ball subtraction. Objects at specified sizes were detected, and their fluorescence was normalized using Hoechst 33342 nuclear staining by dividing the number of cells (for MEFs and A375 cells) or the mean cell fluorescence (for HEK293T cells). For Western blot analysis, cells were seeded into 6-well plates at a density of 10,344 cells/cm^2. Two days after seeding, cells were starved for one hour in serum-free medium and then stimulated with 5 or 20 ng/ml BMP4 or serum-free medium for one hour before lysis. The antibodies used for the phospho-Smad immunofluorescence and Western blot analyses are listed in Supplementary Table 10.

Transcriptomics. Transcriptomes were analyzed via RNAseq. To this end, 500,000 cells were serum-starved for 1 h prior to cell lysis. RNA was isolated using the Dynabeads mRNA DIRECT Purification Kit (Fisher Scientific GTF AB) according to the manufacturer's instructions. The purity and integrity of the RNA preparations were confirmed with an RNA 6000 Nano kit and an Agilent Bioanalyzer (Agilent Technologies). Sequencing was performed at SciLifeLab (Uppsala, Sweden) with an Ion Technology sequencer Ion Proton (Fisher Scientific GTF AB). Reads were aligned using STAR and bowtie2 software, and HTSeq was used to generate counts.

Correlations between LRIG1-GFP variants and BMP-induced phosphorylation of Smad1/5. Cells were transiently transfected using Fugene 6 as described above. *pLRIG1-GFP* encoding full-length LRIG1 fused to GFP has been described, previously[66]. *pLRIG1-Δcyto-GFP* encoding the extracellular/luminal and transmembrane parts, together with the first three cytosolic amino acids (YQT), of LRIG1 fused to GFP was generated by cloning the corresponding PCR-amplified *LRIG1* fragment into the *pEGFP-N1* (Clontech Laboratories) expression vector. PCR was used to generate *pLRIG1-3XFLAG* and *pLRIG1-ecto-3XFLAG* by amplifying the regions corresponding to the full length and the ectodomain of LRIG1, respectively, from an *LRIG1* cDNA (GenBank accession no. AF381545) and cloning these fragments into *p3XFLAG-CMV-13* (Sigma-Aldrich Sweden AB). The integrity of *pLRIG1-3XFLAG* and *pLRIG1-ecto-3XFLAG* were confirmed by DNA sequencing using a Big Dye Terminator v 3.1 cycle sequencing kit (Fisher Scientific GTF AB) and a 3730xl DNA analyzer (Fisher Scientific GTF AB). Twenty-four hours after transfection, Smad1/5 phosphorylation immunofluorescence assays were performed using 20 ng/ml BMP4 for 20 min. Both GFP fluorescence and pSmad1/5 immunofluorescence were quantified simultaneously using a Trophos plate runner.

RTK array. To compare the RTK phosphorylation levels, whole-cell lysates containing 150 μg protein were analyzed using a Human Phospho-RTK Array Kit (R&D Systems Europe Ltd., Abingdon, UK; catalog # ARY014) according to the manufacturer's instructions and quantified using ChemiDoc Touch Imaging System and Image lab software (Bio-Rad Laboratories AB).

***C. elegans* analyses**. All *C. elegans* strains used are described in WormBase (www.wormbase.org). N2 Bristol was used as the wild-type strain in all cases. Worms were maintained at 20 °C on standard nematode growth medium (NGM) agar and with *E. coli* OP50 as a food source. One-day-old adult worms from staged plates were stained with Oil Red O as described previously[67] and imaged at a midplane, with the mouth and the pharyngeal lumen in focus, using the 10x objective on a DIC-equipped Olympus BX51 microscope. Color images were then subtracted for background (rolling ball radius: 50.0 pixels) and converted to 8-bit CIELAB using Fiji software[68]. Quantifications were performed in the "a" channel by selecting an area averaging ~2,500 μm^2 between the posterior end of the pharynx and the anterior border of the gonad to avoid the signal coming from the oocytes. The local background signal for each measurement was then subtracted. All genotypes under study were analyzed in parallel, and each experimental round was normalized to the combined mean signal, which was calculated from all genotype means.

Statistics and reproducibility of cell and animal experiments. All Student's *t*-tests were 2-sided. All statistical analyses of cell and animal experiments were performed using GraphPad Prism 8 software (La Jolla, CA, USA), and the *P*-values <0.05 were considered significant. When cell lines were compared, in general, data were obtained from at least four biological replicates per genotype and three experimental repeats performed on separate days. The exact number of replicates are presented in the individual figure legends.

UK Biobank population characteristics. The UK Biobank is a large project with genotyped and well-phenotyped individuals comprising approximately 500,000 participants[69]. In this study, we excluded participants who did not have BMI measurements or of any of the other outcome variables of interest (done at the stage of that particular analysis) and those who had ambiguous information on sex (discordance between self-reported and genetically encoded sex). The final sample size was 398,810 participants of Caucasian ancestry. Supplementary Table 3 shows the participants' characteristics, and Supplementary Table 4 shows the *LRIG1* tag SNP information. This study was conducted using publicly available data from the UK Biobank, and therefore the current analyses did not require specific ethical approval. The reference for the approved UK Biobank project is ukb18274.

Tag SNP identification. We used *snptag*, an online tool of SNPinfo, (https://snpinfo.niehs.nih.gov/snpinfo/snptag.html) to identify tag SNPs from the nine *LRIG1* SNPs that were significant in the genome-wide association (GWA) analysis. In our tag SNP selection, the population was restricted to the Utah residents with Northern and Western European ancestry from the CEPH collection (CEU); the linkage disequilibrium (LD) threshold was set at r2 = 0.8, the maximum distance (bp) between SNPs for calculating the LD was 250,000 bp, and the minor allele frequency (MAF) range was 0.01 to 0.5. Three tag SNPs and one non-synonymous SNP were identified from the search (Supplementary Fig. 6).

Association with type 2 diabetes and BMI. To isolate the effect of BMI attributed to genetic variance, we regressed out the effects of age, age squared, sex, batch effect, and the first ten genetic principal components separately for men and women and extracted the BMI residuals. The residuals were then used in subsequent analyses as (a) untransformed residuals, and, (b) as inverse-normal transformed residuals. Each case is indicated in the respective analyses. We used these transformed residuals to create interaction terms with each of the tag SNPs and investigated the associations between these interaction terms and the risk of type 2 diabetes for each of the SNPs using logistic regression models adjusted for age, batch effect (array type used for genotyping, UK BiLEVE or UKBB Axiom) and the first ten genetic principal components. Type 2 diabetes was diagnosed by a doctor in this population and was parameterized as a binary outcome, coded "Yes" or "No", for these analyses. We also tested the association of each of the SNPs with type 2 diabetes using logistic regression models adjusted for untransformed BMI residuals and another model without BMI, in addition to the covariates mentioned above in the first model. We investigated the relationship between BMI and each of the SNPs using a simple linear regression model with untransformed BMI residuals as the outcome. All analyses were based on an additive genetic model.

Association with liver fat and biochemical measures of adiposity. To investigate the relationship between these tag SNPs and liver fat percentage (LF%) among 3,858 UK Biobank participants who had liver fat measurements, we extracted LF% residuals separately in men and women using multiple linear regression models adjusted for age, age squared, sex, BMI, array batch, and the first ten genetic principal components. The residuals were then inverse-normal transformed, and a simple linear regression model was fitted to test the association between each of the SNPs with LF% residuals as an outcome. For each of the two biochemical measures (baseline total cholesterol (mmol/L) and triglyceride levels (mmol/L)), we extracted residuals for participants with the relevant measures as outlined above. Cholesterol was normally distributed, so the residuals were not further transformed, but those for triglycerides were inversely normal transformed. The association between each of the measures and each of the tag SNPs was modeled using a simple linear regression model with the respective residuals as the outcome. In all analyses, the minor allele was used as the reference allele.

GENiAL study participants. The GENetics of Adipocyte Lipolysis (GENiAL) cohort included 273 men and 718 women and have been described previously[70]. Briefly, subjects in the GENiAL cohort were recruited by local advertisement to examine the regulation of fat cell function. Fifty-seven percent of the participants were obese (defined as BMI ≥ 30 kg/m^2). They all lived in Stockholm County, Sweden and were at least second-generation Swedes. One hundred ninety-four participants had type 2 diabetes, hypertension, or dyslipidemia alone or in different combinations. None were treated with insulin, glitazones, or glucagon-like-peptide analogs. Data on clinical variables are summarized elsewhere[70]. The study was approved by the local ethics committee at the Huddinge University Hospital (D. no. 167/02, 2002-06-03) and was explained in detail to each participant. Informed consent was obtained from all participants. Included in this study were 948 subjects from the GENiAL cohort with adipose morphology data available[70].

Clinical examination. The GENiAL participants came to the hospital's clinical research center the morning after an overnight fast. Their heights, weights, and waist-to-hip ratios (WHRs) were measured. Each participant's body fat content was measured by bioimpedance, and their total body fat mass was indirectly calculated using a formula based on age, sex and BMI[71]. A venous blood sample was obtained for extraction of DNA and clinical chemistry by the hospital's accredited routine clinical chemistry laboratory. Subcutaneous adipose tissue (SAT) was obtained by a needle aspiration biopsy lateral to the umbilicus as previously described[72].

Adipose tissue phenotyping. SAT samples were rapidly rinsed in sodium chloride (9 mg/ml) before removal of visual blood vessels and cell debris and were subsequently subjected to collagenase treatment to obtain isolated adipocytes as described[73]. The mean weight and volume of remaining cells were determined as described[74]. A curve fit of the relationship between mean adipocyte volume and estimated abdominal subcutaneous fat mass was performed[75]. The difference between the measured and expected mean adipocyte volume at the corresponding total fat mass determines adipose morphology. If the measured adipocyte volume is larger than expected, SAT hypertrophy prevails, whereas the opposite is true for hyperplasia. Thus, this measure of adipose morphology is independent of total fat mass.

Genetic analysis of the GENiAL cohort. The genetic analysis of the GENiAL cohort has been described previously[70]. After quality control, 894 samples were available for analysis. Genetic association analysis was conducted in PLINK[76], using linear regression, assuming an additive genetic model, and adjusting for population structure (PCs1-3), age, and sex.

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

Data availability

The transcriptomic RNAseq datasets generated and analyzed during the current study are available in the NCBI BioProject repository, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA684140. Relevant data and/or materials not present are available upon reasonable request from C.He. (carl.herdenberg@umu.se) and/or H.H. (hakan.hedman@umu.se). Source data from the UK Biobank are available upon approved application through the UKBB data application management system. Source data for the main figures are presented in Supplementary Data 1.

Received: 14 March 2020; Accepted: 17 December 2020;
Published online: 19 January 2021

References

1. Guo, D., Holmlund, C., Henriksson, R. & Hedman, H. The LRIG gene family has three vertebrate paralogs widely expressed in human and mouse tissues and a homolog in Ascidiacea. *Genomics* **84**, 157–165 (2004).
2. Del Rio, T., Nishitani, A. M., Yu, W. M. & Goodrich, L. V. In vivo analysis of Lrig genes reveals redundant and independent functions in the inner ear. *PLoS Genet.* **9**, e1003824 (2013). 2.
3. Wang, Y., Poulin, E. J. & Coffey, R. J. LRIG1 is a triple threat: ERBB negative regulator, intestinal stem cell marker and tumour suppressor. *Br. J. Cancer* **108**, 1765–1770 (2013).
4. Lindquist, D., Kvarnbrink, S., Henriksson, R. & Hedman, H. LRIG and cancer prognosis. *Acta Oncol.* **53**, 1135–1142 (2014).
5. Gur, G. et al. LRIG1 restricts growth factor signaling by enhancing receptor ubiquitylation and degradation. *EMBO J.* **23**, 3270–3281 (2004).
6. Laederich, M. B. et al. The leucine-rich repeat protein LRIG1 is a negative regulator of ErbB family receptor tyrosine kinases. *J. Biol. Chem.* **279**, 47050–47056 (2004).
7. Stutz, M. A., Shattuck, D. L., Laederich, M. B., Carraway, K. L. 3rd & Sweeney, C. LRIG1 negatively regulates the oncogenic EGF receptor mutant EGFRvIII. *Oncogene* **27**, 5741–5752 (2008).
8. Li, F., Ye, Z. Q., Guo, D. S. & Yang, W. M. Suppression of bladder cancer cell tumorigenicity in an athymic mouse model by adenoviral vector-mediated transfer of LRIG1. *Oncol. Rep.* **26**, 439–446 (2011).
9. Powell, A. E. et al. The pan-ErbB negative regulator Lrig1 is an intestinal stem cell marker that functions as a tumor suppressor. *Cell* **149**, 146–158 (2012).
10. Sheu, J. J. et al. LRIG1 modulates aggressiveness of head and neck cancers by regulating EGFR-MAPK-SPHK1 signaling and extracellular matrix remodeling. *Oncogene* **33**, 1375–1384 (2014).
11. Mao, F. et al. Lrig1 is a haploinsufficient tumor suppressor gene in malignant glioma. *Oncogenesis* **7**, 13 (2018).
12. Gumienny, T. L. et al. Caenorhabditis elegans SMA-10/LRIG is a conserved transmembrane protein that enhances bone morphogenetic protein signaling. *PLoS Genet.* **6**, e1000963 (2010).
13. Gleason, R. J. et al. C. elegans SMA-10 regulates BMP receptor trafficking. *PLoS ONE* **12**, e0180681 (2017).
14. Wang, R. N. et al. Bone Morphogenetic Protein (BMP) signaling in development and human diseases. *Genes Dis.* **1**, 87–105 (2014).
15. Heldin, C. H. & Moustakas, A. Signaling receptors for TGF-β family members. *Cold Spring Harb. Perspect. Biol.* **8**, a022053 (2016).
16. Savage-Dunn, C. & Padgett, R. W. The TGF-β family in *Caenorhabditis elegans*. *Cold Spring Harb. Perspect. Biol.* **9**, a022178 (2017).
17. Bragdon, B. et al. Bone morphogenetic proteins: a critical review. *Cell Signal.* **23**, 609–620 (2011).
18. Rahman, M. S., Akhtar, N., Jamil, H. M., Banik, R. S. & Asaduzzaman, S. M. TGF-β/BMP signaling and other molecular events: regulation of osteoblastogenesis and bone formation. *Bone Res.* **3**, 15005 (2015).
19. Yadin, D., Knaus, P. & Mueller, T. D. Structural insights into BMP receptors: specificity, activation and inhibition. *Cytokine Growth Factor Rev.* **27**, 13–34 (2016).
20. Sieber, C., Kopf, J., Hiepen, C. & Knaus, P. Recent advances in BMP receptor signaling. *Cytokine Growth Factor Rev.* **20**, 343–355 (2009).
21. Kozawa, O., Hatakeyama, D. & Uematsu, T. Divergent regulation by p44/p42 MAP kinase and p38 MAP kinase of bone morphogenetic protein-4-stimulated osteocalcin synthesis in osteoblasts. *J. Cell Biochem.* **84**, 583–589 (2002).
22. Broege, A. et al. Bone morphogenetic proteins signal via SMAD and mitogen-activated protein (MAP) kinase pathways at distinct times during osteoclastogenesis. *J. Biol. Chem.* **288**, 37230–37240 (2013).
23. Brazil, D. P., Church, R. H., Surae, S., Godson, C. & Martin, F. BMP signalling: agony and antagony in the family. *Trends Cell Biol.* **25**, 249–264 (2015).
24. Wang, Y., Rimm, E. B., Stampfer, M. J., Willett, W. C. & Hu, F. B. Comparison of abdominal adiposity and overall obesity in predicting risk of type 2 diabetes among men. *Am. J. Clin. Nutr.* **81**, 555–563 (2005).
25. Twig, G. et al. Body-mass index in 2.3 million adolescents and cardiovascular death in adulthood. *N. Engl. J. Med.* **374**, 2430–2440 (2016).
26. Pearson-Stuttard, J. et al. Worldwide burden of cancer attributable to diabetes and high body-mass index: a comparative risk assessment. *Lancet Diabetes Endocrinol.* **6**, e6–e15 (2018).
27. Choe, S. S., Huh, J. Y., Hwang, I. J., Kim, J. I. & Kim, J. B. Adipose tissue remodeling: its role in energy metabolism and metabolic disorders. *Front Endocrinol.* **7**, 30 (2016).
28. Hammarstedt, A., Gogg, S., Hedjazifar, S., Nerstedt, A. & Smith, U. Impaired adipogenesis and dysfunctional adipose tissue in human hypertrophic obesity. *Physiol. Rev.* **98**, 1911–1941 (2018).
29. Cristancho, A. G. & Lazar, M. A. Forming functional fat: a growing understanding of adipocyte differentiation. *Nat. Rev. Mol. Cell Biol.* **12**, 722–734 (2011).
30. Tang, Q. Q. & Lane, M. D. Adipogenesis: from stem cell to adipocyte. *Annu Rev. Biochem.* **81**, 715–736 (2012).
31. Huang, H. et al. BMP signaling pathway is required for commitment of C3H10T1/2 pluripotent stem cells to the adipocyte lineage. *Proc. Natl Acad. Sci. USA* **106**, 12670–12675 (2009).
32. Gustafson, B. et al. BMP antagonists regulate human white and beige adipogenesis. *Diabetes* **64**, 1670–1681 (2015).
33. McKay, R. M., McKay, J. P., Avery, L. & Graff, J. M. *C. elegans*: a model for exploring the genetics of fat storage. *Dev. Cell.* **4**, 131–142 (2003).
34. Yu, Y., Mutlu, A. S., Liu, H. & Wang, M. C. High-throughput screens using photo-highlighting discover BMP signaling in mitochondrial lipid oxidation. *Nat. Commun.* **8**, 865 (2017).
35. Clark, J. F., Meade, M., Ranepura, G., Hall, D. H. & Savage-Dunn, C. *Caenorhabditis elegans* DBL-1/BMP regulates lipid accumulation via interaction with insulin signaling. *G3 (Bethesda).* **8**, 343–351 (2018).
36. Hellström, M. et al. Cardiac hypertrophy and decreased high-density lipoprotein cholesterol in Lrig3-deficient mice. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **310**, R1045–R1052 (2016).
37. Chawla, A., Schwarz, E. J., Dimaculangan, D. D. & Lazar, M. A. Peroxisome proliferator-activated receptor (PPAR) gamma: adipose-predominant expression and induction early in adipocyte differentiation. *Endocrinology* **135**, 798–800 (1994).
38. Rosen, E. D. et al. PPAR gamma is required for the differentiation of adipose tissue in vivo and in vitro. *Mol. Cell.* **4**, 611–617 (1999).
39. Yeh, W. C., Cao, Z., Classon, M. & McKnight, S. L. Cascade regulation of terminal adipocyte differentiation by three members of the C/EBP family of leucine zipper proteins. *Genes Dev.* **9**, 168–181 (1995).
40. Rosen, E. D. & MacDougald, O. A. Adipocyte differentiation from the inside out. *Nat. Rev. Mol. Cell Biol.* **7**, 885–896 (2006).
41. Zamani, N. & Brown, C. W. Emerging roles for the transforming growth factor-{beta} superfamily in regulating adiposity and energy expenditure. *Endocr. Rev.* **32**, 387–403 (2011).
42. Dennler, S. et al. Direct binding of Smad3 and Smad4 to critical TGF beta-inducible elements in the promoter of human plasminogen activator inhibitor-type 1 gene. *EMBO J.* **17**, 3091–3100 (1998).
43. Roberts, A. F., Gumienny, T. L., Gleason, R. J., Wang, H. & Padgett, R. W. Regulation of genes affecting body size and innate immunity by the DBL-1/BMP-like pathway in *Caenorhabditis elegans*. *BMC Dev. Biol.* **10**, 61 (2010).
44. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018). GIANT Consortium.

45. Gao, H. et al. Early B cell factor 1 regulates adipocyte morphology and lipolysis in white adipose tissue. *Cell Metab.* **19**, 981–992 (2014).
46. Riba, A. et al. Explicit modeling of siRNA-dependent on- and off-target repression improves the interpretation of screening results. *Cell Syst.* **4**, 182–193.e4 (2017).
47. Danforth, E. Jr. Failure of adipocyte differentiation causes type II diabetes mellitus? *Nat. Genet.* **26**, 13 (2000).
48. Abraira, V. E. et al. Cross-repressive interactions between Lrig3 and netrin 1 shape the architecture of the inner ear. *Development* **135**, 4091–4099 (2008).
49. Graf, D., Malik, Z., Hayano, S. & Mishina, Y. Common mechanisms in development and disease: BMP signaling in craniofacial development. *Cytokine Growth Factor Rev.* **27**, 129–139 (2016).
50. Chang, W. et al. Bmp4 is essential for the formation of the vestibular apparatus that detects angular head movements. *PLoS Genet.* **4**, e1000050 (2008).
51. Suzuki, Y. et al. Targeted disruption of LIG-1 gene results in psoriasiform epidermal hyperplasia. *FEBS Lett.* **521**, 67–71 (2002).
52. Jensen, K. B. et al. Lrig1 expression defines a distinct multipotent stem cell population in mammalian epidermis. *Cell. Stem Cell.* **4**, 427–439 (2009).
53. Wong, V. W. et al. Lrig1 controls intestinal stem-cell homeostasis by negative regulation of ErbB signalling. *Nat. Cell Biol.* **14**, 401–408 (2012).
54. Piccirillo, S. G. et al. Bone morphogenetic proteins inhibit the tumorigenic potential of human brain tumour-initiating cells. *Nature* **444**, 761–765 (2006).
55. Lee, J. et al. Epigenetic-mediated dysfunction of the bone morphogenetic protein pathway inhibits differentiation of glioblastoma-initiating cells. *Cancer Cell.* **13**, 69–80 (2008).
56. Caja, L. et al. Snail regulates BMP and TGFβ pathways to control the differentiation status of glioma-initiating cells. *Oncogene* **37**, 2515–2531 (2018).
57. Kodaira, K. et al. Purification and identification of a BMP-like factor from bovine serum. *Biochem. Biophys. Res. Commun.* **345**, 1224–1231 (2006).
58. Herrera, B. & Inman, G. J. A rapid and sensitive bioassay for the simultaneous measurement of multiple bone morphogenetic proteins. Identification and quantification of BMP4, BMP6 and BMP9 in bovine and human serum. *BMC Cell Biol.* **10**, 20 (2009).
59. Rafidi, H. et al. Leucine-rich repeat and immunoglobulin domain-containing protein-1 (Lrig1) negative regulatory action toward ErbB receptor tyrosine kinases is opposed by leucine-rich repeat and immunoglobulin domain-containing protein 3 (Lrig3). *J. Biol. Chem.* **288**, 21593–21605 (2013).
60. Rondahl, V. et al. Lrig2-deficient mice are protected against PDGFB-induced glioma. *PLoS ONE* **8**, e73635 (2013).
61. TODARO, G. J. & GREEN, H. Quantitative studies of the growth of mouse embryo cells in culture and their development into established lines. *J. Cell Biol.* **17**, 299–313 (1963).
62. Faraz, M., Herdenberg, C., Holmlund, C., Henriksson, R. & Hedman, H. A protein interaction network centered on leucine-rich repeats and immunoglobulin-like domains 1 (LRIG1) regulates growth factor receptors. *J. Biol. Chem.* **293**, 3421–3435 (2018).
63. Nilsson, J. et al. Cloning, characterization, and expression of human LIG1. *Biochem. Biophys. Res. Commun.* **284**, 1155–1161 (2001).
64. Diab, J. et al. Lipidomics in ulcerative colitis reveal alteration in mucosal lipid composition associated with the disease state. *Inflamm. Bowel Dis.* **25**, 1780–1787 (2019).
65. Korchynskyi, O. & ten Dijke, P. Identification and functional characterization of distinct critically important bone morphogenetic protein-specific response elements in the Id1 promoter. *J. Biol. Chem.* **277**, 4883–4891 (2002).
66. Nilsson, J., Starefeldt, A., Henriksson, R. & Hedman, H. LRIG1 protein in human cells and tissues. *Cell Tissue Res.* **312**, 65–71 (2003).
67. O'Rourke, E. J., Soukas, A. A., Carr, C. E. & Ruvkun, G. *C. elegans* major fats are stored in vesicles distinct from lysosome-related organelles. *Cell Metab.* **10**, 430–435 (2009).
68. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
69. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
70. Lundbäck, V., Kulyté, A., Arner, P., Strawbridge, R. J. & Dahlman, I. Genome-wide association study of diabetogenic adipose morphology in the GENetics of adipocyte lipolysis (GENiAL) cohort. *Cells* **9**, 1085 (2020).
71. Gallagher, D. et al. How useful is body mass index for comparison of body fatness across age, sex, and ethnic groups? *Am. J. Epidemiol.* **143**, 228–239 (1996).
72. Kolaczynski, J. W. et al. A new technique for biopsy of human abdominal fat under local anaesthesia with Lidocaine. *Int J. Obes. Relat. Metab. Disord.* **18**, 161–166 (1994).
73. Löfgren, P., Hoffstedt, J., Näslund, E., Wirén, M. & Arner, P. Prospective and controlled studies of the actions of insulin and catecholamine in fat cells of obese women following weight reduction. *Diabetologia* **48**, 2334–2342 (2005).
74. Hirsch, J. & Gallian, E. Methods for the determination of adipose cell size in man and animals. *J. Lipid Res.* **9**, 110–119 (1968).
75. Andersson, D. P., Arner, E., Hogling, D. E., Rydén, M. & Arner, P. Abdominal subcutaneous adipose tissue cellularity in men and women. *Int J. Obes. (Lond.).* **41**, 1564–1569 (2017).
76. Loh, P. R. et al. Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

Acknowledgements

We thank Jonas Nilsson (Umeå University) for constructing the *pLRIG1-Δcyto-GFP* vector, Mahmood Faraz (Umeå University) for constructing the inducible *LRIG* vectors, Bhavna Rani (Umeå University) for assembling the *pLRIG1-3XFLAG* and *pLRIG1-ecto-3XFLAG* vectors, Richard Padgett (Rutgers University, NJ, USA) for nematode strains, Charlotte Nordström, Yvonne Jonsson, and Annika Holmberg for technical assistance, Mehmet Dinçer Inan for help with Oil Red O staining of *C. elegans* worms, Jonathan Gilthorpe (Umeå University) for valuable input regarding the Trophos and Seahorse analyses, the Swedish Metabolomics Centre (Umeå University) for help with lipid analysis, Neli Tsereteli (Lund University Diabetes Center) for assistance with data visualization, and Alaitz Poveda (Lund University Diabetes Center) for help with data preparation. We also thank all participants and staff involved in the GENiAL study. This work was supported by grants from the Swedish Cancer Society (CAN 2016/344 and CAN 2018/546), Kempe Foundation (JCK-1829), the Cancer Research Foundation in Northern Sweden (AMP 16-799), Lion's Cancer Research Foundation at Umeå University (LP 16-2134 and LP 15-2057), the Strategic Research Program in Diabetes at Karolinska Institutet, the Swedish Research Council, Novo Nordisk Foundation, the Swedish Diabetes Association, and by the regional agreement between Umeå University and Västerbotten County Council on the cooperation in the field of Medicine, Odontology and Health (ALF, RV 836951). The work undertaken by PWF, PM, and AP was supported by the European Research Council (CoG-2015_681742_NASCENT), the Swedish Research Council (Distinguished Young Researchers Award in Medicine), the Swedish Heart-Lung Foundation, and the Novo Nordisk Foundation. RJS was supported by a UKRI Innovation at HDR-UK Fellowship (MR/S003061/1).

Author contributions

C.He. and H.H. conceived the idea, planned the experiments, and wrote the manuscript. H.H. and R.H. supervised the project. C.He. developed the MEF lines, pSmad immunoassay, and adipogenic assay, and performed the in vitro experiments. O.B. and S.T. investigated the connection between *sma-10* and fat accumulation in *C. elegans*. C.Ho. contributed to the initial MEF triple KO experiments. A.A. investigated noncanonical BMP signaling and PDGF-driven cell proliferation and performed the BMP assays (Figs. 2c, d and 5d). P.M.M. and P.W.F. investigated the associations between *LRIG1* gene variants and BMI and diabetes. R.J.S., I.D., and P.A. investigated the correlation between *LRIG1* gene variants and adipose tissue morphology. All authors reviewed the final draft.

Funding

Open Access funding provided by University Of Umea.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42003-020-01613-w.

Correspondence and requests for materials should be addressed to H.H.

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

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

Paper III

PLOS MEDICINE

RESEARCH ARTICLE

Predicting and elucidating the etiology of fatty liver disease: A machine learning modeling and validation study in the IMI DIRECT cohorts

Naeimeh Atabaki-Pasdar[1], Mattias Ohlsson[2,3], Ana Viñuela[4,5,6], Francesca Frau[7], Hugo Pomares-Millan[1], Mark Haid[8], Angus G. Jones[9], E. Louise Thomas[10], Robert W. Koivula[1,11], Azra Kurbasic[1], Pascal M. Mutie[1], Hugo Fitipaldi[1], Juan Fernandez[1], Adem Y. Dawed[12], Giuseppe N. Giordano[1], Ian M. Forgie[12], Timothy J. McDonald[9,13], Femke Rutters[14], Henna Cederberg[15], Elizaveta Chabanova[16], Matilda Dale[17], Federico De Masi[18], Cecilia Engel Thomas[17], Kristine H. Allin[19,20], Tue H. Hansen[19,21], Alison Heggie[22], Mun-Gwan Hong[17], Petra J. M. Elders[23], Gwen Kennedy[24], Tarja Kokkola[25], Helle Krogh Pedersen[19], Anubha Mahajan[26], Donna McEvoy[22], Francois Pattou[27], Violeta Raverdy[27], Ragna S. Häussler[17], Sapna Sharma[28,29], Henrik S. Thomsen[16], Jagadish Vangipurapu[25], Henrik Vestergaard[19,30], Leen M. 't Hart[14,31,32], Jerzy Adamski[8,33,34], Petra B. Musholt[35], Soren Brage[36], Søren Brunak[18,37], Emmanouil Dermitzakis[4,5,6], Gary Frost[38], Torben Hansen[19,39], Markku Laakso[25,40], Oluf Pedersen[19], Martin Ridderstråle[41], Hartmut Ruetten[7], Andrew T. Hattersley[9], Mark Walker[22], Joline W. J. Beulens[14,42], Andrea Mari[43], Jochen M. Schwenk[17], Ramneek Gupta[18], Mark I. McCarthy[11,26,44,45], Ewan R. Pearson[12], Jimmy D. Bell[10], Imre Pavo[46], Paul W. Franks[1,47]*

1 Genetic and Molecular Epidemiology Unit, Department of Clinical Sciences, Lund University, Malmö, Sweden, **2** Computational Biology and Biological Physics Unit, Department of Astronomy and Theoretical Physics, Lund University, Lund, Sweden, **3** Center for Applied Intelligent Systems Research, Halmstad University, Halmstad, Sweden, **4** Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland, **5** Institute for Genetics and Genomics in Geneva, University of Geneva Medical School, Geneva, Switzerland, **6** Swiss Institute of Bioinformatics, Geneva, Switzerland, **7** Sanofi-Aventis Deutschland, Frankfurt am Main, Germany, **8** Research Unit Molecular Endocrinology and Metabolism, Helmholtz Zentrum München, Neuherberg, Germany, **9** Institute of Biomedical and Clinical Science, College of Medicine and Health, University of Exeter, Exeter, United Kingdom, **10** Research Centre for Optimal Health, School of Life Sciences, University of Westminster, London, United Kingdom, **11** Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom, **12** Division of Population Health and Genomics, School of Medicine, University of Dundee, Ninewells Hospital, Dundee, United Kingdom, **13** Blood Sciences, Royal Devon and Exeter NHS Foundation Trust, Exeter, United Kingdom, **14** Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, Amsterdam UMC, Amsterdam, the Netherlands, **15** Department of Endocrinology, Abdominal Centre, Helsinki University Hospital, Helsinki, Finland, **16** Department of Diagnostic Radiology, Copenhagen University Hospital Herlev Gentofte, Herlev, Denmark, **17** Affinity Proteomics, Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Solna, Sweden, **18** Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark, **19** Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, **20** Center for Clinical Research and Prevention, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark, **21** Department of Cardiology and Endocrinology, Slagelse Hospital, Slagelse, Denmark, **22** Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom, **23** Department of General Practice, Amsterdam Public Health Research Institute, Amsterdam UMC, Amsterdam, the Netherlands, **24** Immunoassay Biomarker Core Laboratory, School of Medicine, University of Dundee, Ninewells Hospital, Dundee, United Kingdom, **25** Internal Medicine, Institute of Clinical Medicine, University of Eastern Finland, Kuopio, Finland, **26** Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, **27** University of Lille, Inserm, UMR 1190, Translational Research in Diabetes, Department of Endocrine Surgery, CHU Lille, Lille, France, **28** German Center for Diabetes Research, Neuherberg, Germany, **29** Unit of Molecular Epidemiology, Institute of Epidemiology II, Helmholtz Zentrum München, Neuherberg, Germany, **30** Steno Diabetes Center Copenhagen, Gentofte, Denmark, **31** Department of Cell and Chemical Biology, Leiden University Medical Center, Leiden, the Netherlands, **32** Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center,

Citation: Atabaki-Pasdar N, Ohlsson M, Viñuela A, Frau F, Pomares-Millan H, Haid M, et al. (2020) Predicting and elucidating the etiology of fatty liver disease: A machine learning modeling and validation study in the IMI DIRECT cohorts. PLoS Med 17(6): e1003149. https://doi.org/10.1371/journal.pmed.1003149

Academic Editor: Dominik Heider, University of Marburg, GERMANY

Received: January 16, 2020

Accepted: May 22, 2020

Published: June 19, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pmed.1003149

Data Availability Statement: Data cannot be shared publicly due to a need to maintain the confidentiality of patient data. Interested

researchers may contact DIRECTdataaccess@dundee.ac.uk to request and obtain relevant data.

Funding: The work leading to this publication has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement n°115317 (DIRECT), resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. NAP is supported in part by Henning och Johan Throne-Holsts Foundation, Hans Werthén Foundation, an IRC award from the Swedish Foundation for Strategic Research and a European Research Council award ERC-2015-CoG - 681742_NASCENT. HPM is supported by an IRC award from the Swedish Foundation for Strategic Research and a European Research Council award ERC-2015-CoG - 681742_NASCENT. AGJ is supported by an NIHR Clinician Scientist award (17/0005624). RK is funded by the Novo Nordisk Foundation (NNF18OC0031650) as part of a postdoctoral fellowship, an IRC award from the Swedish Foundation for Strategic Research and a European Research Council award ERC-2015-CoG - 681742_NASCENT. AK, PM, HF, JF and GNG are supported by an IRC award from the Swedish Foundation for Strategic Research and a European Research Council award ERC-2015-CoG - 681742_NASCENT. TJM is funded by an NIHR clinical senior lecturer fellowship. S.Bru acknowledges support from the Novo Nordisk Foundation (grants NNF17OC0027594 and NNF14CC0001). ATH is a Wellcome Trust Senior Investigator and is also supported by the NIHR Exeter Clinical Research Facility. JMS acknowledges support from Science for Life Laboratory (Plasma Profiling Facility), Knut and Alice Wallenberg Foundation (Human Protein Atlas) and Erling-Persson Foundation (KTH Centre for Precision Medicine). MIM is supported by the following grants; Wellcome (090532, 098381, 106130, 203141, 212259); NIH (U01-DK105535). PWF is supported by an IRC award from the Swedish Foundation for Strategic Research and a European Research Council award ERC-2015-CoG - 681742_NASCENT. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: PWF is a consultant for Novo Nordisk, Lilly, and Zoe Global Ltd., and has received research grants from numerous diabetes drug companies. HR is an employee and shareholder of Sanofi. MIM: The views expressed

Leiden, the Netherlands, **33** Lehrstuhl für Experimentelle Genetik, Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt, Technische Universität München, Freising, Germany, **34** Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, **35** Diabetes Division, Research and Development, Sanofi, Frankfurt, Germany, **36** MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom, **37** Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, **38** Section for Nutrition Research, Department of Metabolism, Digestion and Reproduction, Imperial College London, London, United Kingdom, **39** Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark, **40** Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland, **41** Clinical Pharmacology and Translational Medicine, Novo Nordisk, Søborg, Denmark, **42** Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands, **43** Institute of Neuroscience, National Research Council, Padua, Italy, **44** NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, United Kingdom, **45** OMNI Human Genetics, Genentech, South San Francisco, California, United States of America, **46** Eli Lilly Regional Operations, Vienna, Austria, **47** Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts, United States of America

* Paul.Franks@med.lu.se

Abstract

Background

Non-alcoholic fatty liver disease (NAFLD) is highly prevalent and causes serious health complications in individuals with and without type 2 diabetes (T2D). Early diagnosis of NAFLD is important, as this can help prevent irreversible damage to the liver and, ultimately, hepatocellular carcinomas. We sought to expand etiological understanding and develop a diagnostic tool for NAFLD using machine learning.

Methods and findings

We utilized the baseline data from IMI DIRECT, a multicenter prospective cohort study of 3,029 European-ancestry adults recently diagnosed with T2D ($n = 795$) or at high risk of developing the disease ($n = 2,234$). Multi-omics (genetic, transcriptomic, proteomic, and metabolomic) and clinical (liver enzymes and other serological biomarkers, anthropometry, measures of beta-cell function, insulin sensitivity, and lifestyle) data comprised the key input variables. The models were trained on MRI-image-derived liver fat content (<5% or ≥5%) available for 1,514 participants. We applied LASSO (least absolute shrinkage and selection operator) to select features from the different layers of omics data and random forest analysis to develop the models. The prediction models included clinical and omics variables separately or in combination. A model including all omics and clinical variables yielded a cross-validated receiver operating characteristic area under the curve (ROCAUC) of 0.84 (95% CI 0.82, 0.86; $p < 0.001$), which compared with a ROCAUC of 0.82 (95% CI 0.81, 0.83; $p < 0.001$) for a model including 9 clinically accessible variables. The IMI DIRECT prediction models outperformed existing noninvasive NAFLD prediction tools. One limitation is that these analyses were performed in adults of European ancestry residing in northern Europe, and it is unknown how well these findings will translate to people of other ancestries and exposed to environmental risk factors that differ from those of the present cohort. Another key limitation of this study is that the prediction was done on a binary outcome of liver fat quantity (<5% or ≥5%) rather than a continuous one.

in this article are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health. MIM has served on advisory panels for Pfizer, NovoNordisk and Zoe Global, has received honoraria from Merck, Pfizer, Novo Nordisk and Eli Lilly, and research funding from Abbvie, Astra Zeneca, Boehringer Ingelheim, Eli Lilly, Janssen, Merck, NovoNordisk, Pfizer, Roche, Sanofi Aventis, Servier, and Takeda. As of June 2019, MIM is an employee of Genentech, and a holder of Roche stock. AM is a consultant for Lilly and has received research grants from several diabetes drug companies.

Abbreviations: ALT, alanine transaminase; AST, aspartate transaminase; DBP, diastolic blood pressure; EFS, ensemble feature selection; FLI, fatty liver index; HBA1c, hemoglobin A1C; HSI, hepatic steatosis index; MMTT, mixed-meal tolerance test; MRI, magnetic resonance imaging; NAFLD, non-alcoholic fatty liver disease; NAFLD-LFS, non-alcoholic fatty liver disease liver fat score; NASH, non-alcoholic steatohepatitis; OGTT, oral glucose tolerance test; QC, quality control; ROCAUC, receiver operating characteristic area under the curve; SBP, systolic blood pressure; T2D, type 2 diabetes; TG, triglycerides.

Conclusions

In this study, we developed several models with different combinations of clinical and omics data and identified biological features that appear to be associated with liver fat accumulation. In general, the clinical variables showed better prediction ability than the complex omics variables. However, the combination of omics and clinical variables yielded the highest accuracy. We have incorporated the developed clinical models into a web interface (see: https://www.predictliverfat.org/) and made it available to the community.

Trial registration

ClinicalTrials.gov NCT03814915.

Author summary

Why was this study done?

- Globally, about 1 in 4 adults have non-alcoholic fatty liver disease (NAFLD), which adversely affects energy homeostasis (in particular blood glucose concentrations), blood detoxification, drug metabolism, and food digestion.
- Although numerous noninvasive tests to detect NAFLD exist, these typically include inaccurate blood-marker tests or expensive imaging methods.
- The purpose of this work was to develop accurate noninvasive methods to aid in the clinical prediction of NAFLD.

What did the researchers do and find?

- The analyses applied machine learning methods to data from the deep-phenotyped IMI DIRECT cohorts (*n* = 1,514) to identify sets of highly informative variables for the prediction of NAFLD. The criterion measure was liver fat quantified from MRI.
- We developed a total of 18 prediction models that ranged from very inexpensive models of modest accuracy to more expensive biochemistry- and/or omics-based models with high accuracy.
- We found that models using measures commonly collected in either clinical settings or research studies proved adequate for the prediction of NAFLD.
- The addition of detailed omics data significantly improved the predictive utility of these models. We also found that of all omics markers, proteomic markers yielded the highest predictive accuracy when appropriately combined.

What do these findings mean?

- We envisage that these new approaches to predicting fatty liver may be of clinical value when screening at-risk populations for NAFLD.

- The identification of specific molecular features that underlie the development of NAFLD provides novel insights into the disease's etiology, which may lead to the development of new treatments.

Introduction

Non-alcoholic fatty liver disease (NAFLD) is characterized by the accumulation of fat in hepatocytes in the absence of excessive alcohol consumption. NAFLD is a spectrum of liver diseases, with its first stage, known as simple steatosis, defined as liver fat content ≥5% of total liver weight. Simple steatosis can progress to non-alcoholic steatohepatitis (NASH), fibrosis, cirrhosis, and eventually hepatocellular carcinoma. In NAFLD, triglycerides (TG) accumulate in hepatocytes, and liver insulin sensitivity is diminished, promoting hepatic gluconeogenesis, thereby raising the risk of type 2 diabetes (T2D) or exacerbating the disease pathology in those with diabetes [1–5]. Growing evidence also links an increased risk of cardiovascular events with NAFLD [6,7].

The prevalence of NAFLD is thought to be around 20%–40% in the general population in high-income countries, with numbers growing worldwide, imposing a substantial economic and public health burden [8–11]. However, the exact prevalence of NAFLD has not been clarified, in part because liver fat is difficult to accurately assess. Liver biopsy, magnetic resonance imaging (MRI), ultrasound, and liver enzyme tests are often used for NAFLD diagnosis, but the invasive nature of biopsies, the high cost of MRI scans, the non-quantitative nature and low sensitivity of conventional ultrasounds, and the low accuracy of liver enzyme tests are significant limitations [12–14]. To address this gap, several liver fat prediction indices have been developed, but none of these has sufficiently high predictive ability to be considered a gold standard [12].

The purpose of this study was to use machine learning to identify novel molecular features associated with NAFLD and combine these with conventional clinical variables to predict NAFLD. Our models include variables that are likely to be informative of disease etiology, some of which may be of use in clinical practice.

Methods

Participants (IMI DIRECT)

The primary data utilized in this study were generated within the IMI DIRECT consortium, which includes persons with diabetes (n = 795) and without diabetes (n = 2,234). All participants provided informed written consent, and the study protocol was approved by the regional research ethics committees for each clinical study center. Details of the study design and the core characteristics are provided elsewhere [15,16].

Measures (IMI DIRECT)

A T2*-based multiecho technique was used to derive liver fat content from MRI [17,18], and the percentage values were categorized as fatty (≥5%) or non-fatty (<5%) to define the outcome variable. We elected not to attempt quantitative prediction of liver fat content, as this would require a much larger dataset to be adequately powered. A frequently-sampled 75-g oral glucose tolerance test (OGTT) or a frequently sampled mixed-meal tolerance test (MMTT) was performed, from which measures of glucose and insulin dynamics were calculated, as previously described [15,16]. Of 3,029 IMI DIRECT participants, 50% (n = 1,514) had the liver fat

Table 1. Characteristics of IMI DIRECT participants in the non-diabetes, diabetes, and combined cohorts separated for individuals with fatty liver versus non-fatty liver.

Characteristics	Non-diabetes cohort		Diabetes cohort		Combined cohort	
	Fatty liver	Non-fatty liver	Fatty liver	Non-fatty liver	Fatty liver	Non-fatty liver
N (percent)	344 (34)	667 (66)	296 (59)	207 (41)	640 (42)	874 (58)
Age (years)	61 (56, 66)	62 (56, 66)	62 (55, 67)	63 (58, 69)	61 (56, 66)	62 (56, 67)
Sex, *n* (percent female)	62 (18)	134 (20)	130 (44)	86 (42)	192 (30)	220 (25)
Weight (kg)	90.75 (81.50, 100.25)	81.40 (75.67, 89.60)	92.85 (81.47, 103.75)	80.80 (73.00, 93.55)	91.20 (81.50, 102.00)	81.40 (74.03, 90.17)
Waist circumference (cm)	105 (98, 112)	97 (91, 103)	107 (97, 115)	97 (90, 107)	106 (98, 113)	97 (91, 103)
BMI (kg/m^2)	29.23 (26.91, 32.05)	26.69 (24.75, 28.71)	31.47 (28.37, 35.35)	27.64 (25.53, 31.07)	30.05 (27.53, 33.52)	26.85 (24.91, 29.23)
SBP	134.70 (125.30, 143.00)	129.33 (120.00, 140.00)	131 (122.00, 139.33)	127.67 (117.67, 138.33)	132.67 (124.00, 142.00)	128.83 (119.33, 140.00)
DBP	83.50 (79.33, 89.83)	80.67 (75.67, 86.00)	76.67 (72.00, 84.00)	72.67 (67.17, 80.67)	81.33 (5.33, 87.33)	80.00 (73.33, 84.67)
HbA1c (mmol/mol)	38 (36, 40)	37 (35, 39)	47 (44, 51)	45 (42, 48)	41 (37, 46)	38 (36, 41)
Fasting glucose (mmol/l)	5.90 (5.60, 6.30)	5.70 (5.40, 6.00)	7.20 (6.30, 7.90)	6.70 (5.80, 7.60)	6.30 (5.80, 7.20)	5.80 (5.40, 6.30)
Fasting insulin (pmol/l)	75.60 (54.30, 104.40)	44.10 (27.75, 66.00)	115.80 (75.80, 167.80)	60.20 (40.85, 82.90)	90.90 (61.20, 133.90)	48.60 (30.00, 69.60)
2-hour glucose (mmol/l)	6.55 (5.37, 8.20)	5.70 (4.70, 6.80)	9.00 (6.90, 10.65)	7.90 (6.20, 9.90)	7.40 (5.90, 9.60)	6.00 (4.90, 7.50)
2-hour insulin (pmol/l)	345.60 (198.40, 566.20)	169.80 (100.20, 274.20)	489.30 (297.40, 700.50)	271.00 (166.40, 418.10)	403.20 (236.60, 643.50)	190.70 (110.80, 317.60)
Triglycerides (mmol/l)	1.49 (1.13, 2.09)	1.12 (0.86, 1.47)	1.49 (1.01, 1.99)	1.12 (0.86, 1.48)	1.49 (1.08, 2.02)	1.12 (0.86, 1.47)
ALT (units/l)	21 (14, 29)	15 (10, 20)	25 (19, 33)	20 (16, 24)	23 (16, 32)	16 (12, 22)
AST (units/l)	29 (24, 37)	25 (21, 30)	24 (20, 30)	22 (19, 27)	26 (22, 33)	24 (20, 29)
Alcohol intake, *n* for “never,” “occasionally,” “regularly”	21, 68, 255	91, 133, 443	52, 81, 163	38, 45, 124	73, 149, 418	129, 178, 567
Liver fat	8.80 (6.60, 13.00)	2.20 (1.50, 3.30)	11.10 (7.30, 15.82)	2.70 (1.95, 4.00)	9.50 (6.80, 14.30)	2.40 (1.60, 3.50)

Values are median (interquartile range) unless otherwise specified.

ALT, alanine transaminase; AST, aspartate transaminase; BMI, body mass index; DBP, diastolic blood pressure; HbA1c, hemoglobin A1C; SBP, systolic blood pressure.

https://doi.org/10.1371/journal.pmed.1003149.t001

MRI data (503 with diabetes and 1,011 without diabetes). The distribution of the liver fat data among different centers and cohorts is shown in S1 and S2 Figs.

The list of the clinical input (predictor) variables (n = 58), including anthropometric measurements, plasma biomarkers, and lifestyle factors, are shown in S1 Table. These clinical variables were controlled for center effect by deriving residuals from a linear model including each clinical variable in each model; these residuals were then inverse normalized and used in subsequent analyses. Inverse normal transformation is a nonparametric method that replaces the data quantiles by quantiles from the standard normal distribution in order to reduce the impact of outliers and deviation from a normal distribution.

A detailed overview of participant characteristics for the key variables is shown in Table 1 for all IMI DIRECT participants with MRI data. There were no substantial differences in characteristics between these participants and those from IMI DIRECT who did not have MRI data (see S2 Table).

Genetic, transcriptomic, proteomic, and metabolomic datasets were used as input omics variables in the analyses. Buffy coat was separated from whole blood, and DNA was then extracted and genotyped using the Illumina HumanCore array (HCE24 v1.0); genotype imputation was performed using the Haplotype Reference Consortium (HRC) and 1000 Genomes (1KG) reference panels. Details of the quality control (QC) steps for the genetic data are described elsewhere [16]. Transcriptomic data were generated using RNA sequencing from fasting whole blood. Only protein-coding genes were included in the analyses, as reads per kilobase of transcript per million mapped reads (RPKM). The targeted metabolomic data of fasting plasma samples were generated using the Biocrates AbsoluteIDQ p150 Kit. Additionally, untargeted LC/MS-based metabolomics was used to cover a broader spectrum of metabolites. A combination of technologies and quantitative panels of protein assays were used to generate "targeted" proteomic data. This included Olink proximity extension assays [19], sandwich immunoassay kits using Luminex technology (Merck Millipore and R&D Systems, Sweden), microfluidic ELISA assays (ProteinSimple, US [20]), protein analysis by Myriad RBM (Germany), and hsCRP analysis (MLM Medical Labs, Germany). In addition, protein data were generated by single-binder assays using highly multiplexed suspension bead arrays [21]. This approach (denoted "exploratory" proteomics) included a combination of antibodies targeting proteins selected by the consortium given published and unpublished evidence for association with glycemia-related traits. More information about data generation and QC of the transcriptomic, proteomic, and metabolomic data is provided in S1 Text. Technical covariates for transcriptomics include guanine-cytosine mean content, insert size, analysis lane and RNA integrity number, cell composition, date, and center. Technical covariates for proteomics were center, assay, plate number, and plate layout ($n = 4$), and for the targeted metabolites the technical covariates were center and plate. These technical covariates were used to correct the omics data, and the residuals were then extracted from these models and inverse normalized prior to further analyses.

Feature selection (IMI DIRECT)

We developed a series of NAFLD prediction models composed of variables that are available within clinical settings, as well as those not currently available in most clinics (see S3 Table). We had 2 strategies for selecting the clinical variables. For models 1–3, we selected variables based on clinical accessibility and their established association with fatty liver from existing literature without applying statistical procedures for data reduction. For model 4, a pairwise Pearson correlation matrix was used for feature selection of the clinical variables by placing a pairwise correlation threshold of $r > 0.8$, and we then selected the variables we considered most accessible among those that were collinear. Feature selection was undertaken in the combined cohort (diabetes and non-diabetes) in order to maximize sample size and statistical power. Of 1,514 participants with liver fat data, 1,049 had all necessary clinical and multi-omics data for a complete case analysis. We used k-nearest neighbor [22] imputation with k equal to 10 as a means to reduce the loss of sample size, but found that this did not materially improve predictive power in subsequent analyses, so we decided not to include these imputed data. An overview of the pairwise correlations among the clinical variables available in these 1,049 IMI DIRECT participants is presented in Fig 1.

The high-dimensionality nature of omics data also necessitated data reduction using the feature selection tool LASSO prior to building the model. LASSO is a regression analysis method that minimizes the sum of least squares in a linear regression model and shrinks selected beta coefficients (β_j) using penalties (Eq 1). Minimizing the value from Eq 1, LASSO excludes the least informative variables and selects those features of most importance for the outcome of interest (y) in a sample of n cases, each of which consists of m parameters. The penalty applied by λ can be any value from 0 to positive infinity and is determined through a cross-validation step [26].

$$\sum\nolimits_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \times \sum\nolimits_{j=1}^{m} |\beta_j| \qquad (1)$$

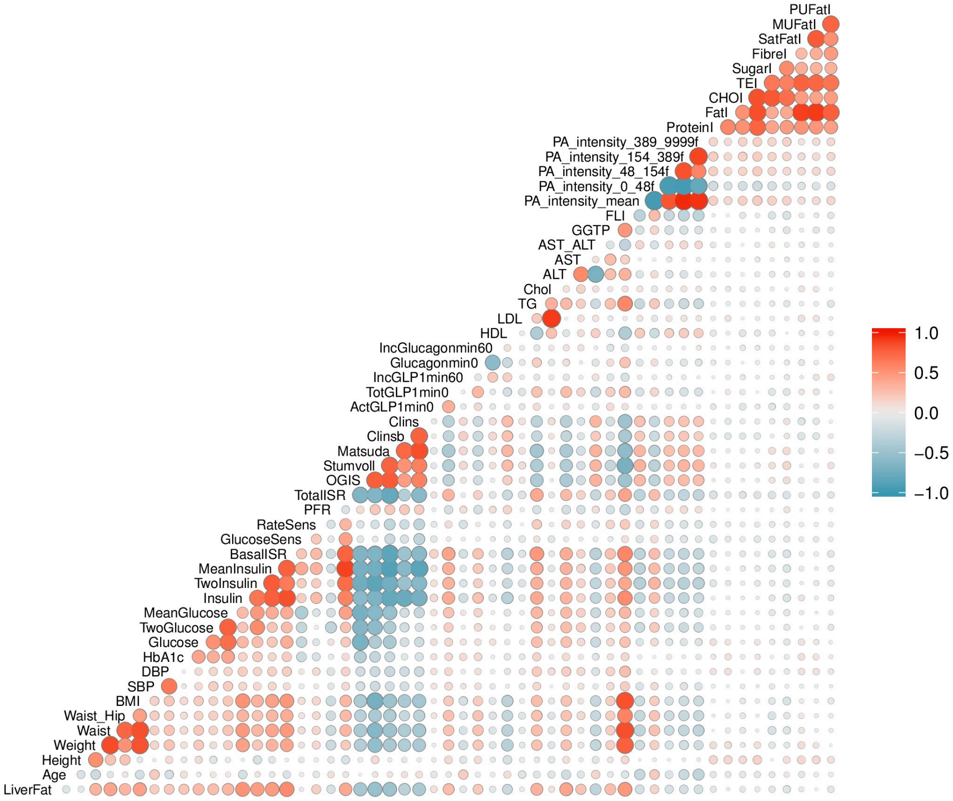


Fig 1. Pearson pairwise correlation matrix of clinical variables (data are inverse normal transformed) in the cohort combining participants with and without diabetes in IMI DIRECT (*n* = 1,049). The magnitude and direction of the correlation are reflected by the size (larger is stronger) and color (red is positive and blue is negative) of the circles, respectively. ActGLP1min0, concentration of fasting active GLP-1 in plasma; ALT, alanine transaminase; AST, aspartate transaminase; AST_ALT, AST to ALT ratio; BasalISR, insulin secretion at the beginning of the oral glucose tolerance test/mixed-meal tolerance test; BMI, body mass index; CHOI, total daily intake of dietary carbohydrates; Chol, total cholesterol; Clins, mean insulin clearance during the oral glucose tolerance test/mixed-meal tolerance test, calculated as (mean insulin secretion)/(mean insulin concentration); Clinsb, insulin clearance calculated from basal values as (insulin secretion)/(insulin concentration); DBP, mean diastolic blood pressure; FatI, total daily intake of dietary fats; FLI, fatty liver index; FibreI, total daily intake of dietary Association of Official Analytical Chemists (AOAC) fiber; GGTP, gamma-glutamyl transpeptidase; Glucagonmin0, fasting glucagon concentration; Glucose, fasting glucose from venous plasma samples; GlucoseSens, glucose sensitivity, slope of the dose–response relating insulin secretion to glucose concentration; HbA1c, hemoglobin A1C; HDL, fasting high-density lipoprotein cholesterol; IncGLP1min60, 1-hour GLP-1 increment; IncGlucagonmin60, 1-hour glucagon increment; Insulin, fasting insulin from venous plasma samples; LDL, fasting low-density lipoprotein cholesterol; Matsuda, insulin sensitivity index according to the method of Matsuda et al. [23]; MeanGlucose, mean glucose during the oral glucose tolerance test/mixed-meal tolerance test; MeanInsulin, mean insulin during the oral glucose tolerance test/mixed-meal tolerance test; MUFatI, daily intake of dietary monounsaturated fats; OGIS, oral glucose insulin sensitivity index according to the method of Mari et al. [24]; PA_intensity_0_48f, number of values in high-pass-filtered vector magnitude physical activity at ≥0 and ≤48; PA_intensity_154_389f, number of values in high-pass-filtered vector magnitude physical activity at ≥154 and ≤389; PA_intensity_389_9999f, number of values in high-pass-filtered vector magnitude physical activity at ≥389 and ≤9,999; PA_intensity_48_154f, number of values in high-pass-filtered vector magnitude physical activity at ≥48 and ≤154; PA_intensity_mean, mean high-pass-filtered vector magnitude physical activity intensity; PFR, potentiation factor ratio; ProteinI, total daily intake of dietary proteins; PUFatI, daily intake of dietary polyunsaturated fats; RateSens, rate sensitivity (parameter characterizing early insulin secretion); SatFatI, daily intake of dietary saturated fats; SBP, mean systolic blood pressure; Stumvoll, insulin sensitivity index according to the method of Stumvoll et al. [25]; SugarI, total daily intake of dietary; TEI, total daily energy intake based on validated multi-pass food habit questionnaire; TG, fasting triglycerides; TotalISR, integral of insulin secretion during the whole oral glucose tolerance test/mixed-meal tolerance test; TotGLP1min0, concentration of fasting total GLP-1 in plasma; TwoGlucose, 2-hour glucose after oral glucose tolerance test/mixed-meal tolerance test; TwoInsulin, 2-hour insulin; Waist_Hip, waist to hip ratio.

https://doi.org/10.1371/journal.pmed.1003149.g001

To minimize bias (for example by overfitting), we randomly divided the dataset and used 70% ($n = 735$) for feature selection and 30% ($n = 314$) for the model generation (see below). We selected these thresholds for partitioning the dataset in order to maximize the power to select the informative features. Stratified random sampling [27] based on the outcome variable was undertaken in order to preserve the distribution of the liver fat categories in the 2 feature selection and model generation sets. We selected LASSO, as a nonlinear data reduction tool might lead to overfitting owing to the high dimensionality of omics data. LASSO was conducted with package glmnet in R [28] with a 10-fold cross-validation step for defining the λ parameter that resulted in the minimum value for the mean square error of the regression model.

Feature selection using LASSO was undertaken in each omics dataset (genetic, transcriptomic, proteomic, and metabolomic) using 70% of the available data (models 5–18). For the genetic dataset, we first performed a genome-wide association study (GWAS) prior to LASSO in order to identify single nucleotide polymorphisms (SNPs) tentatively associated with liver fat accumulation ($p < 5 \times 10^{-6}$). LASSO was then applied to these index variants for feature selection in 70% of the study sample. The individual SNP association analysis was conducted with RVTESTS v2.0.2 [29], which applies a linear mixed model with an empirical kinship matrix to account for familial relatedness, cryptic relatedness, and population stratification. Only common variants with minor allele frequency (MAF) greater than 5% contributed to the kinship matrix. Liver fat data were log-transformed and then adjusted for age, age^2, sex, center, body mass index (BMI), and alcohol consumption. These values were then inverse normal transformed and used in the GWAS analyses. We limited our analysis to genetic MAF > 1% and imputation quality score > 0.3. S3 and S4 Figs show the resulting Manhattan plot, depicting each SNP's association with liver fat percentage and the quantile–quantile (QQ) plot of the GWAS results for liver fat. For the genetic data, 23 SNPs were selected out of the 108 SNPs with p-values $< 5 \times 10^{-6}$. For the transcriptomics, 93 genes were selected out of 16,209 protein-coding genes. In the exploratory and targeted proteomics, 22 out of 377 and 48 out of 483 proteins were selected, respectively. In the targeted and untargeted metabolomic data, 25 out of 116 and 39 out of 172 metabolites were selected by LASSO, respectively.

Model training and evaluation

The remaining 30% of the data was used to develop the binary prediction models for fatty liver (yes/no) with selected features used as input variables. We utilized the random forest supervised machine learning method, which is an aggregation of decision trees built from bootstrapped datasets (a process called "bagging"). Typically, two-thirds of the data are retained in these bootstrapped datasets, and the remaining third is termed the out of bag (OOB) dataset, which is used to validate the performance of the model. To avoid overfitting and improve generalizability, 5-fold cross-validation was done for resampling the training samples and was repeated 5 times to create multiple versions of the folds. The number of trees was set to 1,000 to provide an accurate and stable prediction. Receiver operating characteristic (ROC) curves were used to evaluate model performance by measuring the area under the curve (AUC). A ROC curve uses a combination of sensitivity (true positive rate) and specificity (true negative rate) to assess prediction performance. In our analysis, the random forest model is used to derive probability estimates for the presence of fatty liver. In order to make a class prediction, it is necessary to impose a cutoff above which fatty liver is deemed probable and below which it is considered improbable. The choice of cutoff influences both sensitivity and specificity for a given prediction model. We considered the effect of different cutoffs on these performance measurements. Additionally, we calculated the F1 score, which is the harmonic mean of precision (positive predictive value) and sensitivity, derived as follows:

$$\text{F1 score} = \frac{2 \times \text{sensitivity} \times \text{precision}}{(\text{sensitivity} + \text{precision})} \qquad (2)$$

Balanced accuracy was also evaluated, which is the proportion of individuals correctly classified (true positives and true negatives) within each class individually. Measurements of sensitivity, specificity, F1 score, and balanced accuracy were computed and compared at different cutoffs for the diabetes, non-diabetes, and combined cohorts. The variable importance was also determined via a "permutation accuracy importance" measure using random forest analysis. In brief, for each tree, the prediction accuracy was calculated in the OOB test data. Each predictor variable was then permuted, and the accuracy was recalculated. The difference in the accuracies was averaged over all the trees and then normalized by the standard error. Thus, the measure for variable importance is the difference in prediction accuracy before and after the permutation for each variable [30]. In addition, we used the ensemble feature selection (EFS) method to determine the normalized importance value of all features [31]. With this approach, we do not rely on only random forest for the importance ranking, and we can build the cumulative importance values from different methods including Spearman's rank correlation test, Pearson's product moment correlation test, beta-values of logistic regression, the error-rate-based variable importance measure, and the Gini-index-based variable importance measure. Statistical analyses were undertaken using R software version 3.2.5 [32], and the random forest models were built using the caret package [33]. Fig 2 shows an overview of the different stages involved in the data processing and model training.

Comparison with other fatty liver indices

Given the accessible data within the IMI DIRECT cohorts, several existing fatty liver indices could be calculated and compared with the IMI DIRECT prediction models. These included

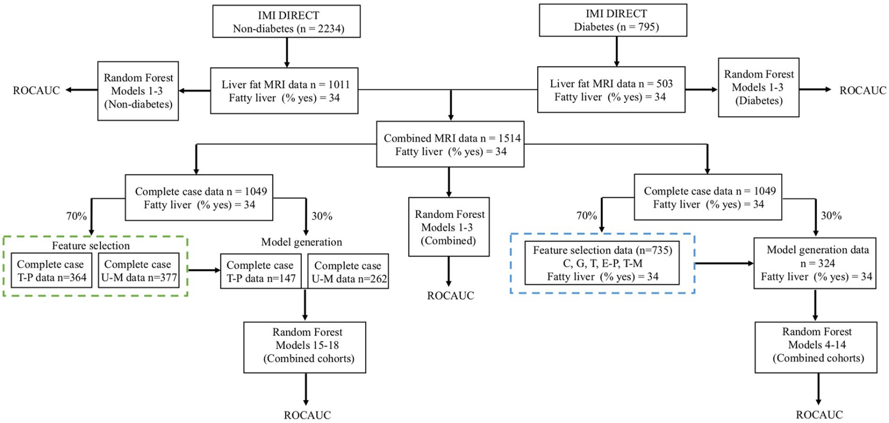


Fig 2. Overview of the different stages involved in data processing and model training. Data sources: clinical (C), genetic (G), transcriptomic (T), exploratory proteomic (E-P), targeted proteomic (T-P), targeted metabolomic (T-M), and untargeted metabolomic (U-M). The green and blue dashed boxes illustrate the feature selection step, the details of which can be found in S5 Fig. ROCAUC, receiver operating characteristic area under the curve.

https://doi.org/10.1371/journal.pmed.1003149.g002

the fatty liver index (FLI) [34], hepatic steatosis index (HSI) [35], and the NAFLD liver fat score (NAFLD-LFS) [36].

FLI. The FLI is commonly used to estimate the presence or absence of fatty liver (categorized into fatty [$\geq$60 FLI units] or non-fatty liver [$<$60 FLI units]) [34]. The FLI uses data on TG, waist circumference, BMI, and serum gamma-glutamyl transpeptidase (GGTP) and is calculated as follows:

$$\mathrm{FLI} = \frac{\mathrm{e}^{((0.953\times\ln(\mathrm{TG}))+\ (0.139\times\mathrm{BMI})\ +\ (0.718\times\ln(\mathrm{GGTP}))\ +\ (0.053\times\mathrm{Waist})\ -\ 15.745\ )} \times 100}{(1\ +\ \mathrm{e}^{((0.953\times\ln\ \mathrm{TG}))+\ (0.139\times\mathrm{BMI})\ +\ (0.718\times\ln(\mathrm{GGTP}))\ +\ (0.053\times\mathrm{Waist})\ -\ 15.745\ )})} \tag{3}$$

NAFLD-FLS. NAFLD-FLS was calculated using fasting serum (fs) insulin, aspartate transaminase (AST), alanine transaminase (ALT), T2D, and metabolic syndrome (MS) (defined according to the International Diabetes Federation [37]) to provide an estimate of liver fat content. A NAFLD-FLS value above −0.64 is considered to indicate the presence of NAFLD:

$$\begin{aligned}\text{NAFLD-LFS} &= -2.89 + 1.18 \times \mathrm{MS}\,(\text{yes } 1,\ \text{no } 0) + 0.45 \times \mathrm{T2D}\,(\text{yes } 2,\ \text{no } 0)\\ &\quad + 0.15 \times \text{fs Insulin}\end{aligned} \tag{4}$$

HSI. The HSI uses BMI, sex, T2D diagnosis (yes/no), and the ratio of ALT to AST and is calculated as follows:

$$\mathrm{HSI} = 8 \times \frac{\mathrm{ALT}}{\mathrm{AST}} + \mathrm{BMI}(+2 \text{ if T2D yes}, +2 \text{ if female}) \tag{5}$$

HSI values above 36 are deemed to indicate the presence of NAFLD.

External validation (UK Biobank cohort)

The UK Biobank cohort [38] was used to validate the clinical prediction models (models 1 and 2) derived using IMI DIRECT data (UK Biobank application ID: 18274). The same protocol and procedure have been used to quantify MRI-derived liver fat in IMI DIRECT and UK Biobank [18]. In addition, we validated the FLI and HSI using UK Biobank data. Field numbers for the UK Biobank variables used in the validation step can be found in the S4 Table. The data analysis procedures used for the UK Biobank validation analyses mirror those used in IMI DIRECT (as described above).

This study is reported as per the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guideline (S1 STROBE Checklist) and the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guideline (S1 TRIPOD Checklist).

Results

The following section describes fatty liver prediction models that are likely to suit different scenarios. We focus on a basic model (model 1), which includes variables that are widely available in both clinical and research settings. Models 2 and 3 focus on variables that could in principle be accessed within the clinical context, but that are not routinely available in the clinical setting at this time. Model 4 includes clinical variables, more detailed measures of glucose and insulin dynamics, and physical activity. Models 5 to 18 are more advanced models that include omics predictor variables alone or in combination with clinical predictor variables. See S3 Table for a full description of models.

Clinical models 1–3

We developed models 1–3 for NAFLD prediction, graded by perceived data accessibility for clinicians. These models were developed on the full dataset without applying any statistical procedures for feature selection. Model 1 includes 6 non-serological input variables: waist circumference, BMI, systolic blood pressure (SBP), diastolic blood pressure (DBP), alcohol consumption, and diabetes status. Model 2 includes 8 input variables: waist circumference, BMI, TG, ALT, AST, fasting glucose (or hemoglobin A1C [HbA1c] if fasting glucose is not available), alcohol consumption, and diabetes status. Model 3 includes 9 variables: waist circumference, BMI, TG, ALT, AST, fasting glucose, fasting insulin, alcohol consumption, and diabetes status. Clinical models 1–3 along with the FLI, HSI, and NAFLD-LFS were applied to the non-diabetes and diabetes cohort datasets separately, as well as to the combined cohort dataset; the ROCAUC results are presented in Fig 3. Model 1 yielded a ROCAUC of 0.73 (95% CI 0.72, 0.75; $p <$ 0.001) in the combined cohort. Adding serological variables to model 2 (with either fasting glucose or HbA1c) for the combined cohort yielded a ROCAUC of 0.79 (95% CI 0.78, 0.80; $p <$ 0.001). Model 3 (fasting insulin added) yielded a ROCAUC of 0.82 (95% CI 0.81, 0.83; $p <$ 0.001) in the combined cohort. The FLI, HSI, and NAFLD-LFS had ROCAUCs of 0.75 (95% CI 0.73, 0.78; $p < 0.001$), 0.75 (95% CI 0.72, 0.77; $p < 0.001$), and 0.79 (95% CI 0.76, 0.81; $p <$ 0.001), respectively, in the combined cohort. The predictive performance of clinical models 1–3, FLI, HSI, and NAFLD-LFS in the non-diabetes and diabetes cohorts is presented in S5 Table.

Performance metrics

We further investigated sensitivity, specificity, balanced accuracy, and F1 score (a score considering sensitivity and precision combined). These measurements were calculated for different cutoffs applied to the output of the random forest model (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9) using clinical models 1–3 in the diabetes, non-diabetes, and combined cohorts. The performance metrics for models 1 and 2 are presented in S6 and S7 Figs, and the metrics for model 3 are presented in Fig 4. We aimed to find the optimal cutoff for these models based on the cross-validated balanced accuracy. The highest balanced accuracy for models 1–3 in the non-diabetes, diabetes, and combined cohorts was observed at cutoffs of 0.4, 0.6, and 0.4, respectively (see Table 2).

Measurements of sensitivity, specificity, F1 score, and balanced accuracy were computed for the FLI, HSI, and NAFLD-LFS and compared with those of clinical models1–3. These measurements were computed at the optimal cutoff values for these indices: −0.640 for NAFLD-LFS, 60 for the FLI, and 36 for the HSI. A comprehensive overview of the prediction models' performance metrics for all of the fatty liver indices listed above is shown in Table 2.

Validation in UK Biobank and IMI DIRECT

Liver fat data were available in 4,617 UK Biobank participants (1,011 with $\geq$5% liver fat and 3,606 with $<$5% liver fat). Of these individuals, 4,609 had all the required variables to replicate clinical model 1. To perform model 2, with either fasting glucose or HbA1c, 3,807 participants had data available for a complete case analysis. Given the limited availability of variables in the UK Biobank dataset, only models 1 and 2 of the NAFLD prediction models we developed could be externally validated. To facilitate this validation analysis, the random forest models developed in the IMI DIRECT cohorts were used to predict the liver fat category (participants with fatty liver versus non-fatty liver) for the UK Biobank participants. The performance of the FLI and HSI was also tested in the UK Biobank cohort. We validated both models 1 and 2 in the UK Biobank cohort with a similar ROCAUC as seen in the IMI DIRECT dataset. The

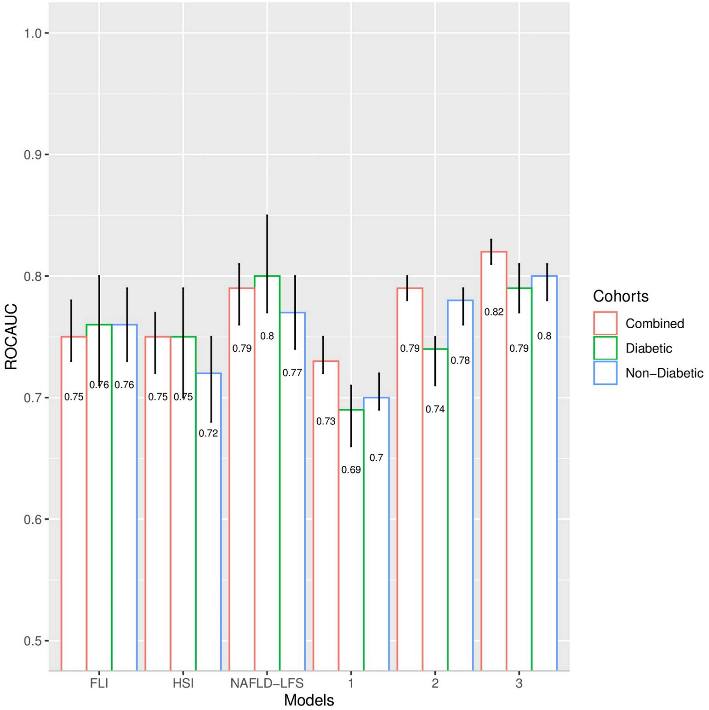


Fig 3. Receiver operating characteristic area under the curve (ROCAUC) with 95% confidence interval (error bars) for clinical models 1–3, fatty liver index (FLI), hepatic steatosis index (HSI), and non-alcoholic fatty liver disease liver fat score (NAFLD-LFS) in the IMI DIRECT cohorts. Model 1 includes 6 non-serological input variables: waist circumference, body mass index(BMI), mean systolic blood pressure, mean diastolic blood pressure, alcohol consumption, and diabetes status. Model 2 includes 8 input variables: waist circumference, BMI, fasting triglycerides (TG), alanine transaminase (ALT), aspartate transaminase (AST), fasting glucose (or hemoglobin A1C if fasting glucose is not available), alcohol consumption, and diabetes status. Model 3 includes 9 variables: waist circumference, BMI, TG, ALT, AST, fasting glucose, fasting insulin, alcohol consumption, and diabetes status. The FLI uses TG, waist circumference, BMI, and gamma-glutamyl transpeptidase. NAFLD-FLS was calculated using fasting insulin, AST, ALT, type 2 diabetes (T2D), and metabolic syndrome defined according to the International Diabetes Federation. The HSI uses BMI, sex, T2D diagnosis (yes/no), and the ratio of ALT to AST.

https://doi.org/10.1371/journal.pmed.1003149.g003

ROCAUCs were 0.71 (95% CI 0.69, 0.73; $p < 0.001$), 0.79 (95% CI 0.77, 0.80; $p < 0.001$), and 0.78 (95% CI 0.76, 0.79; $p < 0.001$) for model 1, model 2 with fasting glucose, and model 2 with HbA1c, respectively. The FLI had a ROCAUC of 0.78 (95% CI 0.76, 0.80; $p < 0.001$), which is similar to the ROCAUC of model 2. The HSI yielded a ROCAUC of 0.76 (95% CI 0.75, 0.78; $p < 0.001$).

Measurements of sensitivity, specificity, F1 score, and balanced accuracy were also computed at the optimal cutoff values for these models: 0.4 for clinical models 1 and 2, 60 for the FLI, and 36 for the HSI (see Table 2).

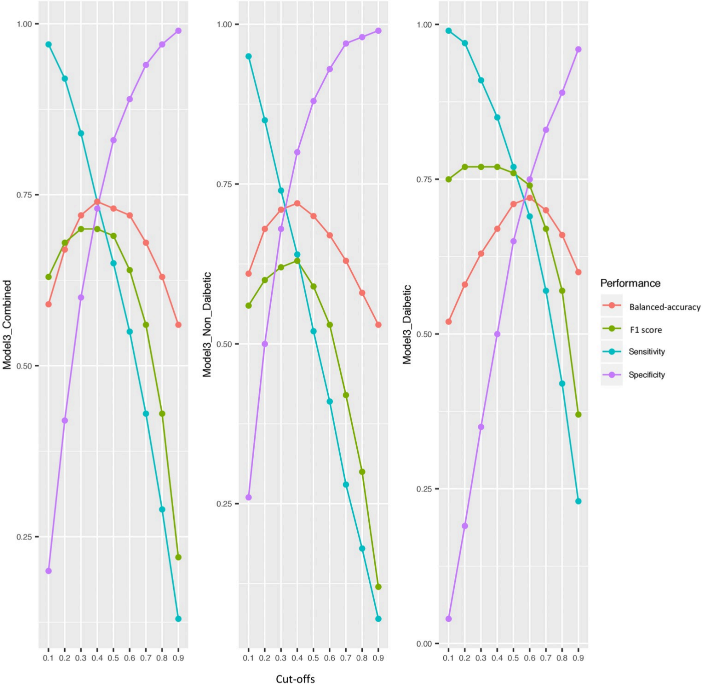


Fig 4. Measurements of sensitivity, specificity, F1 (a score considering sensitivity and precision combined), and balanced accuracy at different cutoffs for model 3 in the diabetes, non-diabetes, and combined cohorts of IMI-DIRECT. The measurements are calculated by defining the predicted probabilities of fatty liver equal to or above these cutoffs as fatty liver, and below as non-fatty liver. Model 3 includes 9 variables: waist circumference, body mass index, fasting triglycerides, alanine transaminase, aspartate transaminase, fasting glucose, fasting insulin, alcohol consumption, and diabetes status.

https://doi.org/10.1371/journal.pmed.1003149.g004

Clinical model 4 and omics models 5–14

More advanced models using omics data were also developed. These models were generated using the omics features selected by LASSO in the combined cohort. The models include only omics or include omics plus 22 clinical variables as the input variables. Twenty-one of these clinical variables were selected based on the pairwise Pearson correlation matrix: BMI, waist circumference, SBP, DBP, alcohol consumption, ALT, AST, GGTP, HDL, TG, fasting glucose, 2-hour glucose, HbA1c, fasting insulin, 2-hour insulin, insulin secretion at the beginning of the carbohydrate challenge test (OGTT or MMTT), 2-hour oral glucose insulin sensitivity index (OGIS), mean insulin clearance during the OGTT/MTT, fasting glucagon concentration, fasting plasma total GLP-1 concentration, and mean physical activity intensity. Diabetes status (non-diabetes/diabetes) was also included as a clinical predictor in the models, given that analyses were undertaken in the combined diabetes and non-diabetes cohort. The ROCAUCs for models 4–14 are shown in Fig 5. The clinical model with the 22 selected clinical

Table 2. An overview of the prediction models' performance metrics for clinical models 1–3, fatty liver index (FLI), hepatic steatosis index (HIS), and non-alcoholic fatty liver disease liver fat score (NAFLD-LFS) in the IMI DIRECT and UK Biobank datasets.

Cohort and model	Cutoff	Sensitivity	Specificity	F1 score	Balanced accuracy
Non-diabetes (IMI DIRECT)					
Model 1	0.4	0.51	0.75	0.51	0.63
Model 2	0.4	0.60	0.79	0.59	0.69
Model 3	0.4	0.64	0.80	0.63	0.72
FLI	60	0.89	0.41	0.58	0.65
HSI	36	0.62	0.68	0.55	0.65
NAFLD-LFS	−0.64	1	0.04	0.51	0.52
Diabetes (IMI DIRECT)					
Model 1	0.6	0.63	0.64	0.67	0.64
Model 2	0.6	0.65	0.68	0.69	0.67
Model 3	0.6	0.69	0.75	0.74	0.72
FLI	60	0.77	0.54	0.73	0.66
HSI	36	0.83	0.48	0.75	0.65
NAFLD-LFS	−0.64	1	0.01	0.73	0.50
Combined (IMI DIRECT)					
Model 1	0.4	0.67	0.65	0.62	0.66
Model 2	0.4	0.72	0.69	0.67	0.71
Model 3	0.4	0.74	0.73	0.70	0.74
FLI	60	0.84	0.44	0.64	0.64
HSI	36	0.71	0.63	0.64	0.67
NAFLD-LFS	−0.64	1	0	0.58	0.50
UK Biobank					
Model 1	0.4	0.49	0.78	0.43	0.63
Model 2	0.4	0.67	0.74	0.52	0.71
FLI	60	0.62	0.76	0.50	0.69
HSI	36	0.66	0.72	0.50	0.69

Model 1 includes 6 non-serological input variables: waist circumference, body mass index (BMI), mean systolic blood pressure, mean diastolic blood pressure, alcohol consumption, and diabetes status. Model 2 includes 8 input variables: waist circumference, BMI, fasting triglycerides (TG), alanine transaminase (ALT), aspartate transaminase (AST), fasting glucose (or hemoglobin A1C if fasting glucose is not available), alcohol consumption, and diabetes status. Model 3 includes 9 variables: waist circumference, BMI, TG, ALT, AST, fasting glucose, fasting insulin, alcohol consumption, and diabetes status. The FLI uses TG, waist circumference, BMI, and gamma-glutamyl transpeptidase. NAFLD-FLS was calculated using fasting insulin, AST, ALT, type 2 diabetes (T2D), and metabolic syndrome defined according to the International Diabetes Federation. The HSI uses BMI, sex, T2D diagnosis (yes/no), and the ratio of ALT to AST.

https://doi.org/10.1371/journal.pmed.1003149.t002

variables (model 4) yielded a ROCAUC of 0.79 (95% CI 0.76, 0.81; $p < 0.001$). Omics models with only the genetic (model 5), transcriptomic (model 7), proteomic (model 9), and targeted metabolomic (model 11) data as input variables resulted in ROCAUCs of 0.67 (95% CI 0.65, 0.70; $p < 0.001$), 0.72 (95% CI 0.69, 0.74; $p < 0.001$), 0.74 (95% CI 0.71, 0.76; $p < 0.001$), and 0.70 (95% CI 0.67, 0.72; $p < 0.001$), respectively. Including all the omics variables in one model (model 13) resulted in a ROCAUC of 0.82 (95% CI 0.80, 0.84; $p < 0.001$). Adding the clinical variables to each omics model improved the prediction ability; models with the clinical variables plus genetic (model 6), transcriptomic (model 8), exploratory proteomic (model 10), and targeted metabolomic (model 12) data resulted in ROCAUCs of 0.82 (95% CI 0.80, 0.84; $p < 0.001$), 0.81 (95% CI 0.79, 0.83; $p < 0.001$), 0.80 (95% CI 0.78, 0.83; $p < 0.001$), and 0.80 (95% CI 0.77, 0.82; $p < 0.001$), respectively. The highest performance was observed for model 14 (ROCAUC of 0.84; 95% CI 0.82, 0.86; $p < 0.001$). The variable importance for model 14 from

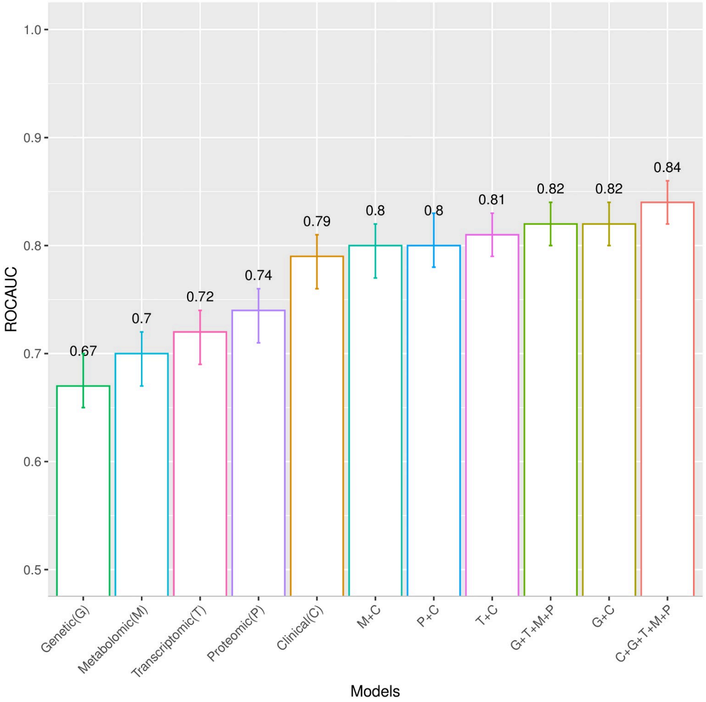


Fig 5. Receiver operating characteristic area under the curve (ROCAUC) with 95% confidence interval for the clinical model and the omics separately or in combination with the clinical model in the IMI DIRECT combined cohort. Clinical (C), model 4, with the 22 selected clinical variables. Genetic (G), model 5, with 23 SNPs. C+G, model 6, with clinical plus genetic variables. Transcriptomic (T), model 7, with 93 protein-coding genes. T+C, model 8, with transcriptomic plus clinical variables. Proteomic (P), model 9, with 22 proteins from exploratory proteomics. P+C, model 10, with proteomic plus clinical variables. Metabolomic (M), model 11, with 25 metabolites from targeted metabolomics. M+C, model 12, with metabolomic plus clinical variables. G+T+M+P, model 13, with all omics together. C+G+T+M+P, model 14, with all the omics combined with the clinical model.

https://doi.org/10.1371/journal.pmed.1003149.g005

the permutation accuracy importance measure, presented in Fig 6, shows that measures of insulin secretion rank amongst those having the highest variable importance of all input variables. Moreover, the importance list derived from EFS, shown in S20 Fig, is highly consistent with that derived from the random forest analysis. Rankings for the individual clinical and omics variables from the permutation accuracy importance measure and EFS are presented in S8–S19 Figs. The minor inconsistencies in results from the 2 approaches are likely to reflect the ability of the random forest analysis to detect variables that interact with others, which the linear methods are not designed to detect.

Additional proteomic and metabolomic analyses (models 15–18)

Data from targeted proteomic and untargeted metabolomic data were further utilized to develop the omics models separately or in combination with the clinical data. However,

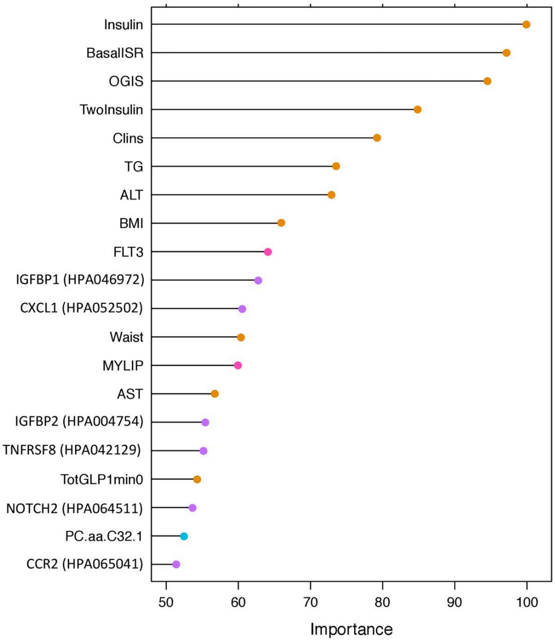


Fig 6. Variable importance for the advanced model 14 with 185 omics and clinical input variables (clinical = 22, genetic = 23, transcriptomic = 93, exploratory proteomic = 22, and targeted metabolomic = 25). The *y*-axis shows the top 20 predictors in the model. The *x*-axis shows the variable importance calculated, via a permutation accuracy importance measure using random forest analysis, as the difference in prediction accuracy before and after the permutation for each variable scaled by the standard error. ALT, alanine transaminase; AST, aspartate transaminase; BasalISR, insulin secretion at the beginning of the oral glucose tolerance test/mixed-meal tolerance test; BMI, body mass index; Clins, mean insulin clearance during the oral glucose tolerance test/mixed-meal tolerance test calculated as (mean insulin secretion)/(mean insulin concentration); FLT3, fetal liver tyrosine kinase-3; Insulin, fasting insulin from venous plasma samples; MYLIP, myosin regulatory light chain interacting protein; OGIS, oral glucose insulin sensitivity index according to the method of Mari et al. [24]; TG, fasting triglycerides; TotGLP1min0, concentration of fasting total GLP-1 in plasma; TwoInsulin, 2-hour insulin after oral glucose tolerance test/mixed meal tolerance test.

https://doi.org/10.1371/journal.pmed.1003149.g006

as some participants lacked these omics data, their models were developed using a smaller data subset and were, hence, not included in the advanced (model 14) analyses. The complete case analysis was primarily defined on the availability of the 22 selected clinical variables ($n = 1{,}049$). Within this complete case set, 511 had a complete set of untargeted metabolomic data, and 686 had a complete set of targeted proteomic data. The models with targeted proteomic data only and with proteomic and clinical variables combined resulted in ROCAUCs of 0.81 (95% CI 0.78, 0.84; $p < 0.001$) and 0.84 (95% CI 0.81, 0.87; $p < 0.001$), respectively. The untargeted metabolomic model alone had a ROCAUC of 0.66 (95% CI 0.63, 0.69; $p < 0.001$), which increased to 0.78 (95% CI 0.75, 0.80; $p < 0.001$) when the 22 clinical variables were added.

Discussion

Using data from the IMI DIRECT consortium, we developed 18 diagnostic models for early-stage NAFLD. These models were developed to reflect different scenarios within which they might be used: These included both clinical and research settings, with the more complex (and less accessible) models having the greatest predictive ability. The models were successfully validated in the UK Biobank where data permitted such analysis (clinical models 1 and 2). Overall, the basic clinical variables proved to be stronger predictors of fatty liver than more complex omics data, although adding omics data yielded the most powerful model, with very good cross-validated predictive ability (ROCAUC = 0.84).

NAFLD is etiologically complex, rendering its prevention and treatment difficult, and diagnosis can require invasive and/or relatively expensive procedures. Thus, noninvasive and cost-effective prediction models with good sensitivity and specificity are much needed. This is especially important because if NAFLD is detected early, treatment through lifestyle interventions can be highly effective [39]. However, simple steatosis is usually asymptomatic, and many patients only come to the attention of hepatologists when serious complications arise [40].

To date, several prediction models have been developed to facilitate the diagnosis of steatosis (thoroughly reviewed elsewhere [13]). The FLI is one of the most well-established and commonly used fatty liver indices, initially developed using ultrasound-derived hepatic steatosis data [34]. The FLI yielded similar predictive performance in the diabetes and non-diabetes cohorts of IMI DIRECT (both ROCAUCs approximately 0.75).

Though commonly used for liver fat prediction, the FLI has a similar discriminative ability as waist circumference alone [41]. Better discrimination can be obtained by incorporating additional serological and hemostatic measures, which is the case with NAFLD-LFS [14], the SteatoTest [42], and the HSI [35], for example. Notwithstanding the added complexity and cost of these scores, the FLI, HSI, and NAFLD-LFS yielded similar predictive ability in a series of liver-biopsy-diagnosed NAFLD cases (n = 324) [36].

Omics technologies have been used in a small number of studies to identify molecular biomarkers of NAFLD [43–45]. These include tests utilizing genetic data such as FibroGENE for staging liver fibrosis [46], and tests using metabolomic data derived from liver tissue to differentiate simple hepatitis from NASH [47], as well as a multi-component NAFLD classifier using genomic, proteomic, and phenomic data [45]. Machine learning models based on lipidomic, glycomic, and free fatty acid data were also developed for the diagnosis of NASH and liver fibrosis [48,49]. In a recent retrospective case series of patients with obesity, EFS was applied for feature selection, using a set of sociodemographic and serum variables to predict the presence or absence of NASH [50].

Using data from IMI DIRECT, we explored the predictive ability of genetic, transcriptomic, proteomic, and metabolomic data from blood in the diagnosis of NAFLD. The top 20 features of each omics model are presented in S9–S14 Figs. The details of the LASSO selected features are summarized in S7 Table. Reassuringly, several of the features that ranked highest have been previously described for their association with liver fat content or closely related traits; these include *PNPLA3* gene variants [44,51], fetal liver tyrosine kinase-3 (FLT3) transcripts [52], IGFBP1 [53–55] and lipoprotein lipase (Lpl) [56] proteins, and the metabolite glutamate [57]. In the analysis of the targeted metabolites, phosphatidylcholines (including PC.aa.C32, PC.aa.C38, PC.aa.C40, and PC.aa.C42), glycerophospholipids, and valine were amongst the highest-ranked metabolites that are known for their correlation with NAFLD and metabolic disorders [58,59]. For exploratory proteomics, the most important variables were proteins secreted into the blood, expressed by the liver as well as those leaking from the blood cells [60]. The prediction model that only included targeted proteomic data (model 15) performed well

(ROCAUC = 0.81), rendering it an interesting candidate biomarker for future clinical tests. Among the top 20 most important proteins were many secreted into the blood or leaked by the liver, as well as the pancreas, fat, or muscle tissue [61].

Our intention by including all features in the same model (model 14) was to maximize predictive power by leveraging interactions between features. Moreover, we explored the value of boosting ensemble algorithms for each data source. The purpose of this was to enhance predictions. We trained a stochastic gradient boosting algorithm for each data source separately and then applied a weighted averaging on the probabilities of observations. The optimal weighting was observed at 0.5 for the clinical data and 0.125 for each omics data layer (i.e., genetic, transcriptomic, exploratory proteomic, and targeted metabolomic). The ensemble prediction model of omics and clinical datasets resulted in a ROCAUC of 0.83 (95% CI 0.78, 0.87; $p < 0.001$), which is not materially different from the ROCAUC derived for the advanced model 14 (described in the Results), which includes all the omics and clinical features in a single model (ROCAUC of 0.84; 95% CI 0.82, 0.86; $p < 0.001$). The models developed here may be useful for screening for NAFLD, and this should be evaluated in future clinical studies.

In order to stratify people into groups of those unlikely and likely to have NAFLD, the latter of whom might subsequently undergo more invasive and/or costly clinical assessments, it would be important for the prediction model to have high sensitivity. However, the predictive utility of a given model can be further improved by selecting model cutoffs that optimize sensitivity or specificity, as the 2 metrics rarely perform optimally at the same cutoff. This issue was apparent for models 1–3 in the current analyses, where we selected cutoffs that maximized balanced accuracy (considering both sensitivity and specificity); these features are especially important in screening algorithms, where the cost of false negatives can be high. Models 1–3 resulted in higher sensitivity in the diabetes cohort than the non-diabetes cohort, whereas the specificity was higher in the non-diabetes and combined cohorts than in the diabetes cohort.

The linear LASSO method was used to minimize overfitting that can occur with high-dimensionality data, while random forest analysis was used to identify nonlinear associations where data structure allowed. We also considered several other machine learning approaches including generalized linear model, stochastic gradient boosting, support vector machines, and *k*-nearest neighbor, and the random forest analysis yielded similar or better results compared with any of these other approaches (see S6 Table).

A limitation of the analytical approach used here is that the methods required a complete case analysis, which diminishes sample size considerably; although imputing missing data here helped preserve sample size, it did not improve the prediction ability of the models, and we hence elected to use the complete case analysis. Heavy alcohol consumption is a key determinant of fatty liver, but is unlikely to be a major etiological factor in IMI DIRECT owing to the demographics of this cohort. Nevertheless, a further limitation of this analysis is that alcohol intake was self-reported and may lack validity. To address this limitation, we removed all self-reported heavy alcohol consumers from the UK Biobank cohort and undertook sensitivity analyses, but this did not materially affect the results.

Here we considered lifestyle variables, but not medications. The use of medicines affecting liver fat is likely to be less in the non-diabetes than in the diabetes cohort, yet the models fit better in the latter, suggesting that glucose-lowering medication use in the IMI DIRECT cohorts did not have a major detrimental impact on prediction model performance.

A further consideration for future work is the impact lifestyle and medications are likely to have on the prediction of NAFLD. Furthermore, this study was undertaken in people of European ancestry, and the extent to which the results will generalize to other ethnic groups is unknown. Moreover, the prediction is for a binary liver fat outcome ($<5\%$ or $\geq 5\%$), and neither fully quantifies liver fat volume nor elucidates the degree of liver damage (cirrhosis). These key limitations of the current work will be the focus of future research.

Our finding that a model focused on proteomic data yielded high predictive utility may warrant further investigation. Our analysis also suggests that insulin sensitivity and beta-cell dysfunction may be involved in liver fat accumulation, which are at present not considered as features of conventional NAFLD risk models.

In summary, we have developed prediction models for NAFLD that may have utility for clinical diagnosis and research investigations alike. A web interface for the diagnosis of NAFLD was developed using the findings described above (https://www.predictliverfat.org), which renders clinical models 1–3 developed here accessible for the wider community of clinicians and researchers.

Supporting information

S1 Fig. Violin plot showing the distribution of liver fat percentage for the diabetes and non-diabetes cohorts of IMI DIRECT.
(TIFF)

S2 Fig. Distribution of liver fat percentage among the different centers contributing to the IMI DIRECT cohorts.
(TIFF)

S3 Fig. Manhattan plot showing SNPs associated with liver fat level (approximately 18 million imputed SNPs) in the IMI DIRECT cohorts. The chromosomal position is plotted on the *x*-axis, and the statistical significance of association for each SNP is plotted on the *y*-axis. Red line indicates genome-wide significance level (5×10^{-8}).
(TIFF)

S4 Fig. Quantile–quantile (QQ) plot showing results of genome-wide association study (GWAS) for liver fat content in the IMI DIRECT consortium (1,514 individuals). The *x*-axis illustrates the expected distribution of *p*-values from the association test across all SNPs, and the *y*-axis shows the observed *p*-values.
(TIFF)

S5 Fig. Details of the feature selection step for models 4–14 and models 15–18 using the IMI DIRECT data. Models 4–14 (blue box); models 15–18 (green box).
(TIFF)

S6 Fig. Measurements of sensitivity, specificity, F1 score (a score considering sensitivity and precision combined), and balanced accuracy at different cutoffs for model 1 in the diabetes, non-diabetes, and combined cohorts of IMI DIRECT.
(TIFF)

S7 Fig. Measurements of sensitivity, specificity, F1 score (a score considering sensitivity and precision combined), and balanced accuracy at different cutoffs for model 2 in the diabetes, non-diabetes, and combined cohorts of IMI DIRECT.
(TIFF)

S8 Fig. Variable importance for the clinical model via a permutation accuracy importance measure. The *y*-axis shows the top 20 predictors in the model. The *x*-axis shows the variable importance, calculated using random forest analysis as the difference in prediction accuracy before and after the permutation for each variable scaled by the standard error. ALT, alanine transaminase; AST, aspartate transaminase; BasalISR, insulin secretion at the beginning of the

OGTT/MMTT; BMI, body mass index; Clins, mean insulin clearance during the OGTT/MMTT calculated as (mean insulin secretion)/(mean insulin concentration); DBP, diastolic blood pressure; Diabetes_status2, non-diabetes/diabetes; GGTP, gamma-glutamyl transpeptidase; Glucagonmin0, fasting glucagon concentration; Glucose, fasting glucose from venous plasma samples; HbA1c, hemoglobin A1C; HDL, fasting high-density lipoprotein cholesterol; Insulin, fasting insulin from venous plasma samples; OGIS, oral glucose insulin sensitivity index according to the method of Mari et al. [24]; PA_intensity_mean, mean high-pass-filtered vector magnitude physical activity intensity; SBP, systolic blood pressure; TG, fasting triglycerides; TotGLP1min0, concentration of fasting total GLP-1 in plasma; TwoGlucose, 2-hour glucose after OGTT/MMTT; TwoInsulin, 2-hour insulin.
(TIFF)

S9 Fig. Variable importance for the genetic model via a permutation accuracy importance measure. The *y*-axis shows the top 20 predictors in the model. The *x*-axis shows the variable importance, calculated using random forest analysis as the difference in prediction accuracy before and after the permutation for each variable scaled by the standard error.
(TIFF)

S10 Fig. Variable importance for the transcriptomic model via a permutation accuracy importance measure. The *y*-axis shows the top 20 predictors in the model. The *x*-axis shows the variable importance, calculated using random forest analysis as the difference in prediction accuracy before and after the permutation for each variable scaled by the standard error.
(TIFF)

S11 Fig. Variable importance for the exploratory proteomic model via a "permutation accuracy importance" measure. The *y*-axis shows the top 20 predictors in the model. The *x*-axis shows the variable importance, calculated using random forest as the difference in prediction accuracy before and after the permutation for each variable scaled by the standard error.
(TIFF)

S12 Fig. Variable importance for the targeted metabolomic model via a permutation accuracy importance measure. The *y*-axis shows the top 20 predictors in the model. The *x*-axis shows the variable importance, calculated using random forest analysis as the difference in prediction accuracy before and after the permutation for each variable scaled by the standard error.
(TIFF)

S13 Fig. Variable importance for the targeted proteomic model via a permutation accuracy importance measure. The *y*-axis shows the top 20 predictors in the model. The *x*-axis shows the variable importance, calculated using random forest analysis as the difference in prediction accuracy before and after the permutation for each variable scaled by the standard error.
(TIFF)

S14 Fig. Variable importance for the untargeted metabolomic model via a permutation accuracy importance measure. The *y*-axis shows the top 20 predictors in the model. The *x*-axis shows the variable importance, calculated using random forest analysis as the difference in prediction accuracy before and after the permutation for each variable scaled by the standard error.
(TIFF)

S15 Fig. Variable importance for the clinical model derived from ensemble feature selection (EFS). The *y*-axis shows the 22 clinical variables ordered by importance value. The *x*-axis shows

the cumulative importance values, calculated via an ensemble of feature selection methods including Spearman's rank correlation test (S_cor), Pearson's product moment correlation test (P_cor), beta-values of logistic regression (LogReg), error-rate-based variable importance measure (ER_RF), and Gini-index-based variable importance measure (Gini_RF). ALT, alanine transaminase; AST, aspartate transaminase; BasalISR, insulin secretion at the beginning of the OGTT/MMTT; BMI, body mass index; Clins, mean insulin clearance during the OGTT/ MMTT calculated as (mean insulin secretion)/(mean insulin concentration); DBP, diastolic blood pressure; Diabetes status, non-diabetes/diabetes; GGTP, gamma-glutamyl transpeptidase; Glucagonmin0, fasting glucagon concentration; Glucose, fasting glucose from venous plasma samples; HbA1c, hemoglobin A1C; HDL, fasting high-density lipoprotein cholesterol; Insulin, fasting insulin from venous plasma samples; OGIS, oral glucose insulin sensitivity index according to the method of Mari et al. [24]; PA_intensity_mean, mean high-pass-filtered vector magnitude physical activity intensity; SBP, systolic blood pressure; TG, fasting triglycerides; TotGLP1min0, concentration of fasting total GLP-1 in plasma; TwoGlucose, 2-hour glucose after OGTT/MMTT; TwoInsulin, 2-hour insulin.
(TIF)

S16 Fig. Variable importance for the genetic model derived from EFS. The *y*-axis shows the 23 genetic variables ordered by importance value. The *x*-axis shows the cumulative importance values, calculated via an ensemble of feature selection methods including Spearman's rank correlation test (S_cor), Pearson's product moment correlation test (P_cor), beta-values of logistic regression (LogReg), error-rate-based variable importance measure (ER_RF), and Gini-index-based variable importance measure (Gini_RF).
(TIFF)

S17 Fig. Variable importance for the transcriptomic model derived from EFS. The *y*-axis shows the 93 transcriptomic variables ordered by importance value. The *x*-axis shows the cumulative importance values, calculated via an ensemble of feature selection methods including Spearman's rank correlation test (S_cor), Pearson's product moment correlation test (P_cor), beta-values of logistic regression (LogReg), error-rate-based variable importance measure (ER_RF), and Gini-index-based variable importance measure (Gini_RF).
(TIFF)

S18 Fig. Variable importance for the exploratory proteomic model derived from EFS. The *y*-axis shows the 22 exploratory proteomic variables ordered by importance value. The *x*-axis shows the cumulative importance values, calculated via an ensemble of feature selection methods including Spearman's rank correlation test (S_cor), Pearson's product moment correlation test (P_cor), beta-values of logistic regression (LogReg), error-rate-based variable importance measure (ER_RF), and Gini-index-based variable importance measure (Gini_RF).
(TIFF)

S19 Fig. Variable importance for the targeted metabolomic model derived from EFS. The *y*-axis shows the 25 targeted metabolomic variables ordered by importance value. The *x*-axis shows the cumulative importance values, calculated via an ensemble of feature selection methods including Spearman's rank correlation test (S_cor), Pearson's product moment correlation test (P_cor), beta-values of logistic regression (LogReg), error-rate-based variable importance measure (ER_RF), and Gini-index-based variable importance measure (Gini_RF).
(TIFF)

S20 Fig. Variable importance for the clinical plus multi-omics model (clinical = 22, genetic = 23, transcriptomic = 93, exploratory proteomic = 22, and targeted metabolomic = 25) derived from EFS. The *y*-axis shows the top 20 predictors in the model. The *x*-axis shows the cumulative importance values, calculated via an ensemble of feature selection methods including Spearman's rank correlation test (S_cor), Pearson's product moment correlation test (P_cor), beta-values of logistic regression (LogReg), error-rate-based variable importance measure (ER_RF), and Gini-index-based variable importance measure (Gini_RF). ALT, alanine transaminase; AST, aspartate transaminase; BasalISR, insulin secretion at the beginning of the OGTT/MMTT; Clins, mean insulin clearance during the OGTT/MMTT calculated as (mean insulin secretion)/(mean insulin concentration); Insulin, fasting insulin from venous plasma samples; OGIS, oral glucose insulin sensitivity index according to the method of Mari et al. [24]; TG, fasting triglycerides; TotGLP1min0, concentration of fasting total GLP-1 in plasma; TwoInsulin, 2-hour insulin after OGTT/MMTT.
(TIFF)

S1 STROBE Checklist. The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist.
(DOCX)

S1 Table. The list of the clinical input variables with the abbreviation used in the analyses and their meaning.
(XLSX)

S2 Table. Characteristics of the study in the non-diabetes, diabetes, and combined cohorts separated for participants from IMI DIRECT who had MRI data versus those who did not have MRI data. Values are median (interquartile range) unless otherwise specified. ALT, alanine transaminase; AST, aspartate transaminase; BMI, body mass index; DBP, diastolic blood pressure; HbA1c, hemoglobin A1C; SBP, systolic blood pressure.
(XLSX)

S3 Table. Variables used to construct each of the NAFLD prediction models developed in IMI DIRECT. ALT, alanine transaminase; AST, aspartate transaminase; BMI, body mass index; DBP, diastolic blood pressure; GGTP, gamma-glutamyl transpeptidase; Glucagonmin0, fasting glucagon concentration; HbA1c, hemoglobin A1C; HDL, fasting high-density lipoprotein cholesterol; MMTT, mixed meal tolerance test; OGIS, oral glucose insulin sensitivity index according to the method of Mari et al. [24]; OGTT, oral glucose tolerance test; PA_intensity_mean, mean high-pass-filtered vector magnitude physical activity intensity; SBP, systolic blood pressure; TotGLP1min0, concentration of fasting total GLP-1 in plasma.
(XLSX)

S4 Table. UK Biobank field number with the description used in the analyses.
(XLSX)

S5 Table. Receiver operating characteristic area under the curve (ROCAUC) with 95% confidence interval for clinical models 1–3, fatty liver index (FLI), hepatic steatosis index (HSI), and NAFLD liver fat score (NAFLD-LFS) in the non-diabetes and diabetes cohorts of the IMI DIRECT separately.
(XLSX)

S6 Table. Receiver operating characteristic area under the curve (ROCAUC) with 95% confidence interval of each separate dataset obtained from random forest (RF), generalized linear model (GLM), stochastic gradient boosting (GBM), support vector machine (SVM),

and *k*-nearest neighbor (KNN) analyses in the cross-validated test data of the IMI DIRECT combined cohort.
(XLSX)

S7 Table. The details of the LASSO (least absolute shrinkage and selection operator)–selected features of the omics layers in separate sheets (genetic, transcriptomic, exploratory proteomic, targeted proteomic, targeted Metabolomic, and untargeted metabolomic).
(XLSX)

S1 Text. QC of the transcriptomic, proteomic, and metabolomic variables in the IMI DIRECT datasets.
(DOCX)

S1 TRIPOD Checklist. The TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) checklist.
(DOCX)

Acknowledgments

We thank Mattias Borell for developing, logistical support, and advice related to the web interface. We thank all the participants and study center staff in IMI DIRECT for their contribution to the study. We thank all the participants in the UK Biobank. This research was conducted using the UK Biobank resource (application ID: 18274). For the proteomic analyses, we thank the entire staff of the Human Protein Atlas, the Plasma Profiling Facility at Science for Life Laboratory, and in particular Elin Birgersson, Annika Bendes, and Eni Andersson for technical assistance. We thank C. Prehn (HMGU) for laboratory work related to the metabolomic data.

Author Contributions

Conceptualization: Naeimeh Atabaki-Pasdar, Mattias Ohlsson, Jimmy D. Bell, Imre Pavo, Paul W. Franks.

Data curation: Naeimeh Atabaki-Pasdar, Ana Viñuela, Mark Haid, Angus G. Jones, E. Louise Thomas, Robert W. Koivula, Azra Kurbasic, Juan Fernandez, Adem Y. Dawed, Ian M. Forgie, Timothy J. McDonald, Matilda Dale, Federico De Masi, Mun-Gwan Hong, Tarja Kokkola, Helle Krogh Pedersen, Anubha Mahajan, Sapna Sharma, Jerzy Adamski, Soren Brage, Søren Brunak, Emmanouil Dermitzakis, Gary Frost, Andrea Mari, Jochen M. Schwenk, Ramneek Gupta, Jimmy D. Bell.

Formal analysis: Naeimeh Atabaki-Pasdar, Ana Viñuela, Francesca Frau, Hugo Pomares-Millan.

Funding acquisition: Jerzy Adamski, Søren Brunak, Torben Hansen, Markku Laakso, Oluf Pedersen, Martin Ridderstråle, Hartmut Ruetten, Andrew T. Hattersley, Mark Walker, Mark I. McCarthy, Ewan R. Pearson, Imre Pavo, Paul W. Franks.

Investigation: Robert W. Koivula, Timothy J. McDonald, Femke Rutters, Henna Cederberg, Kristine H. Allin, Alison Heggie, Gwen Kennedy, Henrik Vestergaard, Soren Brage, Søren Brunak, Torben Hansen, Markku Laakso, Oluf Pedersen, Martin Ridderstråle, Hartmut Ruetten, Andrew T. Hattersley, Mark Walker, Joline W. J. Beulens, Ramneek Gupta, Mark I. McCarthy, Ewan R. Pearson, Imre Pavo, Paul W. Franks.

Methodology: Naeimeh Atabaki-Pasdar, Mattias Ohlsson, Ana Viñuela, Pascal M. Mutie, Hugo Fitipaldi, Giuseppe N. Giordano, Elizaveta Chabanova, Cecilia Engel Thomas, Tue H. Hansen, Petra J. M. Elders, Donna McEvoy, Francois Pattou, Violeta Raverdy, Ragna S. Häussler, Henrik S. Thomsen, Leen M. 't Hart, Petra B. Musholt, Andrea Mari, Jochen M. Schwenk, Jimmy D. Bell, Paul W. Franks.

Project administration: Robert W. Koivula, Giuseppe N. Giordano, Ian M. Forgie, Tue H. Hansen, Gwen Kennedy, Jerzy Adamski, Søren Brunak, Torben Hansen, Markku Laakso, Oluf Pedersen, Martin Ridderstråle, Hartmut Ruetten, Andrew T. Hattersley, Mark Walker, Joline W. J. Beulens, Jochen M. Schwenk, Mark I. McCarthy, Ewan R. Pearson, Jimmy D. Bell, Paul W. Franks.

Resources: Federico De Masi, Søren Brunak, Paul W. Franks.

Software: Naeimeh Atabaki-Pasdar, Paul W. Franks.

Supervision: Mattias Ohlsson, Paul W. Franks.

Visualization: Naeimeh Atabaki-Pasdar.

Writing – original draft: Naeimeh Atabaki-Pasdar, Paul W. Franks.

Writing – review & editing: Naeimeh Atabaki-Pasdar, Mattias Ohlsson, Ana Viñuela, Francesca Frau, Hugo Pomares-Millan, Mark Haid, Angus G. Jones, E. Louise Thomas, Robert W. Koivula, Azra Kurbasic, Pascal M. Mutie, Hugo Fitipaldi, Juan Fernandez, Adem Y. Dawed, Giuseppe N. Giordano, Ian M. Forgie, Timothy J. McDonald, Femke Rutters, Henna Cederberg, Elizaveta Chabanova, Matilda Dale, Federico De Masi, Cecilia Engel Thomas, Kristine H. Allin, Tue H. Hansen, Alison Heggie, Mun-Gwan Hong, Petra J. M. Elders, Gwen Kennedy, Tarja Kokkola, Helle Krogh Pedersen, Anubha Mahajan, Donna McEvoy, Francois Pattou, Violeta Raverdy, Ragna S. Häussler, Sapna Sharma, Henrik S. Thomsen, Jagadish Vangipurapu, Henrik Vestergaard, Leen M. 't Hart, Jerzy Adamski, Petra B. Musholt, Soren Brage, Søren Brunak, Emmanouil Dermitzakis, Gary Frost, Torben Hansen, Markku Laakso, Oluf Pedersen, Martin Ridderstråle, Hartmut Ruetten, Andrew T. Hattersley, Mark Walker, Joline W. J. Beulens, Andrea Mari, Jochen M. Schwenk, Ramneek Gupta, Mark I. McCarthy, Ewan R. Pearson, Jimmy D. Bell, Imre Pavo, Paul W. Franks.

References

1. Tilg H, Moschen AR. Insulin resistance, inflammation, and non-alcoholic fatty liver disease. Trends Endocrinol Metab. 2008; 19(10):371–9. https://doi.org/10.1016/j.tem.2008.08.005 PMID: 18929493
2. Sattar N, Gill JM. Type 2 diabetes as a disease of ectopic fat? BMC Med. 2014; 12:123. https://doi.org/10.1186/s12916-014-0123-4 PMID: 25159817
3. Sattar N, Forrest E, Preiss D. Non-alcoholic fatty liver disease. BMJ. 2014; 349:g4596. https://doi.org/10.1136/bmj.g4596 PMID: 25239614
4. Lucas C, Lucas G, Lucas N, Krzowska-Firych J, Tomasiewicz K. A systematic review of the present and future of non-alcoholic fatty liver disease. Clin Exp Hepatol. 2018; 4(3):165–74. https://doi.org/10.5114/ceh.2018.78120 PMID: 30324141
5. Fazel Y, Koenig AB, Sayiner M, Goodman ZD, Younossi ZM. Epidemiology and natural history of non-alcoholic fatty liver disease. Metabolism. 2016; 65(8):1017–25. https://doi.org/10.1016/j.metabol.2016.01.012 PMID: 26997539
6. Targher G, Byrne CD, Lonardo A, Zoppini G, Barbui C. Non-alcoholic fatty liver disease and risk of incident cardiovascular disease: a meta-analysis. J Hepatol. 2016; 65(3):589–600. https://doi.org/10.1016/j.jhep.2016.05.013 PMID: 27212244
7. Mahfood Haddad T, Hamdeh S, Kanmanthareddy A, Alla VM. Nonalcoholic fatty liver disease and the risk of clinical cardiovascular events: a systematic review and meta-analysis. Diabetes Metab Syndr. 2017; 11(Suppl 1):S209–16.

8. Bellentani S. The epidemiology of non-alcoholic fatty liver disease. Liver Int. 2017; 37(Suppl 1):81–4.

9. Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease—meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology. 2016; 64(1):73–84. https://doi.org/10.1002/hep.28431

10. Younossi ZM. Non-alcoholic fatty liver disease—a global public health perspective. J Hepatol. 2019; 70 (3):531–44. https://doi.org/10.1016/j.jhep.2018.10.033

11. Younossi ZM, Tampi R, Priyadarshini M, Nader F, Younossi IM, Racila A. Burden of illness and economic model for patients with nonalcoholic steatohepatitis in the United States. Hepatology. 2019; 69 (2):564–72. https://doi.org/10.1002/hep.30254 PMID: 30180285

12. Castera L, Friedrich-Rust M, Loomba R. Noninvasive assessment of liver disease in patients with nonalcoholic fatty liver disease. Gastroenterology. 2019; 156(5):1264–81.e4. https://doi.org/10.1053/j.gastro.2018.12.036 PMID: 30660725

13. Castera L. Diagnosis of non-alcoholic fatty liver disease/non-alcoholic steatohepatitis: non-invasive tests are enough. Liver Int. 2018; 38(Suppl 1):67–70.

14. Kotronen A, Peltonen M, Hakkarainen A, Sevastianova K, Bergholm R, Johansson LM, et al. Prediction of non-alcoholic fatty liver disease and liver fat using metabolic and genetic factors. Gastroenterology. 2009; 137(3):865–72. https://doi.org/10.1053/j.gastro.2009.06.005 PMID: 19524579

15. Koivula RW, Heggie A, Barnett A, Cederberg H, Hansen TH, Koopman AD, et al. Discovery of biomarkers for glycaemic deterioration before and after the onset of type 2 diabetes: rationale and design of the epidemiological studies within the IMI DIRECT Consortium. Diabetologia. 2014; 57(6):1132–42. https://doi.org/10.1007/s00125-014-3216-x PMID: 24695864

16. Koivula RW, Forgie IM, Kurbasic A, Vinuela A, Heggie A, Giordano GN, et al. Discovery of biomarkers for glycaemic deterioration before and after the onset of type 2 diabetes: descriptive characteristics of the epidemiological studies within the IMI DIRECT Consortium. Diabetologia. 2019; 62(9):1601–15. https://doi.org/10.1007/s00125-019-4906-1 PMID: 31203377

17. Thomas EL, Fitzpatrick JA, Malik SJ, Taylor-Robinson SD, Bell JD. Whole body fat: content and distribution. Prog Nucl Magn Reson Spectrosc. 2013; 73:56–80. https://doi.org/10.1016/j.pnmrs.2013.04.001 PMID: 23962884

18. Wilman HR, Kelly M, Garratt S, Matthews PM, Milanesi M, Herlihy A, et al. Characterisation of liver fat in the UK Biobank cohort. PLoS ONE. 2017; 12(2):e0172921. https://doi.org/10.1371/journal.pone.0172921 PMID: 28241076

19. Assarsson E, Lundberg M, Holmquist G, Bjorkesten J, Thorsen SB, Ekman D, et al. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. PLoS ONE. 2014; 9(4):e95192. https://doi.org/10.1371/journal.pone.0095192 PMID: 24755770

20. Aldo P, Marusov G, Svancara D, David J, Mor G. Simple Plex(TM): a novel multi-analyte, automated microfluidic immunoassay platform for the detection of human and mouse cytokines and chemokines. Am J Reprod Immunol. 2016; 75(6):678–93. https://doi.org/10.1111/aji.12512

21. Drobin K, Nilsson P, Schwenk JM. Highly multiplexed antibody suspension bead arrays for plasma protein profiling. Methods Mol Biol. 2013; 1023:137–45. https://doi.org/10.1007/978-1-4614-7209-4_8 PMID: 23765623

22. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat. 1992; 46(3):175–85.

23. Matsuda M, DeFronzo RA. Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp. Diabetes Care. 1999; 22(9):1462–70. https://doi.org/10.2337/diacare.22.9.1462 PMID: 10480510

24. Mari A, Pacini G, Murphy E, Ludvik B, Nolan JJ. A model-based method for assessing insulin sensitivity from the oral glucose tolerance test. Diabetes Care. 2001; 24(3):539–48. https://doi.org/10.2337/diacare.24.3.539 PMID: 11289482

25. Stumvoll M, Mitrakou A, Pimenta W, Jenssen T, Yki-Järvinen H, Van Haeften T, et al. Use of the oral glucose tolerance test to assess insulin release and insulin sensitivity. Diabetes Care. 2000; 23(3):295–301. https://doi.org/10.2337/diacare.23.3.295 PMID: 10868854

26. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol. 1996; 58:267–88.

27. Setia MS. Methodology series module 5: sampling strategies. Indian J Dermatol. 2016; 61(5):505–9. https://doi.org/10.4103/0019-5154.190118 PMID: 27688438

28. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010; 33(1):1–22. PMID: 20808728

29. Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics. 2016; 32(9):1423–6. https://doi.org/10.1093/bioinformatics/btw079 PMID: 27153000

30. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics. 2007; 8:25. https://doi.org/10.1186/1471-2105-8-25 PMID: 17254353

31. Neumann U, Genze N, Heider D. EFS: an ensemble feature selection tool implemented as R-package and web-application. BioData Min. 2017; 10:21. https://doi.org/10.1186/s13040-017-0142-8 PMID: 28674556

32. R Core Team. R: a language and environment for statistical computing. Version 3.2.5. Vienna: R Foundation for Statistical Computing; 2013.

33. Kuhn M. caret: classification and regression training. Version 6.0–71. Comprehensive R Archive Network; 2016.

34. Bedogni G, Bellentani S, Miglioli L, Masutti F, Passalacqua M, Castiglione A, et al. The Fatty Liver Index: a simple and accurate predictor of hepatic steatosis in the general population. BMC Gastroenterol. 2006; 6:33. https://doi.org/10.1186/1471-230X-6-33 PMID: 17081293

35. Lee JH, Kim D, Kim HJ, Lee CH, Yang JI, Kim W, et al. Hepatic steatosis index: a simple screening tool reflecting nonalcoholic fatty liver disease. Dig Liver Dis. 2010; 42(7):503–8. https://doi.org/10.1016/j.dld.2009.08.002 PMID: 19766548

36. Fedchuk L, Nascimbeni F, Pais R, Charlotte F, Housset C, Ratziu V, et al. Performance and limitations of steatosis biomarkers in patients with nonalcoholic fatty liver disease. Aliment Pharmacol Ther. 2014; 40(10):1209–22. https://doi.org/10.1111/apt.12963 PMID: 25267215

37. Alberti KG, Zimmet P, Shaw J, Group IDF Epidemiology Task Force Consensus Group. The metabolic syndrome—a new worldwide definition. Lancet. 2005; 366(9491):1059–62. https://doi.org/10.1016/S0140-6736(05)67402-8

38. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015; 12(3):e1001779. https://doi.org/10.1371/journal.pmed.1001779 PMID: 25826379

39. Lean ME, Leslie WS, Barnes AC, Brosnahan N, Thom G, McCombie L, et al. Primary care-led weight management for remission of type 2 diabetes (DiRECT): an open-label, cluster-randomised trial. Lancet. 2018; 391(10120):541–51. https://doi.org/10.1016/S0140-6736(17)33102-1 PMID: 29221645

40. Araujo AR, Rosso N, Bedogni G, Tiribelli C, Bellentani S. Global epidemiology of non-alcoholic fatty liver disease/non-alcoholic steatohepatitis: what we need in the future. Liver Int. 2018; 38(Suppl 1):47–51.

41. Motamed N, Sohrabi M, Ajdarkosh H, Hemmasi G, Maadi M, Sayeedian FS, et al. Fatty liver index vs waist circumference for predicting non-alcoholic fatty liver disease. World J Gastroenterol. 2016; 22(10):3023–30. https://doi.org/10.3748/wjg.v22.i10.3023 PMID: 26973398

42. Poynard T, Ratziu V, Naveau S, Thabut D, Charlotte F, Messous D, et al. The diagnostic value of biomarkers (SteatoTest) for the prediction of liver steatosis. Comp Hepatol. 2005; 4:10. https://doi.org/10.1186/1476-5926-4-10 PMID: 16375767

43. Baranova A, Liotta L, Petricoin E, Younossi ZM. The role of genomics and proteomics: technologies in studying non-alcoholic fatty liver disease. Clin Liver Dis. 2007; 11(1):209–20. https://doi.org/10.1016/j.cld.2007.02.003 PMID: 17544980

44. Eslam M, Valenti L, Romeo S. Genetics and epigenetics of NAFLD and NASH: clinical impact. J Hepatol. 2018; 68(2):268–79. https://doi.org/10.1016/j.jhep.2017.09.003 PMID: 29122391

45. Wood GC, Chu X, Argyropoulos G, Benotti P, Rolston D, Mirshahi T, et al. A multi-component classifier for nonalcoholic fatty liver disease (NAFLD) based on genomic, proteomic, and phenomic data domains. Sci Rep. 2017; 7:43238. https://doi.org/10.1038/srep43238 PMID: 28266614

46. Eslam M, Hashem AM, Romero-Gomez M, Berg T, Dore GJ, Mangia A, et al. FibroGENE: a gene-based model for staging liver fibrosis. J Hepatol. 2016; 64(2):390–8. https://doi.org/10.1016/j.jhep.2015.11.008 PMID: 26592354

47. Alonso C, Fernandez-Ramos D, Varela-Rey M, Martinez-Arranz I, Navasa N, Van Liempd SM, et al. Metabolomic identification of subtypes of nonalcoholic steatohepatitis. Gastroenterology. 2017; 152(6):1449–61.e7. https://doi.org/10.1053/j.gastro.2017.01.015 PMID: 28132890

48. Perakakis N, Polyzos SA, Yazdani A, Sala-Vila A, Kountouras J, Anastasilakis AD, et al. Non-invasive diagnosis of non-alcoholic steatohepatitis and fibrosis with the use of omics and supervised learning: a proof of concept study. Metabolism. 2019; 101:154005. https://doi.org/10.1016/j.metabol.2019.154005 PMID: 31711876

49. Katsiki N, Gastaldelli A, Mikhailidis DP. Predictive models with the use of omics and supervised machine learning to diagnose non-alcoholic fatty liver disease: A "non-invasive alternative" to liver biopsy? Metabolism. 2019; 101:154010. https://doi.org/10.1016/j.metabol.2019.154010

50. Canbay A, Kalsch J, Neumann U, Rau M, Hohenester S, Baba HA, et al. Non-invasive assessment of NAFLD as systemic disease—a machine learning perspective. PLoS ONE. 2019; 14(3):e0214436. https://doi.org/10.1371/journal.pone.0214436

51. Danford CJ, Yao ZM, Jiang ZG. Non-alcoholic fatty liver disease: a narrative review of genetics. J Biomed Res. 2018; 32(5):389–400. https://doi.org/10.7555/JBR.32.20180045 PMID: 30355853

52. Al-Fayoumi S, Hashiguchi T, Shirakata Y, Mascarenhas J, Singer JW. Pilot study of the antifibrotic effects of the multikinase inhibitor pacritinib in a mouse model of liver fibrosis. J Exp Pharmacol. 2018; 10:9–17. https://doi.org/10.2147/JEP.S150729 PMID: 29785143

53. Hagstrom H, Stal P, Hultcrantz R, Brismar K, Ansurudeen I. IGFBP-1 and IGF-I as markers for advanced fibrosis in NAFLD—a pilot study. Scand J Gastroenterol. 2017; 52(12):1427–34. https://doi.org/10.1080/00365521.2017.1379556

54. Petaja EM, Zhou Y, Havana M, Hakkarainen A, Lundbom N, Ihalainen J, et al. Phosphorylated IGFBP-1 as a non-invasive predictor of liver fat in NAFLD. Sci Rep. 2016; 6:24740. https://doi.org/10.1038/srep24740 PMID: 27091074

55. Adamek A, Kasprzak A. Insulin-like growth factor (IGF) system in liver diseases. Int J Mol Sci. 2018; 19 (5):1308.

56. Chen Y, Huang H, Xu C, Yu C, Li Y. Long non-coding RNA profiling in a non-alcoholic fatty liver disease rodent model: new insight into pathogenesis. Int J Mol Sci. 2017; 18(1):21.

57. Gaggini M, Carli F, Rosso C, Buzzigoli E, Marietti M, Della Latta V, et al. Altered amino acid concentrations in NAFLD: impact of obesity and insulin resistance. Hepatology. 2018; 67(1):145–58. https://doi.org/10.1002/hep.29465 PMID: 28802074

58. Imhasly S, Naegeli H, Baumann S, von Bergen M, Luch A, Jungnickel H, et al. Metabolomic biomarkers correlating with hepatic lipidosis in dairy cows. BMC Vet Res. 2014; 10:122. https://doi.org/10.1186/1746-6148-10-122 PMID: 24888604

59. Koch M, Freitag-Wolf S, Schlesinger S, Borggrefe J, Hov JR, Jensen MK, et al. Serum metabolomic profiling highlights pathways associated with liver fat content in a general population sample. Eur J Clin Nutr. 2017; 71(8):995–1001. https://doi.org/10.1038/ejcn.2017.43 PMID: 28378853

60. Uhlen M, Karlsson MJ, Hober A, Svensson AS, Scheffel J, Kotol D, et al. The human secretome. Sci Signal. 2019; 12(609):eaaz0274. https://doi.org/10.1126/scisignal.aaz0274

61. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015; 347(6220):1260419. PMID: 25613900

Paper IV

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

About the Author

PASCAL M. MUTIE completed his medical training (MB.ChB) at the university of Nairobi, Kenya and completed his MPH at Lund university, Sweden. Pascal has completed his PhD at the Genetic and Molecular Epidemiology unit at the Lund University Diabetes Center. His thesis focused on causal effects of perturbations in energy homeostasis on cardiometabolic outcomes.

LUND UNIVERSITY

FACULTY OF MEDICINE

Department of Clinical Research

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2022:92
ISBN 978-91-8021-253-3
ISSN 1652-8220

9 789180 212533