Enhancing prediction and causal inference in metabolic dyshomeostasis

Atabaki Pasdar, Naeimeh

2020

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

Enhancing prediction and casual inference in metabolic dyshomeostasis

NAEIMEH ATABAKI-PASDAR
DEPARTMENT OF CLINICAL RESEARCH, MALMÖ | LUND UNIVERSITY 2020

Naeimeh Atabaki-Pasdar completed her BSc in Biomedical Engineering at Amirkabir University of Technology and her MSc in Bioinformatics at Lund University. Naeimeh has completed her PhD at the Genetic and Molecular Epidemiology unit at the Lund University Diabetes Centre. Naeimeh´s doctoral thesis focuses on enhancing prediction and causal inference in metabolic dyshomeostasis.

Printed by Media-Tryck, Lund 2020 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

FACULTY OF MEDICINE

Department of Clinical Research, Malmö

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2020:138
ISBN 978-91-8021-005-8
ISSN 1652-8220

Enhancing prediction and causal inference in metabolic dyshomeostasis

Enhancing prediction and causal inference in metabolic dyshomeostasis

by Naeimeh Atabaki-Pasdar

LUND
UNIVERSITY

DOCTORAL DISSERTATION
Thesis advisors: Paul W. Franks, Mattias Ohlsson
Faculty opponent: Stefan Engblom - Uppsala University

To be presented, with the permission of the Faculty of Medicine of Lund University, for public criticism at the Department of Clinical Research (CRC) room 28-11-026, on Monday, 7th December, 2020 at 13:00

Organization	Document name
LUND UNIVERSITY	**DOCTORAL DISSERTATION**
	Date of disputation: 2020-12-07
Department of Clinical Research	Sponsoring organization: IMI DIRECT
Author(s): Naeimeh Atabaki-Pasdar	

Title and subtitle: Enhancing prediction and causal inference in metabolic dyshomeostasis

Abstract

This thesis is focused on two globally prevalent diseases: i) non-alcoholic fatty liver disease (NAFLD) and ii) type 2 diabetes (T2D), with an overall aim of improving prediction and causal inference in the context of these conditions. Our projects were mainly conducted using IMI DIRECT and UK Biobank datasets including multi-omics data, extensive environmental exposures, and biological intermediates.

In paper I, we utilized structural equation modeling to test the 'twin-cycle' hypothesis concerning interactions between the liver and the pancreas in the etiology of T2D. Furthermore, the association of physical activity with glycemic control was investigated within the twin-cycle hypothesis. Our results showed the association of physical activity with several metabolic traits and factors. Moreover, the mediation effect of basal insulin secretion rate, insulin sensitivity and liver fat was identified from physical activity towards glucose regulation.

In paper II, we developed a series of machine learning-based models for the diagnosis of fatty liver, using different combinations of complex clinical and omics input data, to screen at-risk populations for NAFLD. Beta-cell function and insulin sensitivity appeared to be the most informative predictors in the developed diagnostic models. Furthermore, the derived importance lists of each data set (clinical, genetic, transcriptomic, proteomic, and metabolomic) were highlighting previous findings and suggesting potential molecular features of the NAFLD etiology.

In paper III, Bayesian network and Mendelian randomization approaches were deployed to examine a range of putative causal associations underlying the development of fatty liver. Our analyses identified basal insulin secretion rate and visceral fat as two key drivers. In addition, the sensitivity analysis on diabetes and non-diabetes strata identified a network mostly dominated by dysglycemia in presence of T2D, whereas, it was mainly controlled by excess adiposity in the absence of T2D.

In paper IV, genotype-based recall (GBR) clinical trials, in which the genetic burden of individuals is used in recruiting two groups of participants with a high and low genetic risk score, were simulated and compared with the conventional randomized controlled trials (RCTs) in terms of their statistical power and the required sample sizes. The analysis showed that GBR trials are, under several diverse scenarios, more powerful than conventional RCTs for testing gene-treatment interactions.

Key words: metabolic dyshomeostasis, fatty liver, NAFLD, type 2 diabetes, machine learning, causal inference

Classification system and/or index terms (if any)

Supplementary bibliographical information	Language
https://www.dropbox.com/sh/7ikqq8xxut5kezi/AAB_BleHrF2Qlq9cU62IyXGra?dl=0.$	English
ISSN and key title: 1652-8220, Lund University, Faculty of Medicine Doctoral Dissertation Series 2020:138	ISBN: 978-91-8021-005-8

Recipient's notes	Number of pages	Price
	181	
	Security classification	

Enhancing prediction and causal inference in metabolic dyshomeostasis

by Naeimeh Atabaki-Pasdar

LUND
UNIVERSITY

Faculty of Medicine
Department of Clinical Research

ISBN 978-91-8021-005-8
ISSN 1652-8220
Lund University, Faculty of Medicine Doctoral Dissertation Series 2020:138

Printed in Sweden by Media-Tryck, Lund University, Lund 2020

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

Cover illustration by Hamed Behjat, Amirhosein Dehqani and Naeimeh Atabaki-Pasdar

It always seems impossible until it's done.
— Nelson Mandela

Never Give Up
Keep Going &
Get Over It —

Dedicated to
My parents, Shahin & Reza

Contents

1 2 3 4

[1]Supporting Information for the papers included in this thesis can be found here:
`https://www.dropbox.com/sh/7ikqq8xxut5kezi/AAB_B1eHrF2Qlq9cU62IyXGra?dl=0.$`

[2]All papers and figures are reproduced with permission of their respective publishers.

[4]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.

List of publications

This thesis is based on the following publications, referred to by their Roman numerals:

I **The role of physical activity in metabolic homeostasis before and after the onset of type 2 diabetes: an IMI DIRECT study**
Koivula RW, **Atabaki-Pasdar N**, Giordano GN, White T, Adamski J, Bell JD, Beulens J, Brage S, Brunak S, De Masi F, Dermitzakis ET, Forgie IM, Frost G, Hansen T, Hansen TH, Hattersley A, Kokkola T, Kurbasic A, Laakso M, Mari A, McDonald TJ, Pedersen O, Rutters F, Schwenk JM, Teare HJA, Thomas EL, Vinuela A, Mahajan A, McCarthy MI, Ruetten H, Walker M, Pearson E, Pavo I, Franks PW; IMI DIRECT Consortium.
Diabetologia volume 63, pages744–756(2020)

II **Predicting and elucidating the etiology of fatty liver disease: A machine learning modeling and validation study in the IMI DIRECT cohorts**
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, McDonald TJ, Rutters F, Cederberg H, Chabanova E, Dale M, Masi F, Thomas CE, Allin KH, Hansen TH, Heggie A, Hong MG, Elders PJM, Kennedy G, Kokkola T, Pedersen HK, Mahajan A, McEvoy D, Pattou F, Raverdy V, Häussler RS, Sharma S, Thomsen HS, Vangipurapu J, Vestergaard H, 't Hart LM, Adamski J, Musholt PB, Brage S, Brunak S, Dermitzakis E, Frost G, Hansen T, Laakso M, Pedersen O, Ridderstråle M, Ruetten H, Hattersley AT, Walker M, Beulens JWJ, Mari A, Schwenk JM, Gupta R, McCarthy MI, Pearson ER, Bell JD, Pavo I, Franks PW.
PLoS Med. 17(6):e1003149(2020)

III **Inferring the causal pathways between metabolic processes and liver fat accumulation: an IMI DIRECT study**
Atabaki-Pasdar N *, Pomares-Millan H *, Koivula RW, Agudelo L, Thomas EL, Viñuela A, Giordano GN, Forgie IM, McDonald TJ, Ruetten H, Masi F, Hansen TH, Kokkola T, Mahajan A, Adamski J, Brage S, Brunak S, Dermitzakis E, Frost G, Hansen T, Laakso M, Pedersen O, Ridderstråle M, Hattersley AT, Walker M, Beulens JWJ, McCarthy MI, Schwenk JM, Mari A, Ruetten H, Bell JD, Ohlsson M, Pavo I, Pearson ER, Franks PW.
(manuscript)

IV **Statistical power considerations in genotype-based recall randomized controlled trials**
Atabaki-Pasdar N, Ohlsson M, Shungin D, Kurbasic A, Ingelsson E, Pearson ER, Ali A*, Franks PW*.
Sci Rep 6, 37307 (2016)

Publications not included in this thesis

I **Causal inference in obesity research**

Franks PW, **Atabaki-Pasdar N**
J Intern. Med, 2017

II. **Genome-Wide and Abdominal MRI Data Provide Evidence That a Genetically Determined Favorable Adiposity Phenotype Is Characterized by Lower Ectopic Liver Fat and Lower Risk of Type 2 Diabetes, Heart Disease, and Hypertension**
Ji Y*, Yiorkas AM*, Frau F*, Mook-Kanamori D*, Staiger H*, Thomas EL, **Atabaki-Pasdar N**, Campbell A, Tyrrell J, Jones SE, Beaumont RN, Wood AR, Tuke MA, Ruth KS, Mahajan A, Murray A, Freathy RM, Weedon MN, Hattersley AT, Hayward C, Machann J, Häring HU, Franks P, de Mutsert R, Pearson E, Stefan N*, Frayling TM*, Allebrandt KV*, Bell JD*, Blakemore AI*, Yaghootkar H*
Diabetes, 2019

III. **Genetic studies of MRI liver iron content in 9,800 individuals identify 3 variants in genes regulating hepcidin and yields new insights into its link with hepatic and extrahepatic diseases**
Wilman HR*, Parisinos CA*, **Atabaki-Pasdar N**, Kelly M, Thomas EL, Neubauer S, Mahajan A, Hingorani AD, Patel RS, Hemingway H, Franks PW, Bell JD, Banerjee R, Yaghootkar H
Journal of Hepatology, 2019

IV. **Glucose-dependent insulinotropic peptide and risk of cardiovascular events and mortality: a prospective study**
Jujić A, **Atabaki-Pasdar N**, Nilsson PM, Almgren P, Hakaste L, Tuomi T, Berglund LM, Franks PW, Holst JJ, Prasad RB, Torekov SS, Ravassa S, Díez J, Persson M, Melander O, Gomez MF, Groop L, Ahlqvist E*, Magnusson M*
Diabetalogia, 2020

V. **Association of Established Blood Pressure Loci With 10-Year Change in Blood Pressure and Their Ability to Predict Incident Hypertension**
Poveda A, **Atabaki-Pasdar N**, Ahmad S, Hallmans G, Renström F, Franks PW
JAHA 2020

VI. **An investigation of causal relationships between prediabetes and vascular complications**
Mutie P*, Pomares-Milan H*, **Atabaki-Pasdar N**, Jordan N, Adams R, Tajes J,Giordano GN, Franks PW
Nature Communications, 2020

Acknowledgements

Paul, there are no proper words to convey my deep gratitude and respect for you. Thanks for your endless support, motivation, kindness, trust and immense knowledge. Thanks for guiding and encouraging me to be professional and do the right thing even when the road got rough. It is impossible to imagine having have had a better advisor and mentor for my PhD studies, and even before, for my Master's thesis. Thanks for all these wonderful and rewarding years and for being a great supervisor and friend.

I would also like to express my deepest appreciation to **Mattias Ohlsson**, my amazing co-supervisor, for sharing his expertise and limitless support. Thanks **Mattias** for your patience in giving guidance since the Perl course and during my Master's and PhD studies and also for kindly hosting me at the **Physics Department** on Fridays.

There is no way to express how much it meant to me to have been a member of the **Genetic and Molecular Epidemiology (GAME) unit**. I want to heartily thank my brilliant friends, colleagues and roommates at **CRC 60-12-012** for their kind support and stimulating discussions: **Hugo P & Seb K (my Kontrast buddies), Hugo F, Pascal, Daniel, Neli, Mi** and also to the most humble senior colleagues and friends: **Paul, Pernilla, Alaitz, Nick, Sebastian, Juan and Simon** for lending me generously their expertise and intuition to my scientific and technical problems. You all made work a joyful, exciting and inspiring place for me. I would also like to extend my gratitude and thankfulness to the former members of the team who helped me get started with my PhD, and who remained great friends during these years: **Frida, Angela, Azra, Robert** and **Tibor**.

It was a great delight and rewarding experience to have been a member of the **IMI DIRECT** consortium during my PhD years. I am thankful to the whole consortium, and in particular, I wish to express my gratitude for the advice and support from **Ewan Pearson**, **Imre Pavo**, **Jimmy Bell**, **Robert Koivula**, **Ana Viñuela**, **Adem Dawed**, **Juan Fernandez**, **Ian Forgie**, **Francesca Frau**, **Giuseppe Giordano**, **Ramneek Gupta**, **Angus Jones**, **Azra Kurbasic**, **Anubha Mahajan**, **Andrea Mari**, **Mark McCarthy**, **Hartmut Ruetten**, **Jochen Schwen**, **Louise Thomas** and **Konstantinos Tsirigos**.

My sincere thanks also goes to **Manolis Kellis**, who provided me an opportunity to join his fantastic lab at **MIT** for an inspiring research visit. By spending time at your lab during the last year of my PhD, I truly learned a lot and also met excellent researchers. I would particularly like to single out **Leandro Agudelo** and thank him for his patient support and for all of the opportunities I was given to further my research. I would also like to deeply thank **Patty Purcell** for her generous and precious support throughout this visit.

I would also like to thank my great co-authors whom I have enjoyed working with and have learned from: **Hugo Pomares**, **Robert Koivula**, **Alaitz Poveda**, **Pascal Mutie**, **Ana Viñuela**,

Francesca Frau, **Hanieh Yaghootkar**, **Ashfaq Ali**, **Amra Jujic** and **Martin Magnusson**.

I take this opportunity to express gratitude to all of the Department faculty members of the **LUDC** for their help and support during my PhD. In particular, I wish to express my sincere thanks to **Mattias Borell, Lena Eliasson, Hindrik Mulder** and **Claes Moreau**.

I cannot forget friends in Sweden, **Niloofar**, **Milad**, **Saeedeh**, **Roozbeh**, **Newsha** and **Iman** who provided happy distractions to rest my mind outside of research, cheered me on, and celebrated each accomplishment.

I would like to deeply thank my family: my parents, **Shahin** and **Reza**, and my brothers, **Naeim, Mojtaba and Mohsen**, for supporting me spiritually throughout my PhD studies, and my life in general. Thanks for your unconditional trust and unfailing emotional support.

Finally, I thank with love my adorable companions, my husband **Hamid** and my daughter **Barana**. I could not have completed this dissertation without your support, understanding and caring. Thanks for brightening my life. **Hamid**, I can see your key role in all the achievements I have had since we met. Thanks for continuously encouraging me to be better and to do better!

Naeimeh Atabaki-Pasdar
Boston, Oct 2020

Popular summary

The food we eat and the beverages we drink provide our body with the energy to function properly. Metabolism is a network of chemical reactions that occur in the body cells to transform the molecular substrates into a cell-usable form of energy; either amino acids (protein), fats, or carbohydrates are the nutrients that provide the caloric input. Glucose (sugar) is the most basic form of carbohydrates and an important source of energy which is mainly regulated by insulin and glucagon, both hormones secreted from the pancreas. Insulin acts like a 'key' and activates the cell receptors to uptake glucose from the bloodstream for storage or metabolism when the blood glucose level is elevated, conversely, glucagon regulates the release of glucose when the blood glucose level is lower than it is needed. Glucose regulation and metabolic homeostasis are essential for survival. Severe health consequences or death occur when blood glucose levels reach abnormally high (hyperglycemia) or low (hypoglycemia) levels. Type 2 diabetes (**T2D**) and non-alcoholic fatty liver disease (**NAFLD**), both manifestations of metabolic abnormalities are the two diseases that I focused on my PhD thesis. T2D and NAFLD are globally prevalent diseases that cause serious health complications for the affected individual and impose a substantial economic and clinical burden for many societies around the world. In T2D, blood glucose concentration is chronically elevated ($\geq$7 mmol/L when fasting) and this is primarily due to insulin resistance that the cells no longer respond to secreted insulin. It is not exactly known yet why cells become resistant to insulin, there are many genetic and environmental factors implicated in the onset of the disease. One key risk factor is obesity, which effectively keeps the body in a perpetual feeding state. This may potentially result in a constant exposure of cells to insulin, with a simultaneous response aiming to stabilize the elevated glucose. Over time, cell receptors become desensitized to insulin, i.e., resistant to its effect. NAFLD, as the other focused disease is defined when the liver fat content is $\geq$5% of total liver weight and is *not* caused by excessive alcohol consumption. NAFLD is a spectrum of liver diseases and in its early stages, the disease is relatively benign, and most individuals express no overt symptoms. yet, in some cases, it can progress to a more serious condition where the patient may need a liver transplant. One of the most strongly correlated risk factors for NAFLD is insulin resistance or T2D. Indeed, these conditions often coincide and are strongly related, as described in a myriad of published studies. However, the mechanisms underlying the interplay between these diseases is poorly understood. Thus, understanding the etiology of these diseases might aid their **diagnosis**, **prevention**, and **treatment**. Bearing this in mind, we set about investigating fatty liver and glucose dysregulation within 4 papers included in my PhD thesis.

In paper II, we developed a series of machine learning-based classification models for the **diagnosis** of fatty liver. The diagnostic models could be potentially implemented for screening at-risk populations to detect NAFLD. When an association between two variables is observed (e.g., between a treatment and a disease), the question that comes after is whether

this observed association reflects a causal relationship or not (correlation does not imply causation). While it is often not necessary to understand the causal basis of associations that are useful for prediction and diagnosis, intervening to effectively **prevent** the diseases requires an understanding of causal pathways and mechanisms of action. To address this, in paper III, we used causal inference methods to investigate the causal pathways that underly the development of fatty liver. Within the same context, in paper I, we tested a well-established model on the etiology of T2D, concerning interactions between the liver, pancreas, and adipose tissue (known as the twin-cycle hypothesis). Furthermore, we investigated the effect of physical activity on glucose regulation through causal pathways within the twin-cycle model. To assess the **treatment** efficacy, experimental studies such as randomized controlled trials (RCTs) come to aid. RCTs are study designs where participants are randomly assigned to control and experimental groups to test a therapeutic agent such as a drug, however, RCTs can be extremely expensive and time-consuming, therefore optimizing the design of clinical trials to reduce costs and time to completion is of great importance. Paper IV, describes an innovative clinical trial method, termed 'genotype-based recall' (GBR), where the participants are recruited based on their genetic characteristics. Our findings showed that GBR trials are, under many scenarios, more powerful than conventional RCTs and require a much less number of participants when testing hypotheses.

Much of the work of this thesis has been mainly conducted utilizing data from IMI DIRECT[5] and UK Biobank[6] cohorts. The Innovative Medicines Initiative (IMI) Diabetes Research on Patient Stratification (DIRECT) is a multi-center prospective cohort study of ∼3000 adults from northern Europe. The study had two cohorts: the first cohort with participants free of T2D (n=2127), many of whom were at high risk of the disease, whereas the second cohort enrolled recently diagnosed patients (>6 months and < 3 years) with T2D (n=789). The participants were deeply examined and we had access to a huge amount of data including blood biomarkers, genomics, MRI, and multilevel lifestyle datasets. Besides, UK Biobank is a large long-term biobank study of over 500,000 participants who were recruited between 2006 and 2010 in the United Kingdom with an overall aim to investigate the determinants of a wide range of life-threatening diseases. Having access to these rich, diverse, and big datasets, the fundamental challenge was integrating them to understand T2D and NAFLD, which we aimed to address through novel machine learning, statistical, and bioinformatics methods.

[5]https://www.imi.europa.eu/projects-results/project-factsheets/direct
[6]https://www.ukbiobank.ac.uk

Abbreviations

ADA	American diabetes association
AIC	Akaike information criterion
ALT	alanine transaminase
ASAT	abdominal subcutaneous adipose tissue
AST	aspartate transaminase
AUC	Area Under Curve
BasalISR	insulin secretion at the beginning of the OGTT/MMTT
BilrubinDir	Direct bilirubin
BIC	Bayesian information criterion
BMI	body mass index
BN	Bayesian network
CFI	comparative fit index
Chol	Cholesterol
Clins	mean insulin clearance during the OGTT/MMTT
CT	Computerized tomography
DAG	directed acyclic graph
DBP	mean diastolic blood pressure
DIAGRAM	DIAbetes Genetics Replication And Meta-analysis
DIRECT	Diabetes Research on Patient Stratification
DPP	Diabetes Prevention Program
EFS	ensemble feature selection
ENMO	euclidean norm minus one
ER_RF	error-rate based random forest
FDA	Food and Drug Administration
FFA	free fatty acid
FLI	fatty liver index
FN	false negative
FP	false positive
fsOGTT	frequently sampled OGTT
GBR	genotype- based recall
GGTP	gamma-glutamyl transpeptidase
Gini_RF	Gini-index based random forest
Glucagonmino	fasting glucagon concentration
GlucoseSens	glucose sensitivity
GRS	genetic risk score
GWAS	Genome-wide association studies
H_0	null hypothesis
HbA1c	hemoglobin A1C
HDL	high-density lipoprotein cholesterol

hpfVM	High-pass-filtered vector magnitude
HSI	hepatic steatosis index
HSL	hormone-sensitive lipase
IAAT	intra-abdominal adipose tissue
IGR	impaired glucose regulation
IMI	Innovative Medicines Initiative
IV	instrumental variable
IVW	inverse variance-weighted
LASSO	least absolute shrinkage and selection operator
LDL	low-density lipoprotein cholesterol
LiverInflam	liver inflammation factor
LogReg	beta-values of logistic regression
MAFLD	metabolic associated fatty liver disease
MMTT	mixed meal tolerance test
MR	Mendelian randomization
MRI	magnetic resonance imaging
NAFLD	non-alcoholic fatty liver disease
NAFLD-LFS	non-alcoholic fatty liver disease liver fat score
NASH	non-alcoholic steatohepatitis
NGR	normal glucose regulation
OGIS	oral glucose insulin sensitivity
OGTT	oral glucose tolerance test
PA	physical activity
PancFat	pancreas fat
PancIron	pancreas iron
PCA	principal component analyses
P_cor	Pearson's product moment correlation
RCT	randomized controlled trial
RMSEA	root mean square error of approximation
ROC	receiver operating characteristic
ROCAUC	receiver operating characteristic area under the curve
RSS	residual sum of squares
SAT	subcutaneous adipose tissue
SBP	mean systolic blood pressure
SD	standard deviation
SE	standard error
SEM	structural equation modeling
SNP	single nucleotide polymorphism
S_cor	Spearman's rank correlation
T1D	type 1 diabetes
T2D	type 2 diabetes
TG	triglyceride

TLI	Tucker–Lewis index
TN	true negative
TotGLP1mino	concentration of fasting total GLP-1
TP	true positive
TC	twin-cycle
TC-PA	twin cycle-physical activity
TwoGlucose	2-hour glucose after OGTT/MMTT
TwoInsulin	2-hour insulin
UKBB	UK Biobank
VAT	visceral adipose tissue
WHO	World Health Organization
χ^2	Chi-squared
2SLS	2-stage least squares

Enhancing prediction and causal inference in metabolic dyshomeostasis

Chapter 1

Introduction

Metabolism can be defined as the internal network of chemical reactions that occur inside a cell, which renders the human body in a continuous state of energy flux. Through metabolism, the chemical energy that is stored in molecules is transformed into a form of energy that can be used for cellular processes. Energy derived from ingested food or beverages can be absorbed in forms of either amino acids (protein), fats, or carbohydrates. From a structural point of view, monosaccharides, of which glucose is one, are the most basic form of carbohydrate and an important source of energy, amounts of which in blood being mainly controlled by insulin and glucagon hormones secreted from the pancreas. Insulin promotes the transfer of glucose from the blood into the body's cells for storage or metabolism, whereas glucagon regulates the release of stored glucose (breakdown of the glycogen, known as glycogenolysis), or its *de novo* synthesis by the liver [1–3], a feedback loop involving the pancreas, liver and tissue cells (brain, muscle and adipose), see Figure 1.1. Glucose regulation and metabolic homeostasis *per se* are essential for survival, with severe health consequences or death occurring when blood glucose levels reach abnormally high or low levels [4–6]. Type 2 diabetes (T2D) and non-alcoholic fatty liver disease (NAFLD), both manifestations of metabolic abnormalities [5, 7–10], are the two diseases that I focus on in this thesis. T2D and NAFLD are globally prevalent diseases that often coincide [5, 7, 9, 11–14] and are thought to result from the complex interplay of genetic and environmental factors [15–18]. Their increasing trend has led to a rise in health care costs, as both diseases can progress to serious complications that are extremely costly to treat [7, 9, 14, 19]. In T2D, complications include serious damage to different body parts such as nerves, eyes, heart, brain, blood vessels and kidneys; in NAFLD, the complications include fibrosis, cirrhosis, ascites, esophageal varices, hepatic encephalopathy and liver cancer. Understanding the etiology of these highly prevalent and linked diseases may help in their diagnosis, prevention and treatment. Despite tremendous progress in understanding the risk factors and mechanisms underlying T2D and NAFLD, much remains unknown about their etiology

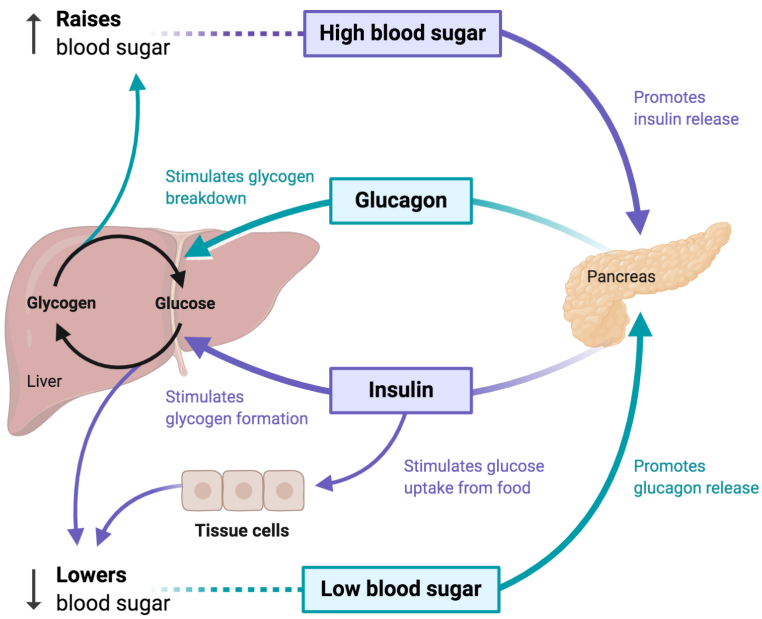

Figure 1.1: The regulation of blood glucose, a feedback loop involving the pancreas, liver and tissue cells. Created with BioRender.com

and pathogenesis.

In the following section, the key traits related to T2D and NAFLD are defined. The second section provides an overview of the mechanisms behind these diseases through a brief literature review. The third section continues the literature review on different approaches and methods used to study these diseases. This first chapter concludes by listing the aims of the papers included in this thesis and how we attempted to enhance the prediction and etiological elucidation of these metabolic diseases through statistical and bioinformatics methods.

1 Diabetes and fatty liver - definition of traits

1.1 Diabetes mellitus

Diabetes mellitus is a name given to a group of disorders in which blood glucose concentration is chronically elevated; causes may include defects in insulin secretion from the pan-

creas, its function, or both. There are two relatively common types of diabetes: type 1 diabetes (T1D) accounting for 10% of diabetes, and T2D, accounting for 90% of all diabetes. T1D is believed to be an autoimmune condition whereby the body attacks and destroys the beta cells of the pancreas, which results in the lack of insulin. T2D, non-autoimmune diabetes, is primarily due to insulin resistance, i.e. when the cells no longer respond to secreted insulin. However, as the disease progresses, the pancreas can also become affected, resulting in decreased insulin production. Various other rare types of diabetes also exist (`https://www.diabetes.co.uk/diabetes-types.html`). The work described in this thesis relates only toT2D. Nevertheless, regardless of the specific type of diabetes, the disease can be viewed as deficiencies of insulin production, insulin action, or both.

According to the American Diabetes Association (ADA), T2D is diagnosed when fasting blood glucose levels are greater than 7 mmol/L. In this hyperglycemic state, the pancreas secretes insulin from beta cells located in the islets of Langerhans. Insulin acts like a key and activates the cell receptors to uptake glucose from the bloodstream and store it, primarily in liver and skeletal muscle cells [20]. Glucose is stored in three ways: glycolysis, where glucose is converted to ATP (the basic unit of energy currency), glycogenesis, which is the formation of glycogen in the liver or muscle tissue, and lipogenesis, which stores energy from glucose or other substrates in the form of lipids or fatty acids in adipose tissue. Each of these processes reduces the concentration of glucose in the blood, thereby helping with glucose regulation (Figure 1.1) [1–3, 20].

There are many genetic [21, 22] and environmental factors that in some cases jointly and others separately causing why cells become resistant to insulin [15, 16]. One key risk factor is obesity, which effectively keeps the body in a perpetual feeding state. This may potentially result in constant exposure of cells to insulin, in an effort to bring down the increased glucose. Over time, cell receptors become desensitized to insulin, i.e. they become resistant to its effect. Furthermore, as cells become saturated with triglycerides (TGs), receptors become more distant from the cell membrane, which inhibits insulin signaling. As a compensatory answer to hepatic and muscle insulin resistance, the pancreas will produce more insulin, a state known as hyperinsulinemia [23]. However, as time progresses the demand for progressively more insulin can, in some cases, lead to exhaustion of pancreatic beta cells.

Family history of diabetes, some ethnicities (e.g. Hispanics, African American and Indian American according to ADA), aging and high birth-weight are among the non-modifiable risk factors for T2D which cannot be changed. Whereas, obesity, physical inactivity and abnormal lipid levels (e.g. low levels of high-density lipoproteins (HDL) and high levels of TGs are among those modifiable risk factors that can be changed to minimize the disease risk, through physical activity and a healthy diet (low-fat, high-fiber and complex carbohydrates) [24]. Physical inactivity, known as one of the key causes of global epidemics of T2D and obesity, is also strongly associated with peripheral insulin resistance [25–28].

Diabetes in and of itself may not be life-threating. However, poorly controlled diabetes, where blood glucose levels are chronically elevated, can cause damage to the small and large vessels in the body, leading to so called 'diabetes complications'. This includes damage to the heart, eyes, kidneys, brain, and peripheral tissues (which can cause gum disease, skin ulceration and ultimately loss of lower limbs). T2D is a chronic disorder that often requires lifelong treatment. As such, early diagnosis and treatment is of utmost importance to prevent serious associated complications [6, 10]. Screening people who are at risk of the disease and monitoring their blood sugar can help with early detection and deploy interventions to prevent the development of the symptoms [29].

1.2 The liver - its function and disfunction

The liver is the body's largest internal organ, weighing around 1.5 *kg*, and is located on the right-hand side of the abdomen. The liver is a vital organ and carries out many distinct major roles in the body including detoxification, making proteins and blood clotting factors, bile production, metabolism, process and storage of the nutrients. The liver has two separate blood supplies; the hepatic artery, which supplies it with oxygen-rich red blood cells, and the portal vein from the intestinal tract, which supplies it with nutrient-rich blood [30]. Ingested food is absorbed from the intestinal tract to be either metabolized, stored or detoxified in the liver. Manufacturing fats, including TGs, cholesterol (Chol) and lipoproteins, is another major function of the liver [1, 2, 30].

A healthy liver cell, known as a *hepatocyte*, has a centrally located nucleus and stores little droplets of lipids in liposomes. In the state of fatty liver, instead of microvesicles of fat, the liposomes become macrovesicles owing to fat accumulation, pushing the hepatocyte's nucleus away from the cell's center, and inhibiting its proper functioning [30]. Fatty liver is defined when the liver fat content is greater or equal to 5% of total liver weight. In the early stages of fatty liver disease, the disease is relatively benign and many people express no overt symptoms; but, in some cases, over time, it can progress to a more serious condition [7, 19].

It is not known why exactly fatty liver develops, but it occurs when the body produces excessive fat mass or when the liver cannot process it properly. One of the most strongly correlated risk factors is insulin resistance. This can be explained by the fact that in the presence of insulin resistance, the activity of hormone-sensitive lipase (HSL) is not suppressed and the adipocyte keeps hydrolyzing TGs into free fatty acids (FFAs) and releasing them into the bloodstream towards the liver. In the liver, these FFAs get stored as TGs and this can lead to hepatic fat accumulation [7, 12, 13].

Excessive alcohol consumption is another risk factor that stimulates fat storage in the liver, a condition known as alcoholic fatty liver disease. If the accumulation of fat in the liver is

not caused by alcohol, the condition is called NAFLD, which is a spectrum of liver diseases, ranging from simple steatosis to non-alcoholic steatohepatitis (NASH), fibrosis, cirrhosis, and hepatocellular carcinoma [31]. NAFLD is known as the most common liver disorder in western countries, affecting a quarter of the global population and occurring frequently alongside obesity and T2D [7, 32]. The condition can be explained by the accumulated TGs in the hepatocytes, which can reduce the sensitivity of the hepatocytes to insulin, increase hepatic gluconeogenesis and may lead to T2D or accelerate diabetes progression in those with the disease.

There is no Food and Drug Administration (FDA) approved medication for NAFLD, but weight loss through a healthy diet, exercise and surgery is in the first line to prevent, manage and treat the disease condition [32, 33]. Liver failure is a life-threatening condition and reaching the advanced stages of the disease limits treatment options significantly. Therefore, preventing or detecting liver disease very early in its pathogenesis is important if the adverse consequences are to be prevented and the disease process reversed.

2 Mechanisms behind NAFLD and T2D

Numerous studies have shown the high correlation between NAFLD, T2D and obesity, all proposing various mechanisms [12, 34–41]. A well-established hypothesis termed the *twin-cycle* model for the etiology of T2D suggests two processes that work in concert to affect glycemic control [38]. The first cycle describes how fat in the liver leads to resistance to insulin suppression of gluconeogenesis and consequently increase the level of fasting glucose, fasting insulin and hepatic lipogenesis. The second cycle involves the pancreas, where increased fatty acids in its islets decrease insulin secretion in response to ingested food or beverages and raise postprandial glucose levels. Over time, and if having a long-term intake of more energy than is expended, the two cycles spin faster and reach the trigger level of rapid onset of clinical diabetes (Figure 1.2).

A recent study on the causal relationship of NAFLD, T2D and obesity suggested sub-phenotyping the diseases for better diagnosis, prevention and treatment [37]. Their results showed that genetically-driven NAFLD is caused by T2D (a late-onset type 1-like diabetes) and central obesity, whereas metabolic NAFLD is promoted by genetic T2D, overall and central obesity (Figure 1.3) [37]. Indeed, the terminology of NAFLD has become controversial recently, as it is referring to the absence of a condition, rather than reflecting the cause of the disease [19].Instead, *MAFLD*, has been proposed by experts as a more appropriate term, which stands for metabolic associated fatty liver disease [42].

Another comprehensive hypothesis, termed 'multiple hit', explains NAFLD pathogenesis in individuals with a particularly heavy genetic predisposition to NAFLD. Insulin resistance,

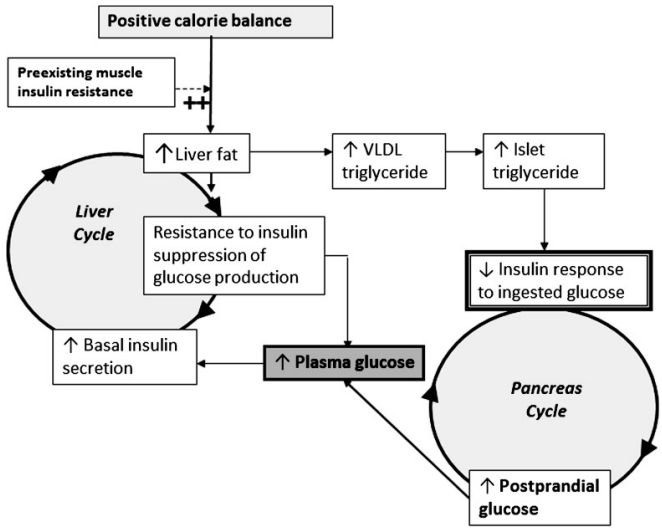

Figure 1.2: The twin-cycle hypothesis of the etiology of type 2 diabetes.
VLDL, very low-density lipoprotein. Figure from Taylor, Roy: Type 2 diabetes: etiology and reversibility [38].

adipose tissue hormones (Adipokines), gut microbiota, genetic and environmental factors are the key parameters considered in the multiple hit model (Figure 1.3) [32, 40]. The original idea was from the 'two hit' hypothesis [43] with the first hit as the promotion of insulin resistance by hepatic fat accumulation, and the second hit triggering inflammation and fibrosis. The multiple hit hypothesis was, then, proposed to be more complete by considering environmental factors that can affect gene expressions and weight gain. This leads to the flux of FFAs to the liver, pancreas and muscle, causing insulin resistance and hepatic *de novo* lipogenesis [32, 40]. Figure 1.4 shows how the pathophysiology of NAFLD is associated with obesity, T2D and metabolic syndrome.

2.1 Genes and the interaction with environmental factors

T2D and NAFLD are moderately heritable with defined genetic bases. In genetics, heritability can be defined as the proportion of phenotypic variation in a population which can be explained by the genetic variation and is not attributed to chance or environmental factors. Genome-wide association studies (GWAS) have been used extensively to define the specific genetic architecture of a given disease, typically using cohort and case-control studies, but occasionally also intervention trials, case series and other settings. In a GWAS, the single nucleotide polymorphisms (SNPs) are tested in association with a trait in order to find a consistent difference among the individuals' trait values and the SNPs [44–46]. Revealing the genetic architecture of these diseases can help with better understanding their biology and prioritizing targets for drug development. For T2D, large-scale GWAS in European-

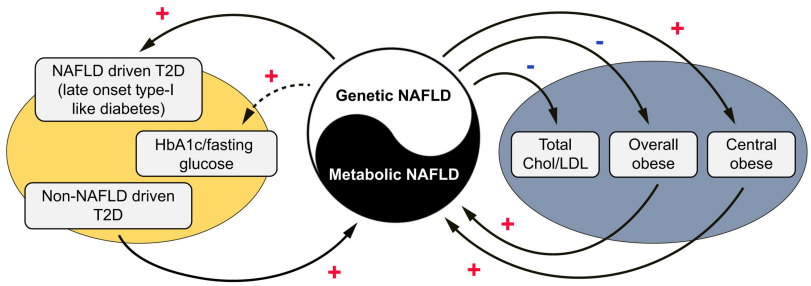

Figure 1.3: The causal relationship of NAFLD, T2D and obesity, suggesting sub-phenotyping the diseases for better diagnosis, prevention and treatment. Genetically-driven NAFLD is caused by T2D (a late-onset type 1-like diabetes) and central obesity, whereas metabolic NAFLD is promoted by genetic T2D, overall and central obesity
NAFLD, non-alcoholic fatty liver disease; T2D, type 2 diabetes; HbA1c, hemoglobin A1c; LDL, low-density lipoprotein; Chol, cholesterol. Figure from Liu et al.: Causal relationships between NAFLD, T2D and obesity have implications for disease subphenotyping [37].

ancestry cohorts have identified robust genetic associations linked with the disease and are gathered in a study called DIAGRAM (DIAbetes Genetics Replication And Meta-analysis) [44, 47]. As for NAFLD, no major consortium of GWAS studies has been formed to date. Nevertheless, variants in multiple loci (e.g. *PNPLA3*, *TM6SF2*, *MBOAT7*, *GCKR*, *HSD17B13*) have been replicated through GWAS [17, 18, 46]. All these genes are encoding proteins that regulate hepatic lipid metabolism [48]. *PNPLA3*, first reported in 2008 [49], is the most validated gene in association with fatty liver and half of those with NAFLD are carriers of at least one variant allele at *rs*738409 in *PNPLA3* [8, 50]. This variant leads to loss of enzymatic activity of hydrolyzed TGs and retinyl esters, as a consequence brings in two-fold greater accumulation of TGs and esters within the lipid droplets of hepatocytes [51]. The *rs*738409 (I148M) variant of the *PNPLA3* gene is prevalent among most populations and as such, is considered as a strong therapeutic target [52].

Neither genetic nor environmental factors are the sole determinants of these conditions; indeed, metabolic diseases are best-considered complex traits and are assumed to stem from environmental exposures acting on a susceptible polygenic background, a notion known as gene-environment interaction [15, 18, 53]. Multiplicative interactions by the epidemiological definition (also known as 'effect modification') occur when the combined genetic and environmental effects are significantly different from the effects of the genetic and environmental factors when considered additively. In fact, it is through the genome channel that non-genomic factors convey their effects on the disease. In order to optimize the treatment and prevention of these complex diseases, it follows that genetic factors should be considered in tandem with environmental factors in etiological studies [16].

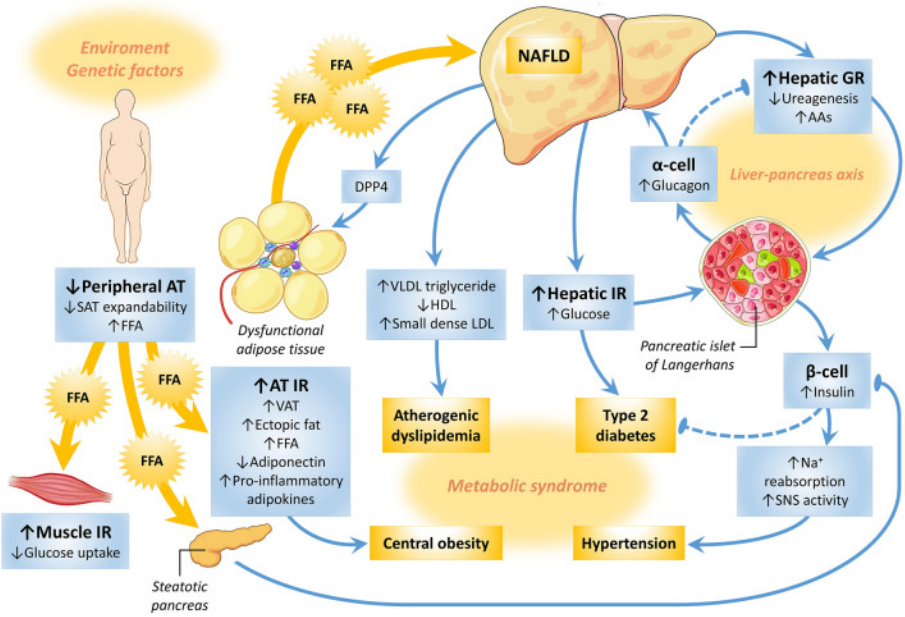

Figure 1.4: "Multiple hit" hypothesis for NAFLD development, showing how the pathophysiology of NAFLD is associated with obesity, T2D and metabolic syndrome.
AT, adipose tissue; SAT, subcutaneous adipose tissue; FFA, free fatty acid; IR, insulin resistance; VAT, visceral adipose tissue; NAFLD, non-alcoholic fatty liver disease; VLDL, very low-density lipoprotein; LDL, low-density lipoprotein; HDL, high-density lipoprotein; GR, glucagon resistance; DPP4, dipeptidyl peptidase 4; AAs, amino acids; SNS, sympathetic nervous system. Figure from Godoy-Matos et al.: NAFLD as a continuum: from obesity to metabolic syndrome and diabetes [32].

2.2 From a diagnostic point of view

Understanding the etiology and the mechanisms of the diseases is considered important also from a diagnostic point of view, as it can help with a more accurate and early diagnosis. The asymptomatic nature of NAFLD in its early stages makes diagnosis difficult. Liver blood tests, considered as first-line diagnostic tools, have a low sensitivity to detect the disease and most patients with NAFLD are left undiagnosed with a normal liver blood result [54]. Computerized tomography (CT) and magnetic resonance imaging (MRI) are good at detecting steatosis especially in cases with liver fat greater than 30%. However, they are not as reliable to detect NASH or fibrosis in the presence of scar tissue and inflammation [55]. The invasive nature of liver biopsy makes it a controversial option for investigating suspected NAFLD patients and is often avoided. Liver biopsy does not seem necessary when the blood test results are abnormal, other important risk factors such as T2D, obesity and dislipidemia are present, and when ultrasound/MRI confirm steatosis [56]. However, for differentiating the stages of NAFLD, a liver biopsy might be needed as an additional diagnostic tool [55]. Several non-invasive prediction models have been defined mainly from clinical or laboratory-based variables to overcome the gap in the diagnosis [57], such as the

fatty liver index (FLI) [58], hepatic steatosis index (HSI) [59], and the NAFLD liver fat score (NAFLD-LFS) [60]. However, none of these have sufficient predictive accuracy to be considered as gold standard and there is a lack of clear guidelines for identifying high-risk patients [7, 14, 61].

3 Observational vs. experimental studies

There are thousands of published observational studies exploring populations and drawing hypotheses on patterns, causes and effects of metabolic diseases [7, 41, 62–64]. A strong association between NAFLD and obesity was observed in a study of obese patients [body mass index (BMI)$> 35 kg/m^2$] with a NAFLD prevalence of 91% in patients undergoing bariatric surgery [65]. Another observational study shows the impact of central obesity in the severity of NAFLD [66]. A study of 3000 Italian individuals with T2D reports a high prevalence of NAFLD (69.5%) detected via ultrasound [67]. Another example is a 4-year longitudinal study of 7,849 individuals without diabetes, that were categorized into four groups by the presence of impaired fasting glucose and NAFLD at baseline. Their results showed a higher incidence of diabetes in the NAFLD group, 9.9% compared with 3.7% in the non-NAFLD group. The higher risk for T2D due to NAFLD was reported as the hazard ratio of 1.33 (95% CI 1.07–1.66) but only present in the impaired fasting glucose group. Their findings suggest an independent effect of NAFLD on T2D in the presence of impaired insulin secretion [68].

To validate the findings of such epidemiological hypotheses, experimental studies such as randomized controlled trials (RCTs) come to aid. RCTs are study designs where participants are randomly assigned to control and experimental groups to test an intervention such as a drug or a treatment. The samples of these groups should be similar as much as possible to ensure that the only expected difference between them to be the studied outcome. Random allocation of the participants into the groups minimizes the risk of biases.

Between 1993 and 1998, Finish Diabetes Prevention Study (DPS) studied overweight, middle-aged individuals with impaired glucose tolerance (n=522) on a RCT, where participants were randomly assigned either to a control group with general lifestyle information or to an intensive lifestyle intervention group, to study the effect of lifestyle intervention on the incidence of T2D for a median of 4 years [69, 70]. Some participants who were still not affected by T2D were followed up for an extra 9 years; intervention group with n=200 and the control group with n=166 participants. Their results showed a greater risk reduction in the progression to T2D for the lifestyle intervention group in comparison with the control group (HR= 0.614, 95% CI 0.478, 0.789) [69].

The ultimate objective of these etiological studies, both in epidemiological and experi-

mental settings, is inferring cause and effect. Although randomized controlled trials are known as the gold standard for inferring causality, they are not always practical and they can also be very expensive and time-consuming. On the other hand, observational studies are prone to biases and confounding, which limit the extent to which causal relationships can be reliably inferred. Bias causes a systematic deviation from the truth by creating a spurious association between exposure and outcome. Moreover, confounding refers to a misleading association between exposure and outcome caused by a third factor (confounder) that is correlated with both exposure and the outcome [71]. Several methods for causal inference that can be applied to epidemiological datasets have been developed in responding to these challenges, including Mendelian randomization (MR) analyses, Bayesian network (BN) analyses, and structural equation modeling (SEM) [72].

In the MR method, genetic variants are used as instrumental variables that proxy an exposure in order to assess its relationship with an outcome. MR method is analogous to RCT by reason of random assortment of the parents' alleles to the offspring. Moreover, DNA sequence remains unchanged throughout the life course. These make the relationship between genotypes and traits resilient to most confounders and reverse causality [73].

In a recent study, the causal relationships between NAFLD, T2D and obesity was investigated through MR analyses [37] (see Figure 1.3 and Chapter 1 - Section 2 for summary of findings). In another recent work, wide-angled MR analyses were conducted on 97 risk factors of T2D and they found 14 risk factors and 15 protective factors for their causal associations with T2D. Their findings suggest considering multiple aspects on obesity, mental health, sleep quality, education level, birth weight and smoking in defining the prevention strategies for T2D [74].

BNs are probabilistic graphical models, illustrated by directed acyclic graphs (DAGs), where the nodes are the random variables and the directed arcs between the nodes presents the dependencies among them. Aiming to define the causal relationship between NAFLD and metabolic syndrome, a simplified BN was suggested on data from bidirectional cohorts (NAFLD $\leftrightarrow$ metabolic syndrome) of a Chinese population where participants were studied for the incidence of NAFLD and metabolic syndrome, between the years of 2005-2011 [75]. Their findings suggested a reciprocal causal association between NAFLD and metabolic syndrome, with a higher effect from metabolic syndrome to NAFLD [75].

SEM is a statistical method, combination of multiple linear regression and factor analysis, to study the structural relationship between variables (both measured and latent). In a study of 2,230 older adults ($\geq$50 years) in the US, SEM was used to test a hypothesis that was defined based on the previously reported risk factors in association with prediabetes [76]. The direct and indirect pathways were studied and their results suggested waist circumference with the strongest direct effect on prediabetes [76].

MR analyses have been used in many observational studies to test causal associations between

different metabolic traits [37, 74, 77–79]. On the contrary, BNs and SEMs despite their potential for defining causal pathways, have infrequently been utilized in the studies of metabolic traits [72].

4 Aims

The existing evidence supports the close association between NAFLD and T2D, proposing several mechanistic hypotheses with many different genetic and environmental components involved. Yet, none of them can completely explain these conditions and the interplay between them. Considering the increasing prevalence of these diseases, which leads to a substantial economic and clinical burden on societies, understanding their etiology may help in the diagnosis, prevention, and treatment of them.

The overall objective of this thesis is to enhance the prediction and etiological elucidation of fatty liver disease and/or diabetes through novel statistical and bioinformatics methods. In this context, machine learning approaches, causal inference methods in observational studies and statistical power calculations for a novel study design were of the specific focus of this thesis. Much of the work is conducted using the IMI DIRECT and UK Biobank datasets and is limited to European ancestry. The overarching aims of the papers included in this thesis are as follows:

- Paper I – in this paper, SEM is utilized to test the previously proposed 'twin-cycle' hypothesis, which seeks to explain the interactions of energy homeostasis, glucose regulation, insulin action, adipose accumulation and liver fat accumulation. Moreover, the association of physical activity with glycemic control is investigated within the twin-cycle hypothesis. The utilized data is from IMI DIRECT diabetic and non-diabetic cohorts.
- Paper II – this paper is focused on developing models for the prediction of fatty liver applying machine learning methods to complex clinical and omic datasets of the IMI DIRECT cohorts. UK Biobank dataset is considered for validation purposes.
- Paper III – in this paper, BN and MR approaches are deployed to examine a range of putative causal associations between metabolic features involved in liver fat accumulation. IMI DIRECT and UK Biobank are used as the primary datasets.
- Paper IV – this paper describes an innovative clinical trial method, termed 'genotype-based recall' (GBR) and illustrates scenarios under which this approach is especially powerful for the assessment of gene-treatment (drug and lifestyle) interactions. Simulation analyses are done with parameters taken from Diabetes Prevention Program (DPP) study.

Chapter 2

Cohorts of studies

1 IMI DIRECT

The Innovative Medicines Initiative (IMI) Diabetes Research on Patient Stratification (DIRECT) consortium is a joint collaboration among 21 European academic institutions and 4 pharmaceutical companies [80] (Figure 2.1). The primary objective of IMI DIRECT is to discover and validate biomarkers of glycemic deterioration before and after the onset of T2D. The project aims to address the gap in diabetes drug development using a stratified medicine approach to the treatment of the disease. Accordingly, the IMI DIRECT investigators established two multi-center prospective cohort studies comprised of around 3000 adults from northern Europe. The first of these studies enrolled people without diabetes, many of whom were at high risk of the disease, whereas the second enrolled patients recently (>6 months and < 3 years) diagnosed with T2D.

The participants were exquisitely phenotyped, covering a broad array of risk factors, intermediate phenotypes and metabolic-related outcomes; standardized protocols were used at each of the seven clinical study centers. Blood samples were assayed for genetics, metabolomics, transcriptomics, miRNAs, proteomics and other biochemical features, fecal microbiome assays were performed, physical activity and sleep were assessed using accelerometry, diet was assessed using self-report and metabolomics, and multi-organ MRI spectroscopy was performed. Blood samples collected during frequently sampled carbohydrate tolerance tests were assayed for glucose, insulin and c-peptide, and these data were subsequently mathematically modelled to quantify glucose and insulin dynamics. Being a new study, bio-sample collection, storage, and processing were highly standardized, with minimal time from sample collection to analysis.

In the non-diabetes cohort, 2127 individuals were recruited from a sampling frame of 24,682

Figure 2.1: The Innovative Medicines Initiative (IMI) Diabetes Research on Patient Stratification (DIRECT) consortium, a joint collaboration among 21 European academic institutions and 4 pharmaceutical companies.

adults who were enrolled in population-based cohorts across Europe. Participants were selected according to the DIRECT-DETECT risk algorithm, based on age, BMI, waist circumference, use of antihypertensive medication, smoking status and family history of T2D [80, 81]. Of these individuals, 67% were within the prediabetes range with impaired glucose regulation (IGR) and the rest were with normal glucose regulation (NGR). Using plasma samples, the inclusion/exclusion was based on the ADA's 2011 criteria, where impaired fasting glucose was defined as 5.6–6.9 mmol/l, impaired glucose tolerance was defined as 7.8–11.0 mmol/l at 2-hours after a 75 g oral glucose load and prediabetes indic-

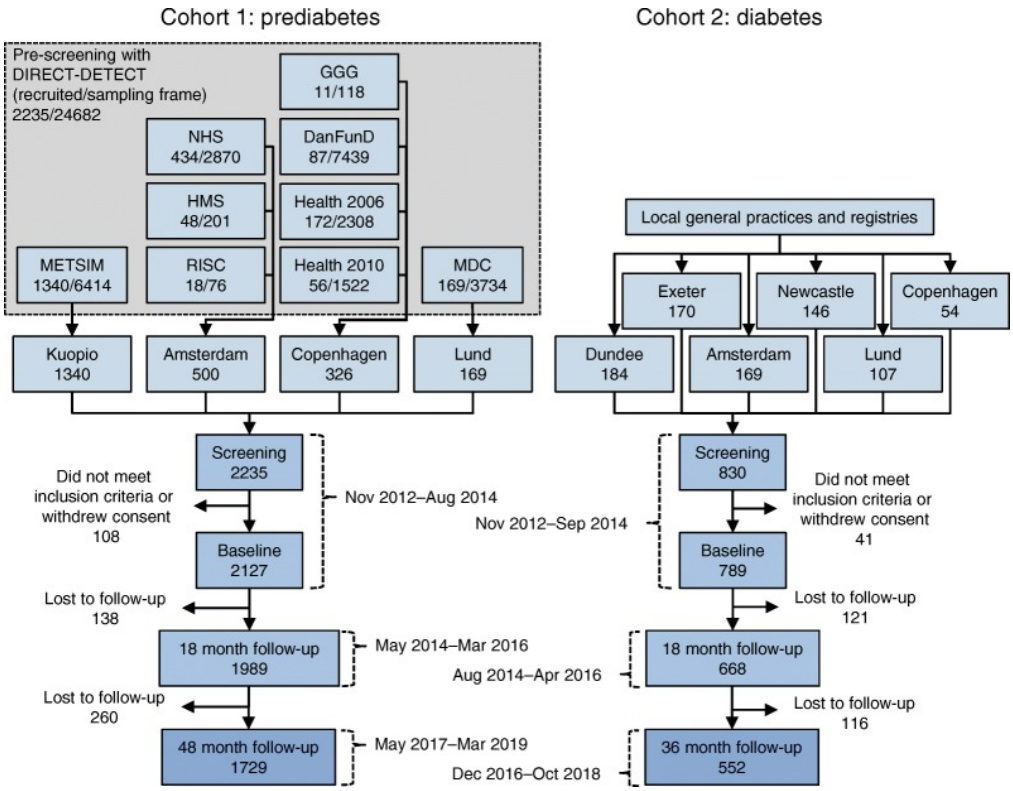

Figure 2.2: Participant flow of the diabetes and non-diabetes cohorts of IMI DIRECT.
DanFunD, Danish Functional Disability study; GGG, Gut, Grain and Greens study; HMS, Hoorn Meal Study; METSIM, Metabolic Syndrome in Men study; NHS, New Hoorn Study; RISC. Figure from koivula et al.: Discovery of biomarkers for glycaemic deterioration before and after the onset of type 2 diabetes: descriptive characteristics of the epidemiological studies within the IMI DIRECT Consortium [81].

ated by elevated hemoglobin A1c (HbA1c) was defined as 40–48 mmol/mol (5.7–6.4%) [82]. Participants were considered as having prediabetes if one or more of their aforementioned values fell within the impaired range. Participants whose values were below the prediabetes range were considered to have NGR, whereas values above the prediabetes range were considered to be indicative of diabetes. In the diabetes cohort, 789 participants, all of whom had a recent clinical diagnosis of T2D, were enrolled. After the baseline examination, the participants were followed-up at 18 months for both cohorts and again at 36 months for the diabetes cohort and at 48 months for the non-diabetes cohort (Figure 2.2).

All participants provided written informed consent at recruitment and the research conformed to the ethical principles of the Helsinki Declaration. With the exception of the UK centers, the study protocol was approved for each clinical center by its regional ethics authority; in the UK, where there were two clinical study centers resided at separate partner institutions, ethics approval was obtained from the national ethics review board [80, 81].

1.1 Biochemistry assays

Venous blood samples were drawn from participants after at least 10 hours of overnight fasting. In the non-diabetes cohort, 75 *g* frequently sampled OGTTs (fsOGTT) were conducted, whereas in the diabetes cohort, frequently sampled mixed meal tolerance tests (MMTTs), with 250 ml Fortisip liquid drink; in the latter, the MMTT, rather than the fsOGTT, was undertaken to minimize the risk of acute hyperglycemia. Venous blood was drawn at fasting and the carbohydrate drink (glucose or Fortisip) was then administered, following which blood was drawn at 5, 30, 45, 60, 90, and 120 minutes, in the non-diabetes cohort, and at 30, 60, 90, and 120 minutes in the diabetes cohort. Glucose, insulin and C-peptide concentrations were assessed at the aforementioned time-points and from these values, measures of glucose and insulin dynamics were derived using mathematical models [83]. These parameters include oral glucose insulin sensitivity (OGIS) according to the method of Mari et al., basal insulin secretion rate, glucose sensitivity (dose-response relationship between glucose and insulin secretion), and mean insulin clearance calculated as (mean insulin secretion)/(mean insulin concentration). Furthermore, HbA1c, glucagon-like peptide-1 (GLP-1), blood lipids (TGs, LDL, HDL, Chol), liver enzymes [alanine transaminase (ALT), aspartate transaminase (AST), gamma-glutamyl transpeptidase (GGTP)] and several other biochemistry assays were assessed through the drawn blood samples.

1.2 Lifestyle

Physical activity was assessed by triaxial accelerometry that participants wore on their non-dominant wrist for 10 days (ActiGraph GT3X+, `https://actigraphcorp.com/`). High-pass-filtered vector magnitude (hpfVM) and euclidean norm minus one (ENMO) of the triaxial acceleration signals were calculated as the intensity of the participants' movements in different directions and time points. The mean value was considered as the overall physical activity measure. The physical activity parameters were calculated from the raw data using the PAMPRO software (version 2015, `https://github.com/Thomite/pampro`) [84].

Diet was assessed through a 24-hour multi-pass dietary record and food habit questionnaire, a method validated as part of the Euroaction study [85]. The three passes (levels of dietary questioning) of this method are: 1st: documenting a usual day's meal, 2nd: participants' reflections on the first pass and add to that if needed, and 3rd: obtaining information on portion size. To assess the overall quality of the diet against healthy eating and diabetes guidelines, food habit questionnaires were also recorded, in conjunction with the dietary record. These assessments were made the day before the study visit. Data was then computationally analyzed by Dietplan-6, a comprehensive food analysis program (version 6.70.43,

2013; Forestfield Software), and for each participant, the micro- and macro-nutrient content were derived. Goldberg's equation [86] was utilized for assessment of participants' under- and over-reporting of energy intake.

1.3 MRI

Liver and pancreas fat and iron were assessed using multiecho MRI technique, simultaneously [87, 88]. In brief, via a bi-exponential curve-fitting model, relative fat and water proportions of the images (from the multiecho sequence) were derived. Fatty tissues were characterized by significant oscillations, in contrast, tissues with no fat had a smooth decay curve [88]. The iron content (mg/g dry weight tissue) is also measured from proton transverse relaxation rates (R2) [89]. Furthermore, adipose tissue (in litres) was measured using MRI at two levels; abdominal subcutaneous adipose tissue (ASAT), and intra-abdominal adipose tissue (IAAT), which is also referred as visceral adipose tissue (VAT) [87]. Given the availability of the MRI equipment at each center, T1-weighted images were obtained at 1.5 or 3T field strengths for almost half of the participants of the two cohorts. Same protocols were applied across the centers to harmonize the methodology. ImageJ software [90] was utilized to convert the raw data into a readable and analyzable format. Core baseline clinical characteristics of IMI DIRECT participants in the diabetes and non-diabetes cohorts can be found in Table 2.1 [81].

1.4 Omics

DNA was extracted using Maxwell 16 Blood DNA purification kits and a Maxwell 16 semi-automated nucleic acid purification system (Promega). Genotyping was carried out using the Illumina HumanCore array (HCE24 v1.0) and then they were called using Illumina's GenCall algorithm. Samples with call rate <97%, low or excess mean heterozygosity, sex discordance and monozygosity were excluded. Extra quality control (QC) steps were applied for having a high-quality data and samples with the following criterion were excluded: variants not mapped to human genome build GRCh37, variants with duplicate chromosome positions, variants deviating from Hardy–Weinberg equilibrium (exact $p < 0.001$) and call rate <99%. Genotype imputation was conducted with 1000 Genomes (1KG) and the Haplotype Reference Consortium (HRC) as the reference panels. Moreover, ethnic outliers (defined as non-European ancestry) were identified by GCTA software (version 1.24.4) [91] and 6 individuals were removed consequently.

For the transcriptomic, mRNA concentration was assessed by Qubit2.0 from Invitrogen. The quality was checked by TapeStation Software (A.01.04) in conjunction with an RNA Screen Tape from Agilent to check the mRNA quality on gel. Quality evaluation of the libraries was performed using Qubit and TapeStation using DNA1000 Screen Tape. The

approved samples were placed in Flow cell PE using the cBOT system from Illumina and were then sequenced on the Illumina HiSec2000 platform using 49 bp paired-end reads and consequently mapped to the GRCh37 reference genome (2001) with GEM [92]. Via package matchASE from the suite QTL-tools, each expression profile (BAM files) were tested against the imputed genotypes to identify the best matching expression-genotype pair. To get the gender information, genes expression in the autosomal region for the chromosome Y was compared with the expression of the XIST gene in the chromosome X. Gender information that was identified by genotype data, RNAseq data, and reported from clinical data were compared in order to confirm each participant's gender.

To profile the plasma proteins, antibody bead array assays [93] were performed using 779 antibodies targeting 385 proteins selected by the consortium for their association with glycemia-related traits. Analyses of the beads were done using the FlexMap 3D instrument (Luminex Corp.) and xPONENT software (version 4.2). To represent the amount of protein binding to each antibody, the median fluorescence intensity (MFI) was used. Moreover, a targeted proteomic data was generated using several panels of protein assays including Olink proximity extension assays [94], sandwich immunoassay kits using Luminex technology (Merck Millipore and R&D Systems), microfluidic ELISA assays (ProteinSimple [95]), protein analysis by Myriad RBM, and hsCRP analysis (MLM Medical Labs).

Using the Biocrates AbsoluteIDQTM p150 kit, the concentration of 163 metabolites of plasma samples were determined. Multiple steps of QC were applied on the data including: analysis of peak shapes, retention times, compound identity, batch effects, study center effects and effects of different phenotypes. Where the missingness rate exceeded 50% of the sample size and/or the metabolites with zero value of concentration were present for 50% of the participants, data was excluded. The coefficient variation (CV) of measurement for each metabolite over all the plates were calculated and those with CV>0.25 were excluded for not having reliable measurements. If a metabolite's concentration was below the limit of detection (LOD) for more than 95% of the individuals, the metabolite was marked for extra consideration in the downstream analyses. Beside targeted metabolites, untargeted liquid chromatography/mass spectrometry (LC/MS)-based metabolomic data was generated to cover a broader spectrum of metabolites. The approximate missingness rate in the untargeted metabolites was around 20-30% and the missing ones were imputed with the 'multivariate imputation by chained equations' (MICE) method by *mice* package in *R* (version 2.2.5).

IMI DIRECT datasets were the main resource in this thesis and papers I–3 were conducted utilizing the IMI DIRECT data as the primary dataset. Additional details on the IMI DIRECT's omics data can be found in paper II supporting information.

2 UK Biobank

UK Biobank (UKBB) is a large prospective study of over 500,000 participants who were recruited between 2006 and 2010 from 22 assessment centers across the UK. UKBB is a non-profit medical research project and was established with initial funding of £62 million [96]. The overall aim of the UKBB project is to investigate the determinants of a wide range of diseases, such as cancer, heart diseases and diabetes in order to improve the prevention, diagnosis and treatment of such life-threatening illnesses. UKBB is considered a unique and special biorepository for medical research owing to its very large sample size, coverage of many different exposures and enhanced phenotyping [96, 97].

Most of the UKBB participants (99.5%) were middle- or old- aged, between 40 and 69 years, at recruitment. During the baseline visit, participants' blood, urine and saliva samples were taken, and participants also provided detailed information about themselves via a touch-screen questionnaire. Functional and physical measurements were also obtained from participants. These include blood pressure, heart rate, grip strength anthropometrics and an eye examination [96].

At the recruitment visit, participants allowed UKBB for anonymous usage of their data by signing an electronic consent. They also agreed to be contacted for the follow-up visits. All participants agreed that UKBB could access and link their health-related records on death and cancer registrations, hospital inpatient/outpatient episodes and primary care information for the follow-up of their health. To correct for regression dilution, which may result from measurement error and fluctuation of exposure within person, and also to facilitate longitudinal analyses, the baseline assessments were repeated a few years after the baseline visit and will be continually repeated every few years (2013—) [96, 98]. Out of the whole cohort, 20,000—25,000 had a repeat of baseline assessments (at 2013) [96].

Between 2011 and 2012, a dietary web questionnaire was answered by 210,000 participants during the follow-up visits to provide the diet data with nutrient intake. Of this, 80,000 individuals responded to the diet questionnaires over three times. Information about other exposures and outcomes that cannot be identified from health records (e.g. occupation and depression) were asked from 350,000 participants via web questionnaires (2014—). Around 100,000 participants agreed to have their physical activity monitored by wearing wrist accelerometers 24-hours for a week (2013—2015) and 20,000 of those had repeated physical activity measures [96].

All 500,000 participants had a wide range of biochemical assays, including cardiovascular, bone and joint, cancer, diabetes, renal and liver biomarkers, measured from their baseline samples (2014—2015). The biomarkers were selected if they had an established association with disease, were of diagnostic value, or were needed to properly characterize the phenotypes, such as lipids for its association with cardiovascular disease, HbA1c for diabetes

diagnosis and liver function tests for liver disease. Genotyping was also performed on the baseline samples of the whole cohort using either UK BiLEVE Axiom Array (n=50,520) or UK Biobank Axiom Array (n = 438,692). The arrays had 95% overlaps of their markers [96].

During the baseline assessment visit, 100,000 participants also agreed to undergo multimodal imaging assessments of the brain (3T), heart (1.5T), and abdomen (1.5T) (all MRI scans), carotid arteries ultrasound and whole-body (bones and joints) dual-energy X-ray absorptiometry (DXA) scans; all scans were undertaken between 2016 and 2019 [96]. Imaging assessment had a pilot phase between 2014 and 2015 on 5,000 participants. From the conducted abdomen MRI (used in this thesis), fat distribution (liver fat, pancreas fat, ASAT and VAT) was derived with the same protocol and procedure as explained above in IMI DIRECT [99].

The UKBB resource can be accessed by both academia and industry researchers to conduct heath- or medical-related research via the UK Biobank Access Management System (AMS). The 2nd and 3rd papers of this thesis were conducted using UKBB data that was accessed through project application number 18274. The main characteristics of the UKBB cohort (only the European ancestry component) are summarized in the supplementary information of paper III.

3 Diabetes Prevention Program (DPP)

DPP was a randomized controlled trial with the objective of determining strategies to prevent or delay T2D development in high-risk individuals. The DPP was conducted at 27 study centers across the US, between 1996 to 2001. All participants had IGR [i.e., 2-hour glucose level $\geq$7.8 $mmol/l$ and <11.1 mmol/l and fasting glucose level of 5.3–6.9 $mmol/l$ (except for the American Indians)], and they should not have had any previous diabetes diagnosis (except during pregnancy), according to World Health Organization (WHO) (1985) and ADA (1997) criteria. Additional inclusion criteria were age$\geq$25 years and BMI$\geq$24.9 kg/m^2 ($\geq$22 for Asian American) [100–103].

After screening 158,000 individuals and performing 30,986 OGTTs, the trial enrolled 3,234 participants; 45% were from minority groups from African American (20%), Asian American (4%), American Indian (5%), Hispanic/Latino (16%) and the remaining 55% were Caucasian. The enrolled individuals had average characteristics (mean ± SD) as follows; fasting glucose, 6.0 ± 0.5 mmol/l; HbA1c, 5.9 ± 0.5%; age, 51 ± 10.7 years (16% <40 years and 20% $\geq$60 years); BMI, 34.0 ± 6.7 kg/m^2. Moreover, 67.7% of the participants were women, and among all, 71% of women and 66% of men had a family history of diabetes of their first-degree relative [100].

The DPP strategy for recruitment was adapted from previous clinical trials and it was specific to each clinical center. The recruitment took three years (1996-1999) and was mainly covered through direct telephone/mail contact, use of media advertisements, through health care systems, work and employment site. At different steps of screening and recruitment, participants provided informed consent following the Helsinki Declaration [104]. Eventually, the enrolled participants were randomly assigned to either intensive lifestyle, medication with Metformin, or the control group. Participants of all three groups had similar characteristics for diabetes risk factors and were followed for an average of 2.8 years (1.8−4.6 years) [100, 103].

In the lifestyle intervention group (n=1079), participants received training in diet and physical activity, designed to induce 7% of body weight loss. To achieve and maintain weight loss, participants were asked to consume less dietary energy, focusing on reduced fat consumption, and to do moderate exercise for at least 150 minutes a week (e.g. walking and biking). In participants randomized to metformin intervention (n=1073) 850 *mg* of metformin was given twice a day; the comparison group (n=1082) received a placebo intervention (sham pills). Both the metformin and placebo groups received standard of care, which include advice about healthy diet and exercise practices. DPP had a fourth arm, which consisted of troglitazone therapy (Rezulin), but this arm was terminated early owing to serious adverse events related to the intervention (death from liver failure) [103].

The primary outcome of the DPP trial was the development of T2D, which was determined through semi-annual OGTTs, with a positive result recorded when fasting glucose exceeded 7 mmol/l and/or when 2-hour glucose exceeded 11.1 mmol/l, consistent with the 1997 ADA criteria. Within six weeks, a second independent positive test was also required to confirm the diagnosis [101, 103]. The DPP trial had multiple secondary outcomes such as cardiovascular disease and the related risk factors, obesity, changes in insulin sensitivity, insulin secretion and lipoprotein sub-fractions [100, 101, 105]. In total, 2,994 DPP participants consented to genetic studies. These participants were initially genotyped for 1,590 SNPs that were selected based on their known association with T2D [105]. Whether these SNPs affect the diabetes incidence and/or they interact with metformin/lifestyle interventions were further investigated [105]. Moreover, genetic analyses of 32 variants, that were known for their associations with lipid concentrations, lipoprotein sizes and dyslipidemia, were examined using DPP data. The cumulative genetic risk score (GRS) effect of these 32 variants and whether the GRS interacts with lifestyle or metformin interventions were tested in a subsequent analysis [106].

After a median intervention period of 3.2 years, the lifestyle and metformin interventions had lowered the risk of developing diabetes by 58% and 32% respectively, compared to the comparison arm. The DPP lifestyle intervention was efficacious regardless of gender and ethnicity, although the greatest benefits were seen in older participants ($\geq$60years), where diabetes incidence decreased by 71% on average. All participants received a revised version

of the DPP's lifestyle intervention program at the end of the trial. Similarly, the metformin group was also efficacious in both genders and across ethnicities. Among those randomized to metformin, participants aged (25−44 years), or had a BMI$\geq$35 kg/m^2 or who were female and with a history of gestational diabetes benefited most [102, 103].

Out of 1590 tested SNPs, the genetic analyses revealed nominal associations of 85 SNPs with diabetes incidence, nominal interactions of 91 SNPs with the metformin intervention and 69 SNPs with the lifestyle intervention. The GWAS findings of lipid traits, published through 2012 [106], replicated most of the previously associated SNPs for baseline lipid traits and suggested a significant relationship between the GRS and most of the lipid traits and lipoprotein subfractions [106]. Interaction effects between lifestyle intervention and the GRS were observed in association with LDL levels and small LDL particle number; participants with a high genetic burden benefited less from lifestyle intervention than those with low genetic burden [106]. In paper IV of this thesis, we have taken parameters from DPP's published works [105, 106] for our simulation analyses.

Table 2.1: Core baseline clinical characteristics of IMI DIRECT participants in the diabetes and non-diabetes cohorts [81]. Data are mean (SD) except for sex, which is n%.

	Non-diabetes cohort		Diabetes cohort	
	Value	n	Value	n
Male sex, %	76	2127	58	789
Age (years)	62 (6.2)	2127	62 (8.1)	787
Height, cm	174 (8)	2127	171 (9.8)	787
Weight (kg)	85 (13)	2127	89 (17)	787
Waist circumference, cm	99 (11)	2127	103 (13)	781
BMI, kg/m2	27.9 (4.0)	2127	30.5 (5.0)	787
Systolic blood pressure, mmHg	131 (15)	2107	131 (16)	664
Diastolic blood pressure, mmHg	81 (9.0)	2107	75 (9.5)	664
HbA1c, mmol/mol	37 (2.9)	2113	46 (5.8)	784
HbA1c, %	5.5 (0.3)	2113	6.4 (0.5)	784
Fasting glucose, mmol/l	5.7 (0.6)	2126	7.2 (1.4)	787
Fasting insulin, pmol/l	78.2 (54.5)	2124	106.6 (70.9)	787
Fasting HDL-cholesterol, mmol/l	1.3 (0.4)	2123	1.2 (0.4)	789
Fasting LDL-cholesterol, mmol/l	3.2 (0.9)	2123	2.3 (1.0)	781
Fasting triacylglycerol, mmol/l	1.4 (0.6)	2123	1.5 (0.9)	789
ALT, U/l	18 (12)	2120	26 (14)	789
AST, U/l	27 (10)	2052	26 (12)	789
Total cholesterol, mmol/l	5.1 (1)	2123	4.2 (1.2)	789
Fasting intact GLP-1 concentration, pg/ml	0.41 (0.59)	2121	0.67 (1.05)	782
Fasting total GLP-1 concentration, pg/ml	6.5 (4.4)	2120	9.4 (9)	780
Fasting glucagon, pg/ml	98 (41)	2116	111 (51)	758
1 h GLP-1 increment, pg/ml	9.3 (12.1)	2103	9.8 (12.5)	774
1 h glucagon increment, pg/ml	−10.7 (38)	2097	−3.9 (51)	746
Mean 2 h glucose, mmol/l	7.7 (1.5)	2126	9.3 (2)	779
Mean 2 h insulin, pmol/l	383 (260)	2126	457 (275)	779
2 h glucose, mmol/l	5.9 (1.6)	2127	8.6 (2.8)	786
2 h insulin, pmol/l	48 (48)	2102	445 (348)	786
Fasting insulin secretion, pmol min−1 m−2	106 (40)	2126	137 (48)	779
Integral of total insulin secretion, nmol/m2 52 (18) 2126 44 (14) 779 Glucose sensitivity, pmol min−1 m−2 (mmol/l)−1	113 (55)	2126	83 (55)	779
Rate sensitivity, pmol m−2 (mmol/l)−1	921 (699)	2126	1124 (1082)	779
Potentiation factor ratio, dimensionless	1.7 (0.6)	2126	1.4 (0.6)	777
Insulin sensitivity (2 h OGIS), ml min−1 m−2	381 (59)	2118	298 (69)	775
Stumvoll insulin sensitivity index, ml min−1 kg−1	7.8 (2.4)	2099	5.5 (2.7)	775
Matsuda insulin sensitivity index, arbitrary units	5 (3.1)	2126	2.9 (2.2)	779
IAAT, l	5.5 (2.4)	956	5.7 (2.2)	374
ASAT, l	6.1 (2.6)	953	8.1 (3.8)	374
TAAT, l	12 (3.9)	953	14 (4.8)	374
Liver fat, %	5 (4.7)	959	8.7 (7.1)	498
Pancreatic fat, %	13 (8.9)	929	11 (7.3)	446
Liver iron content, mg/g tissue	1.3 (0.26)	958	1.4 (0.31)	498
Pancreatic iron content, mg/g tissue	1.3 (0.43)	927	1.2 (0.33)	447
Mean physical activity intensity: hpfVM, mg	37 (10.2)	1714	34 (9.9)	722
Total energy intake, kJ/day	8213 (3142)	2064	7699 (2519)	707
Carbohydrate intake, g/day	223 (96)	2064	213 (78)	707
Fat intake, g/day	79 (39)	2064	72 (33)	707
Protein intake, g/day	99 (44)	2064	87 (31)	707
Sugar intake, g/day	96 (53)	2064	85 (43)	707
Fibre intake, g/day	20 (9.8)	2064	19 (8.4)	707
Saturated fat intake, g/day	29 (16)	2064	26 (14)	707
Monounsaturated fat intake, g/day	27 (17)	2064	24 (13)	707
Polyunsaturated fat intake, g/day	13 (8.2)	2064	12 (8)	707

Chapter 3

Analytical methods

1 Machine learning

We live in a world filled with a lot of data, and machine learning brings the promise of deriving meaning from all these data. Traditionally, data could be analyzed manually, but the very large and complex nature of many modern datasets have motivated more automated approaches. As such, machine learning and automated systems that can learn from data and unlock the hidden insight in data come to aid. Indeed, machine learning makes human tasks less error prone, faster and easier than before and it can help us do tasks that we never could have achieved on our own [107].

In machine learning, we use data (referred to as *training*) to answer questions (referred to as *making predictions or inference*). Through training, data is used for the creation and fine-tuning of a predictive model and the derived model is then utilized to generate predictions in novel datasets to answer the initial questions. Data is the key component of the whole process and the larger the data sample becomes, the more likely the derived predictions will be reliable and accurate.

1.1 Feature selection - LASSO

The number of parameters is another key factor affecting the performance of a model. Adding irrelevant and redundant parameters generally increases overfitting without improving the model's predictive accuracy or power. To this end, feature selection (i.e. the identification of the most informative independent variables that will be used in subsequent models) is considered a critical step in the process of building models. Regularized regression models, where penalization of uninformative coefficients reduce their effect estimates

to zero, can be utilized during the process of feature selection by pulling out the non-zero parameters from the model. The least absolute shrinkage and selection operator (LASSO) is a commonly used regularized regression method that selects the most informative subset of features, which accordingly can be used in constructing the models [108, 109]. In a linear regression setting, with y as the outcome of interest, n as sample size and m as the number of parameters, the residual sum of squares (RSS) is minimized in order to have the least error and optimal fitting estimation of the coefficients, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. However, in LASSO regression, this step is penalized by introducing a constraint (λ) on the sum of the absolute values of the coefficients (β) and adding it to the RSS value, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \times \sum_{j=1}^{m}|\beta_j|$. λ can be defined as any positive constant value and the larger penalty results in further shrinkage of the estimates towards zero [109].

1.2 Supervised learning: regression - classification

Machine learning is a broad field spanning an entire family of analytical techniques. At a high-level, machine learning can be categorized into unsupervised and supervised learnings. Supervised learning involves functions that map an input to an output based on a series of input-output pairs as examples. Furthermore, supervised learning can be sub-categorized into regression and classification [110, 111].

In regression models, the relationship between the outcome target variable (continuous) and the independent predictor variable(s) is estimated through statistical processes. Some of the most common regression models include *linear regression*, *decision tree*, *random forest* and *neural networks*. Linear regression is simply finding a line that fits the data, with its extensions as multiple linear regression (finding a plane of best fit), and polynomial regression (finding a curve for best fit). A decision tree is a tree-like model with nodes representing a decision and branches as the consequence of the taken decision. Random forest is an assembling technique constructed of decision trees [112]. It involves creating multiple decision trees using bootstrap data sets of original data and randomly selecting a subset of variables at each step of the decision tree. The model relies on the *Majority Wins* model of all the decision trees in order to reduce the risk of error from an individual tree. The neural network is another popular method, which is comprised of multi-layered connected units (nodes), inspired by the neurons in the human brain [110].

Decision trees, random forest and neural networks can also be deployed as classification models. In fact, they follow the same logic as regression but the only difference is that the output is discrete rather than continuous. Some of the other most common types of classification models include *logistic regression*, *support vector machine* and *Naive Bayes*. Logistic regression is similar to linear regression but is used to model the probability of a finite number of outcomes typically two (0/1). The support vector machine, as a supervised classification technique, carries an objective to find a hyperplane in n-dimensional space

that can distinctly classify the data points. Naive Bayes acts as a probabilistic model for classification tasks with cuts of the classifier defined based on the Bayes theorem [110, 111].

1.3 Unsupervised learning: clustering - dimensionality reduction

Unlike supervised learning, unsupervised learning is used to draw inferences and find patterns from input data without references to the labeled outcome. Two main methods used in unsupervised learning include clustering and dimensionality reduction. Clustering involves grouping of the data points. Common clustering techniques include *K-means clustering*, *hierarchical clustering*, *mean shift clustering* and *density-based clustering*, while each technique has different methods in finding clusters, they all aim to achieve the same thing. Dimensionality reduction is a process of reducing dimensions of the feature set. Most dimensionality reduction techniques can be categorized as either feature elimination or feature extraction. It is important to make a distinction between feature extraction and feature selection, the former extracts newly defined variables and the latter only subsets the same variables. A popular method of dimensionality reduction is called *principal component analyses (PCA)*, which aims to uncover the low-dimensional patterns of big data through a hierarchical coordinate system and to capture the maximum amount of variance in data [113]. PCA is widely used in population genetics and it helps with identifying patterns of distribution communality in different geographic regions and ethnicities [114].

1.4 Performance metrics and model evaluation

Once the model training is complete, next is to evaluate the model against data that has not been used for training. This is meant to be representative of how the model might perform in the real world. When there is no access to an external data set for model evaluation, the whole data can be divided into separate training and testing sets, prior to the analysis. *Cross-validation* is a common approach for resampling the dataset into training and testing sets in order to evaluate the developed models. The method is usually called *k-fold cross-validation*, where k refers to the number of split groups. K-fold cross-validation is a favored approach as it results in minimally biased model evaluation.

There are different evaluation metrics and based on the defined aim and question one can elect which metric to use. Some of the main evaluation metrics include *Confusion Matrix*, *Accuracy*, *Sensitivity*, *Specificity*, *Precision*, *F1 score* and *Area Under Curve (AUC)*. A confusion matrix is a way of evaluating classification problems such as diagnostic tests. Table 3.1 summarizes a 2-class confusion matrix for a diagnostic test (positive/negative), where rows correspond to what the machine learning model predicted and columns represent the condition as in the reality. Four possible information can be retrieved from this binary diagnostic test including; true positive (TP), where cases are with the positive condition

Table 3.1: A 2-class confusion matrix for a diagnostic test (positive/negative), rows corresponding to what the model predicted and columns representing the condition as in the reality.

	Positive condition	Negative condition
Positive model output	true positive (TP)	false positive (FP)
Negative model output	false negative (FN)	true negative (TN)

and are correctly predicted as positive; false positive (FP), where cases are with the negative condition but are wrongly predicted as positive; false negative (FN), where cases have the positive condition but are wrongly predicted as negative; true negative (TN), where cases are with the negative condition and are correctly predicted as negative.

'Accuracy' is a metric indicating the correctly predicted cases (Equation 3.1). When having imbalanced groups, the balanced accuracy can be measured instead, by averaging the accuracies that are calculated within groups. Sensitivity can be defined as the test's ability to correctly identifying the positive conditions (Equation 3.2), whereas specificity is the test's ability in correctly identifying the negative conditions (Equation 3.3). F1 score is another metric, which is a balanced and harmonic mean of precision (Equation 3.4) and sensitivity, defined as Equation 3.5. A receiver operating characteristic (ROC) curve is a graphical model that plots sensitivity (y-axis) vs. 1-specificity (x-axis) at different discrimination thresholds. Measuring the AUC of a ROC curve can be used in evaluating the overall prediction of a model. The closest point to the upper left corner of the ROC is usually considered as the best cutoff, however, depending on the purpose of the predictions the cut-off can be chosen as a trade-off between specificity and sensitivity.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{3.1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.3}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.4}$$

$$F1score = \frac{2 \times precision \times sensitivity}{precision + sensitivity} \tag{3.5}$$

2 Causal inference

When it is determined that an exposure/treatment is associated with the outcome/disease, the question that comes after is whether this observed association reflects a causal relationship or not (correlation may not equal causation). Even a statistically strong association

between an exposure and the outcome does not necessarily imply causality. A cause must proceed the outcome and must be necessary for the outcome to occur. Causal inference methods help answer questions like 'what is a given treatment's causal effect on a particular individual, measured by a particular outcome?'. Indeed, if we could know the outcome value in the absence of the exposure/treatment (known as the *counterfactual*) [115, 116], then we could assess the causal effect by subtracting the outcome value of the two conditions. However, the fundamental problem is that the counterfactual is impossible to determine. Once someone is treated, we can only see the treatment effect, and it is impossible to go back in time and to know what their outcome would be if they were not treated. It is also important to note that causality is not ascertained by only one component cause, and several pieces of evidence are required to claim a conclusion. Furthermore, different mechanisms may also result in the same outcome or disease, each having direct and indirect causes components involved. Understanding causal mechanisms is important, as eventually based on these, public health decisions are made and interventions are designed. SEMs, BNs and instrumental variables (IVs) (like the MR approach) are examples of different methods deployed in the causal inference of epidemiological studies and have been utilized in paper I and paper III of this thesis.

2.1 Structural Equation Modeling (SEM)

SEM is a multivariate data analysis technique that can be used in testing and establishing causal relations among variables. Mostly, SEMs are used as confirmatory modeling in testing a represented hypothesis. SEM can simultaneously test the measurement and the causal structural relationships among the variables [117]. Covariance-based SEM is one of the most preferred methods for confirming or rejecting theories through hypotheses testing, particularly when the sample size is large enough, the data is normally distributed, and most importantly, the model is correctly specified. Covariance can be defined as the expected (E) or mean value of how much the two random variables (X and Y) jointly change, $E[(X - E[X])(Y - E[Y])]$.

Typically, the *Chi-squared* test (χ^2)) is deployed as a measure of fit by comparing the observed covariance matrix with the implied covariance matrix of the proposed model. There are several other fitting indices, including *root mean square error of approximation (RMSEA)*, *the comparative fit index (CFI)* [118] and *the Tucker–Lewis index (TLI)* [119]. In brief, RMSEA, which is considered as an absolute fit index, measures how far the proposed model is from a best-fitting model (a bigger value corresponds to a worse fit). Whereas, CFI and TLI, defined as incremental (relative) fit indices, measure the difference between the proposed model and the worst-fitting null model (a bigger value corresponds to a better

fit) [120–122]. The computational formulas of RMSEA, CFI and TLI are as follows [120]:

$$RMSEA = \frac{\sqrt{\chi^2 - df}}{\sqrt{df \times (N-1)}} \tag{3.6}$$

$$CFI = \frac{\frac{\chi^2}{df}(null\ model) - \frac{\chi^2}{df}(proposed\ model)}{\frac{\chi^2}{df}(null\ model) - 1} \tag{3.7}$$

$$TLI = \frac{d(null\ model) - d(proposed\ model)}{d(null\ model)}, \tag{3.8}$$

where N represents the sample size, df represents the model's degrees of freedom, and d represents the $\chi^2 - df$ [120].

The variables of an SEM are mainly connected through independent regression equations and are represented through path diagrams, where observed variables are represented as a rectangle or square node, unobserved (latent) variables by a circle or ellipse nodes, and arrows represent the causal association. SEMs have several advantages over regression models. For instance, regression models are restricted to have only a single dependent variable, on the other hand, in SEM analysis, multiple dependent variables can be used. SEM allows for measurement errors and having unobserved variables, whereas regression analysis assumes perfect measurement. SEM can also be used to test mediation or indirect effect in just one step, however, regression analysis can test the mediation effect in a series of steps. All the above are examples of why SEM is a favored method in hypothesis testing. We have utilized this method in testing the twin-cycle hypothesis in the paper I of this thesis.

2.2 Bayesian networks (BN)

A Bayesian network has two components; a qualitative component in the form of a DAG and a numerical component that quantifies the DAG's structure. A DAG is a graphical tool for a visual representation of the causal pathways among a set of random variables (nodes in the graph). The relationships between these nodes are depicted with arrows (edges/arcs) and, as the name suggests, a DAG is a graph that has no cyclic paths between variables [123–125].

In a BN, the arrows encode assumptions about the conditional probability distributions of variables that are specified by model parameters. In brief, P(A|B) represents the conditional probability of event A occurring given that event B has already occurred. Following the Bays theorem, this can be calculated as $\frac{P(A \cap B)}{P(B)}$ and can be presented in a DAG with an arrow from node B (parent) to node A (child). If A and B are independent events then the expression would be P(A|B)= P(A) and there would be no link between the nodes [124, 125].

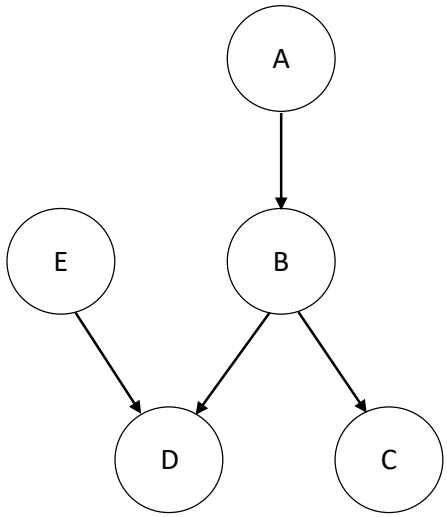

Figure 3.1: A simple example of a Bayesian network structure with a joint probability distribution calculated as $P(A, B, C, D, E) = P(D|B, E) \times P(C|B) \times P(B|A) \times P(E) \times P(A)$.

In general, the two variables in a network are considered as independent in the absence of a direct link between them, conditioning on other intermediate variables *(Markov condition)* [123]. As such, the joint probability distribution of a BN with X = $\{x_1,\ldots,x_n\}$ as variables and $\Theta = \{\theta_1, ..., \theta_n\}$ as parameters can be formulated as a product of conditional probabilities of each node given its parents $pa(x_i)$, and parameters (Equation 3.9). Figure 3.1 shows a simple BN, the joint probability of which, can be calculated as $P(D|B, E) \times P(C|B) \times P(B|A) \times P(E) \times P(A)$, following the Equation 3.9 [123–125].

$$P(X|\Theta) = \prod_{i=1}^{n} P(x_i|pa(x_i), \theta_i) \tag{3.9}$$

Three main classes are considered in learning the structure of a BN including *constraint-based*, *score-based* and *hybrid* algorithms. Via constraint-based algorithms, conditional independence associations are found using statistical tests such as χ^2. Score-based algorithms are based on general optimization techniques aiming to identify the best-fitting and highest-score DAG [72]. The scoring of networks can be done through different scoring methods such as *Bayesian information criterion (BIC)* and *Akaike information criterion (AIC)*, where they both are based on likelihood function with different penalizing terms for the number of model parameters, with BIC imposing a bigger penalty. Hybrid algorithms are combinations of score-based and constraint-based methods trying to find the highest-score network in a limited search space [123, 126].

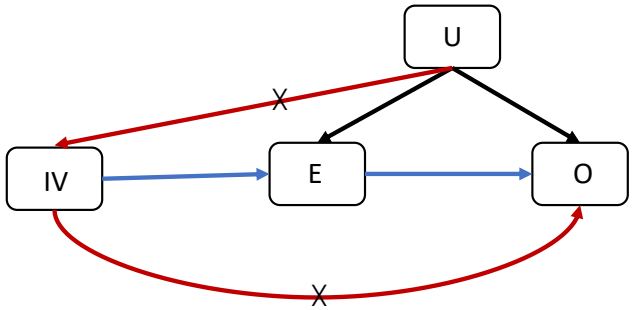

Figure 3.2: Graphical representation of the instrumental variables method (IVs).
The blue lines represents the causal association between exposure and outcome (E-O) using the IV as a proxy for E. Red crossed lines show the violation of the IV criteria, where IV is associated with unmeasured confounders (U) of the E−O association and IV is associated with O not only through E but also through another causal pathway.

BNs, unlike SEMs, can be utilized on data without defining any hypothesis, allowing for learning new knowledge from data [117]. However, it is also possible to force the inclusion or exclusion of some of the edges based on prior knowledge, given that they do not violate the acyclic characteristic of the DAG. In the paper III of this thesis, we used BN to examine causal pathways and to infer causality.

2.3 Instrumental variables (IVs) - Mendelian randomization (MR)

One of the oldest methods for learning about causality, using observational data, is the IVs method that can deal with unmeasured confounding. With the help of an IV, the causal relationship between an exposure (E) and an outcome (O) can be examined. A valid IV should meet some criteria to help with an unbiased estimation of the causal association between E and O (visualized in Figure 3.2) [127]:

- IV should directly be associated with E.
- IV should not be associated with confounders of the E−O association
- IV's association with O should only be through the E and there should not be any other causal pathway from IV to O.

Under these assumptions, the association between E and O can be estimated as the ratio of their associations with the instrument, $\frac{IV-O}{IV-E}$.

MR is an IV approach that uses the genetic variant(s) as instruments (see Chapter 1 - Section 3). Either a single or multiple genetic variants can be deployed as instruments, the latter may improve the statistical power and is a common practice when the exposure is a complex trait. The genetic instruments can be obtained through GWAS databases for their association with the exposure of interest [128, 129].

The genetic variants should follow the IV criteria in order to be valid instruments. The first criterion can be verified by the strength of the genetic variant with the exposure of interest. To verify the second assumption, the relationship between the genetic variant(s) with the possible confounders should be examined. Important to note that it is impossible to test all the confounders and the absence of such associations can only improve our confidence in holding the second assumption. Genetic variant(s) are likely to have associations with the outcome through other biological pathways (known as *horizontal pleiotropy*) and this might affect the certainty of verifying the third assumption. However, when having several genetic instruments, this can be tested through sensitivity analyses such as the *MR-Egger* method. Furthermore, in a similar examination, multiple fitting models can be applied to each instrument and if the results were heterogeneous, it can be an indication that the third criterion is violated. If pleiotropy is present, the violating genetic variant should be excluded or controlled through relevant statistical methods [128–132].

MR approach can be applied using both in-person data or summary statistic data, the former is called *one-sample MR* and the latter is *two-sample MR*. The one-sample MR can be conducted through the *2-stage least squares (2SLS)* regression, where at its first stage, exposure is predicted from the genetic instrument, and at the second stage, the outcome is regressed on the predicted exposure providing a causal estimate that is most likely independent of confounders [133]. The two-sample MR with only a single variant can be estimated through the *Wald ratio* method (Equation 3.10), and the standard error (se) of the causal estimate can be calculated through Equation 3.11 [133].

$$\beta_{E-O} = \frac{\beta_{IV-O}}{\beta_{IV-E}} \pm se \tag{3.10}$$

$$se = |\beta_{E-O}| \sqrt{\left(\frac{se_{IV-E}}{\beta_{IV-E}}\right)^2 + \left(\frac{se_{IV-O}}{\beta_{IV-O}}\right)^2} \tag{3.11}$$

The inverse variance-weighted (IVW) method is a meta-analysis approach that can be utilized for combining the effect estimates of multiple variants into an overall causal estimate (Equation 3.12).

$$\beta_R = \frac{\sum_{j=1}^{m} w_j \hat{\beta_{Rj}}}{\sum_{j=1}^{m} w_j}, \tag{3.12}$$

where $\hat{\beta_{Rj}}$ is the causal effect estimate of variant j derived from the ratio method, w_j as the inverse variance weight of that and m the number of variants [134]. The two-sample MR analysis in conjunction with BN modeling was deployed in paper III of this thesis.

Table 3.2: The table represents the true or false relationships between the real H_0 (columns) and the observed H_0 (rows), that is retrieved as a result of a test; α, type I error; β, type II error.

	True H_0	False H_0
H_0 is not rejected	True negative (1-α)	False negative (β)
H_0 is rejected	False positive (α)	True positive (1-β)

3 Statistical power

In statistics, when testing a hypothesis against a null hypothesis (H_0), we are concerned about two different types of errors; type I error (α), which is the probability of rejecting the H_0 incorrectly, and type II error (β), which is the probability of not rejecting a false H_0. Statistical power is inversely related to the probability of making a type II error ($1 - \beta$) (see Table 3.2) [135]. Through a statistical power analysis, the sample size that is required to ensure a high probability of correctly disproving the H_0 can be quantified.

The main factors that influence the statistical power are α (the significance level), effect size/parameter estimate, which is the magnitude of difference between the null and alternative hypothesis, and the sample size [136, 137]. α level is the statistical significance criterion used in the test and is usually set as 0.05. The smaller α increases the chance of β and results in lower statistical power. The higher the effect size, the larger the difference will be, and the higher the power will be. The sample size used to detect the effect is positively associated with power.

In general, the power of a test can be estimated via formula 3.13, where $Z_{1-\beta}$ is the z-score of power, γ is the effect size, $SE(\gamma)$ is the standard error of the effect size (a function of sample size), $Z_{1-\frac{\alpha}{2}}$ is the z-score of $1 - \frac{\alpha}{2}$ (α as the probability of making type I error), following a standard normal distribution [138].

$$Z_{1-\beta} \approx \frac{\gamma}{SE(\gamma)} - Z_{1-\frac{\alpha}{2}} \tag{3.13}$$

Statistical power calculations can (and typically should be) performed *a priori*, although in some rare cases *post-hoc* power calculations are also justified. *A priori* power analysis determines the sample size that is needed for designing a study, given the desired statistical power (usually 0.8) and the effect size (taken from prior research or a personal assessment). *A priori* power analysis is done before the actual data collection process and initial analyses. On the other hand, *post-hoc*, determines the power of a completed study based on the sample size and the calculated effect size. *A priori* power calculation is usually required for getting a study proposal approved. Designing a study with the recommended sample size will improve our confidence in accepting the study results, regardless of the *p-value*. Whereas, having low power, due to insufficient sample size, will result in failing to reject

the H_0, even though the alternative hypothesis is true [136]. A *post-hoc* power calculation, on the other hand, is generally criticized for not providing much more information beyond the *p-value*. Moreover, a *post-hoc* power calculation is assuming an equal effect size in the sample and in the population, which might lead to a misleading power estimate [138–140].

The concept behind power corresponds to the frequency of rejecting the H_0, meaning that, if the studies are repeated, how often and how strongly we can disprove the H_0. To this aim, study designs can be generated repeatedly via computer simulations and the average of H_0 rejection over these simulations can represent the power. This is known as the *Zero/One method*, where each simulation is scored with 1 given that the p-value for the parameter of interest is less than the significance level (usually α <0.05), and scored with 0 if greater or above the significance level. The scores will then be averaged across the simulation iterations to provide the statistical power value. Following this simulation-based approach, power can also be estimated through formula 3.13, meaning that data will be simulated repeatedly based on the alternative hypothesis and the average over the standard error and effect estimates of these simulations will result in the power estimation (*standard error method*). In order to get an accurate power estimate, a large enough number of simulation iterations is required [138].

Chapter 4

Results and Discussions

This chapter presents summaries of the papers included in this thesis on enhancing prediction and causal inference in metabolic dyshomeostasis. In paper I, the SEM method was used to test the established 'twin-cycle' hypothesis concerning interactions between the liver, pancreas and adipose tissue in the etiology of T2D. In addition, the association of physical activity with glucose regulation within the twin-cycle model was further examined. In paper II, I developed a series of prediction models for the diagnosis of fatty liver with complex input data including clinical and multi-omic datasets. Several machine learning approaches were deployed for feature selection, model training, and model validation. In paper III, BN modeling was used to investigate the causal paths involved in liver fat accumulation and to visualize them with the help of DAGs. Furthermore, multiple two-sample MR analyses were conducted to validate the detected pathways. In paper IV, I undertook simulations to generate power calculations to test the hypotheses that an innovative clinical trial method termed 'GBR' has superior statistical power to detect gene-treatment interactions compared with the conventional RCT approach. In the following, I briefly discuss these four papers and summarize some of the main findings therein. The final chapter concludes with an overall summary of all papers and the future perspectives.

1 Paper I

In paper I, we sought to test the 'twin-cycle' hypothesis (read more on Chapter 1 - Section 2) separately (TC) and when physical activity was added to the model (TC-PA), via the SEM method. As the second author of that paper, my main contribution was introducing the SEM method for testing the aforementioned hypotheses. Throughout the data analysis process, I provided feedback on the analytical approach, but the first author (Koivula) was

the one who implemented the statistical analyses.

This project was conducted utilizing complete case data from the IMI DIRECT diabetes ($n \leq 435$) and non-diabetes ($n \leq 920$) cohorts in two separate analyses. Emulating the twin-cycle hypothesis, we incorporated the following variables into the models: fasting glucose, 2-hour glucose, fasting(basal) insulin secretion rate, liver fat, pancreas fat, TG, mean physical activity intensity assessed by accelerometry, OGIS as an index of insulin sensitivity, and glucose sensitivity, which is insulin secretion in response to glucose. These variables were transformed for normality and residualized for age, center and energy intakes (fat, carbohydrates and protein) for the subsequent analyses.

There is existing evidence that physical activity, glycemic regulation and abdominal fat are mechanistically interrelated [27, 141], but no study prior to ours had quantified the magnitude of the relationships within this mechanistic network. Understanding this might aid the design of lifestyle intervention trials seeking to improve glycemic regulation and related metabolic traits. With this in mind, we extended the twin-cycle model by examining the extent to which physical activity perturbs this metabolic network (Figure 4.1).

In the SEM diagram (Figure 4.1), the direct edges denote the regression coefficients and were estimated for our testing models. The mediated associations were also assessed for the indirect pathways by coefficient product method [143], which is simply multiplying the involved pathways from physical activity towards fasting/2-hour glucose. Model fits were estimated through χ^2, RMSEA, CFI and TLI (read more on Chapter 3 - Subsection 2.1). In addition, we defined random models by repeated permutations of the variables with each other and then compared the average χ^2 of all the shuffled models with the initial model's χ^2 through a t-test. We took this approach as there is no complete agreement on the cut-off values in defining a good-fitting model.

Through the conducted correlation analysis, physical activity showed a significant association with most of the variables of the model which improves our confidence in testing the SEM model in presence of the physical activity. Both the TC and TC-PA models had better goodness of fitting values in comparison with the shuffled and null hypotheses. Most of the direct edges in the TC model were significant and were matching with the twin-cycle hypothesis in terms of direction. Some of the significant observed direct associations for the TC-PA model include:

- both cohorts: physical activity $\rightarrow$ insulin sensitivity (OGIS) $\uparrow$
- both cohorts: physical activity $\rightarrow$ fasting insulin secretion rate $\downarrow$
- both cohorts: physical activity $\rightarrow$ TG $\downarrow$
- only in diabetes cohort: physical activity $\rightarrow$ glucose sensitivity $\downarrow$

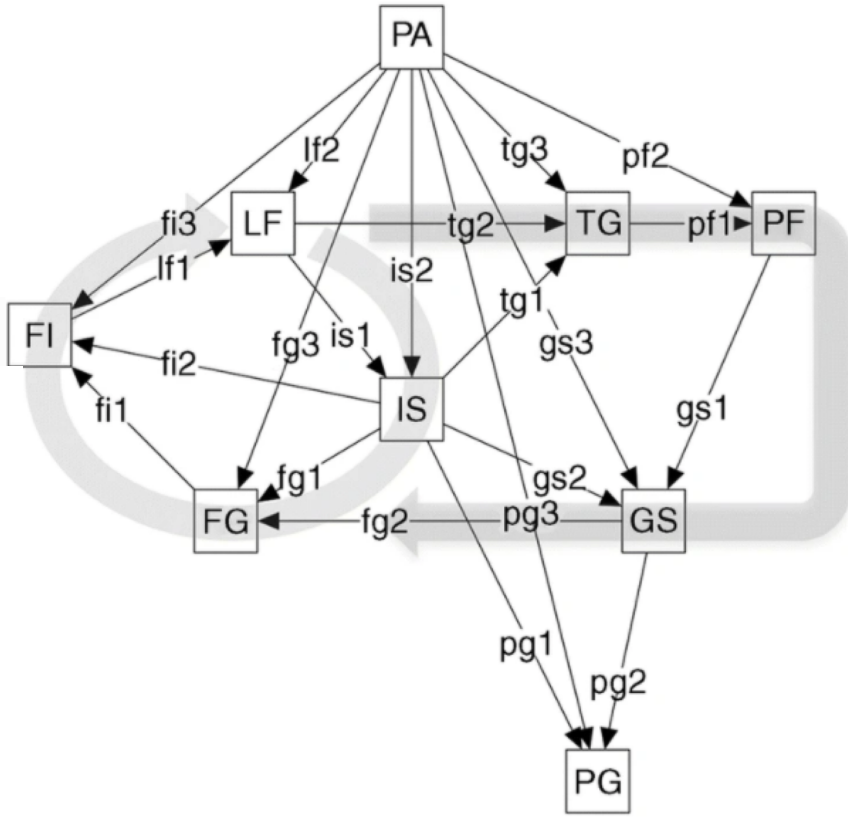

Figure 4.1: The diagram shows the structural equation (SEM) model for the twin-cycle [38] plus physical activity (TC-PA) hypothesis [142]. Nodes (squares) of the SEM are physical activity (PA), liver fat (LF), fasting insulin secretion rate (FI), fasting glucose (FG), insulin sensitivity (IS), triglycerides (TG), pancreas fat (PF), glucose sensitivity (GS) and 2-hour glucose (PG). The arrows represent regression coefficients from exposures towards outcomes.

There were no direct significant association from physical activity to liver fat nor to 2-hour glucose in any of the cohorts.

The indirect and mediated association from physical activity to fasting/2-hour glucose were mainly through the liver fat cycle and the following significant pathways were observed:

- both cohorts: physical activity → insulin sensitivity↑ → fasting/2-hour glucose↓
- non-diabetes cohort only: physical activity → fasting insulin secretion rate↓ → liver fat↓ → insulin sensitivity (OGIS)↑ → fasting/2-hour glucose↓
- non-diabetes cohort only: physical activity → insulin sensitivity (OGIS)↑ → glucose sensitivity↓ → fasting/2-hour glucose↑
- diabetes cohort: physical activity → glucose sensitivity↓ → fasting/2-hour glucose↑

In summary, using diabetes and non-diabetes cohorts of the IMI DIRECT consortium, our SEM analyses on the twin-cycle hypothesis plus physical activity showed the association of physical activity with several metabolic traits and factors. Additionally, the results showed that the effect of physical activity on glucose was mediated through basal insulin secretion rate, insulin sensitivity and liver fat. Our analyses could only validate the liver cycle of the twin-cycle model, whereas the pancreas cycle was not supported by our SEM analyses.

2 Paper II

I completed Paper II during the second and third years of my PhD studies and is the most independent work of all the published papers of this thesis. The purpose of this paper was to develop diagnostic tools for NAFLD as pragmatic alternatives to expensive scanning or biopsy methods and to improve our understanding of the disease etiology. The project was conducted using data from diabetes and non-diabetes cohorts of the IMI DIRECT study. The criterion measure was MRI-derived liver fat percentage stratified at 5%, the below of which defined as individuals without fatty liver, and equal or above as individuals with fatty liver. Figure 4.2-A shows the liver fat percentage across diabetes and non-diabetes cohorts, and panel B shows its distribution among different centers of the IMI DIRECT combined cohorts.

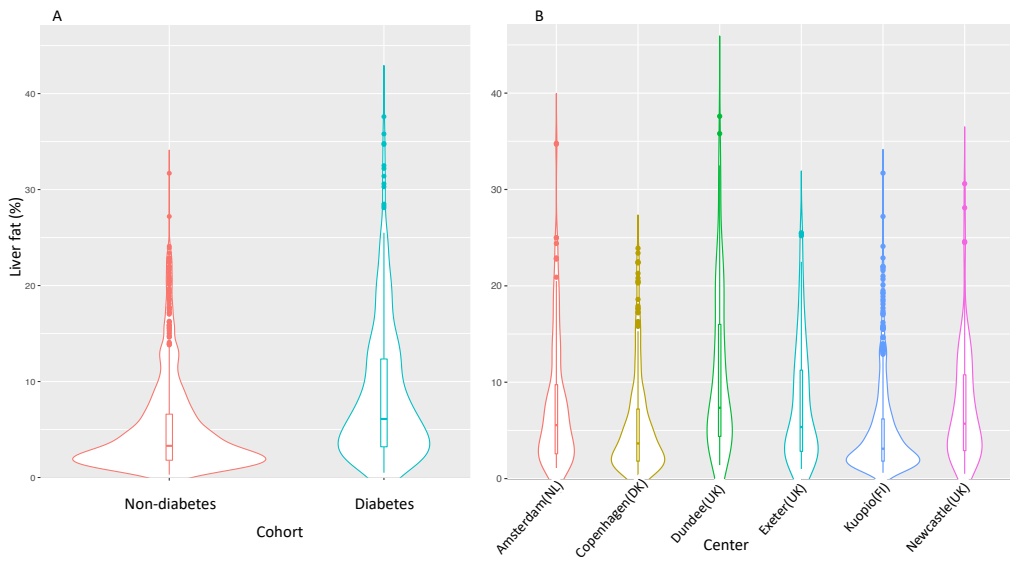

Figure 4.2: Violin plots of liver fat distribution across the diabetes and non-diabetes cohorts (A) and among different centers of the combined cohorts of the IMI DIRECT study (B).

The key input data include multilevel datasets of environmental exposures (triaxial acceler-

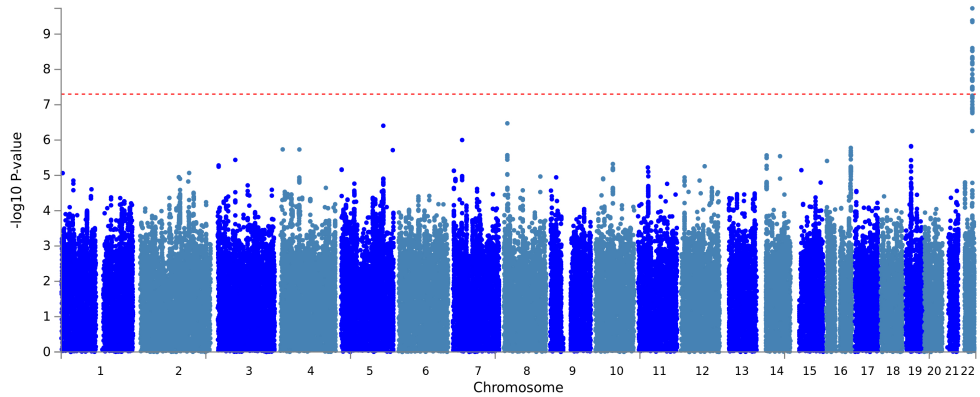

Figure 4.3: Manhattan plot for the association of around 18 million single nucleotide polymorphisms (SNPs) across the 22 autosomal chromosomes (X-axis) with the liver fat content. Y-axis shows the negative logarithm of the association p-values per SNP.

ometry for physical activity, multi-pass dietary record, and food habit questionnaire diet), extensive biological intermediates (glucose, insulin, several measurements of beta-cell function, HbA1c, GLP-1, blood lipids, liver enzymes, and several other biochemistry assays, assessed through the drawn blood samples), and deep omics profiling (genetic, transcriptomic, targeted and exploratory proteomic, targeted and untargeted metabolomic).

To select the most informative features of this high-dimensional input dataset, I deployed two approaches: for the clinical datasets, I considered their Pearson correlation matrices and the accessibility of the input variables within the clinical setting, whereas, for each layer of the omic data, I applied the LASSO method and excluded the least informative variables. For the genetic data, prior to undertaking the LASSO analysis I ran a GWAS analysis (with RVTESTS v2.0.2 [144]) and selected robustly associated SNPs (with p-value $< 5 \times 10^{-6}$) as input variables for the following LASSO analysis. Figure 4.3 displays the Manhattan plot for the association of around 18 million SNPs across the 22 autosomal chromosomes with liver fat content as the dependent variable.

It is important to consider that the feature selection method was applied to 70% of the complete data (n=735) and left the remaining 30% (n=324) for model training. The number of input variables before and after the feature selection step can be seen in the left panel of Figure 4.4. The selected features were then included as input variables for the model training step using machine learning approaches. After training with several machine learning models, I elected to train with random forest (5×5 -fold cross-validated), as it had a better or similar performance amongst all methods. Separate models for each omic type or models combing both the omics variables and clinical measures were developed. The performance of the models was described using ROCAUC (see Figure 4.4, right panel). Of all models,

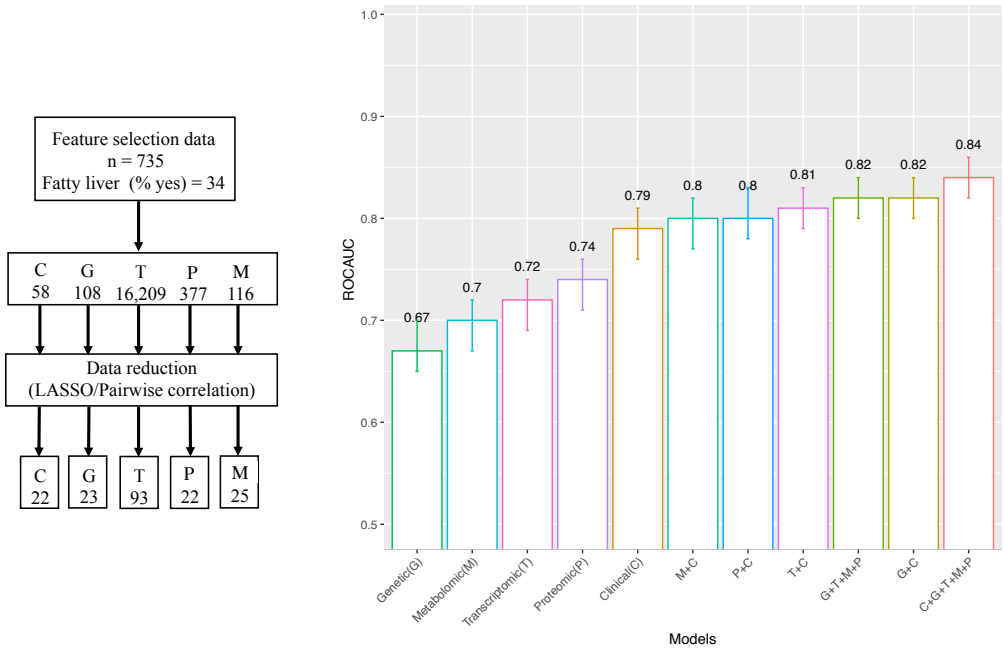

Figure 4.4: Left panel: feature selection step prior to the model building with The least absolute shrinkage and selection operator (LASSO) method for omics and the Pearson pairwise correlation for the clinical variables of the IMI DIRECT combined cohorts. Right panel: model performance measured by the receiver operating characteristic area under the curve (ROCAUC) with a 95% confidence interval for each omic model separately or plus clinical models.
For the clinical data, 22 out of 58; the genetic (G) data, 23 out of 108 single nucleotide polymorphisms (SNPs); for the transcriptomics, 93 out of 16,209 genes; for the exploratory proteomics (P) 22 out of 377 proteins; for the targeted metabolomics (M), 25 out of 116 metabolites were selected and used in training models.

the advanced model comprised of clinical data plus all the omics resulted in the highest ROCAUC (0.84; 95% CI 0.82, 0.86).

The most predictive variables were estimated through the 'permutation accuracy importance' technique, which is calculated based on the decrease in random forest model performance when each variable is randomly permuted (Figure 4.5, right panel). We also utilized the ensemble feature selection (EFS) method (see Figure4.5, left panel), which ranks the most informative predictors by accumulating their importance value from different approaches: Pearson's product moment correlation (P_cor), Spearman's rank correlation (S_cor), beta-values of logistic regression (LogReg), the error-rate based (ER_RF) and the Gini-index based (Gini_RF) variable importance measures of the random forest [145]. Both methods had a very similar importance list, highlighting beta-cell function and insulin sensitivity as the most informative predictors of all omics and clinical variables in the advanced model.

We also developed three clinical models with variables selected based on literature and graded them on their accessibility in the clinical settings. These three models were developed on diabetes, non-diabetes, and combined datasets of the IMI DIRECT and were

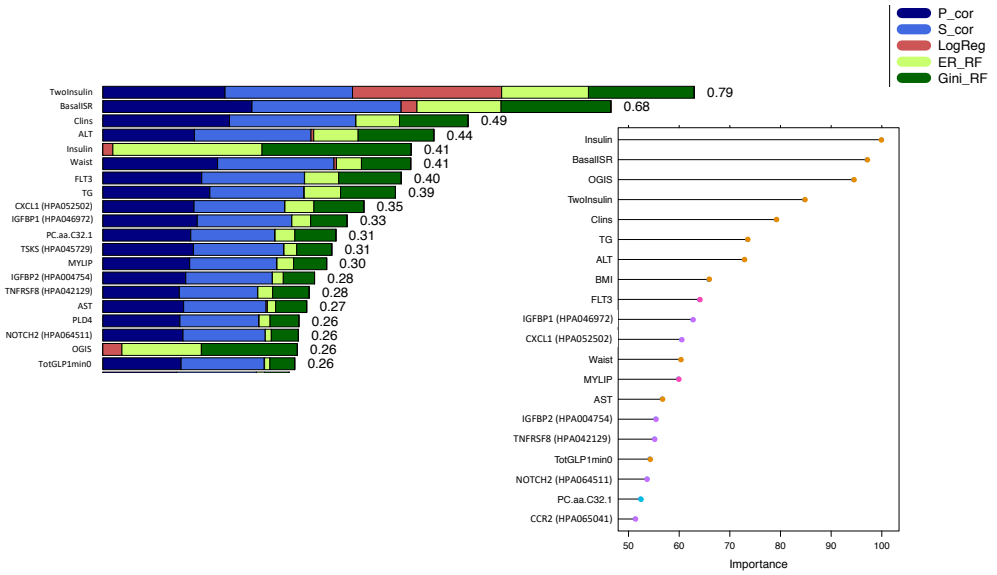

Figure 4.5: Right panel: the most predictive variables were estimated through the "permutation accuracy importance" technique, which is calculated based on the decrease in random forest model performance when each variable is randomly permuted. Left panel: the most informative predictors are ranked via the ensemble feature selection (EFS) method [145], which accumulates the importance value from different approaches including Pearson's product moment correlation (P_cor), Spearman's rank correlation (S_cor), beta-values of logistic regression (LogReg), the error-rate based (ER_RF) and the Gini-index based (Gini_RF) variable importance measures of the random forest.
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); Glucagonmin0, fasting glucagon concentration; Glucose, fasting glucose from venous plasma samples; Insulin, fasting insulin from venous plasma samples; OGIS, oral glucose insulin sensitivity index according to the method of Mari et al.; TG, fasting triglycerides; TotGLP1min0, concentration of fasting total GLP-1 in plasma; TwoInsulin, 2-hour insulin.

compared with three well-known liver fat indices; fatty liver index (FLI), hepatic steatosis index (HSI), and non-alcoholic fatty liver disease liver fat score (NAFLD-LFS)). Amongst all, model 3 had the highest ROCAUC of 0.82 (95% CI 0.81, 0.83) in the combined cohorts, with BMI, waist circumference, ALT, AST, TG, fasting glucose, fasting insulin, alcohol consumption, and diabetes status as the models' input variables.

Several performance metrics including sensitivity, specificity, balanced accuracy, and F1 measure were considered at different probability thresholds in the random forest model's output. Owing to the manner in which we perceived these models might be used in the clinical setting, we were seeking models with both high sensitivity and high specificity; therefore, for our developed models (models 1-3), we suggested cut-offs on the highest observed balanced accuracy (at 0.4, 0.6, and 0.4, for non-diabetes, diabetes and the combined cohorts, respectively).

Considering the availability of data, we could externally validate models 1 and 2 using UKBB datasets which had similar performances to the IMI DIRECT ones, model 1 with

ROCAUC of 0.71 (95% CI 0.69, 0.73) and model 2 with ROCAUC of 0.79 (95% CI 0.77, 0.80). The input variables of model 1 were BMI, waist circumference, alcohol consumption, diabetes status, systolic and diastolic blood pressure (SBP, DBP). Model 2 had BMI, waist circumference, ALT, AST, TG, fasting glucose, alcohol consumption, and diabetes status as input variables.

In order to maximize access to our models, we developed a web interface for NAFLD diagnosis using models 1-3 and made it available online (*https://www.predictliverfat.org*). To conclude, our results suggest the benefits of the proposed models to screen at-risk populations for NAFLD. In addition, the derived importance lists of each data set (clinical, genetic, transcriptomic, proteomic, and metabolomic) were highlighting the previous findings and suggesting potential molecular features of the NAFLD etiology.

3 Paper III

I completed Paper III during the final few months of my PhD studies, with methods and analysis design mainly determined by me. The analyses were performed with another more junior PhD student within my research group. Thus, this project gave me the opportunity to practice supervising another student's work and to share the knowledge I had acquired during my studies with other members of our research group.

The main purpose of this analysis was to study the causal relationships between dysmetabolic processes and liver fat accumulation, which might help with better understanding the etiology of NAFLD and T2D. To assess the causal pathways, we utilized the BN method, and to validate our findings we undertook a series of two-sample MR analyses. We utilized data from IMI DIRECT and UKBB with key input variables including anthropometry, lifestyle, measures of glycemia, insulin and beta-cell function, MRI-derived abdominal and liver fat content, blood lipids, liver enzymes, and several other biochemistry assays.

Before learning the structural pathways by BN, we undertook cluster analyses of the variables in both IMI DIRECT and UKBB datasets, as it was anticipated that the clustered variables would also be linked through the BN approach. The left and right panels of Figure 4.6 illustrate the heatmap cluster among the variables of IMI DIRECT combined cohorts and the UKBB, respectively. Heatmap plots from both datasets suggest abdominal fat content, BMI, waist, and TG as a potential cluster node, with different measures of insulin dynamics only present in the IMI DIRECT cluster.

BN analyses were then undertaken using the score-based approach for building the structure of the network, since they often perform better with small sample sizes [126]. The structure scores were obtained using the BIC method, and the heuristic search towards the highest-score structure was done through the hill-climbing technique [146]. To fit a Gaus-

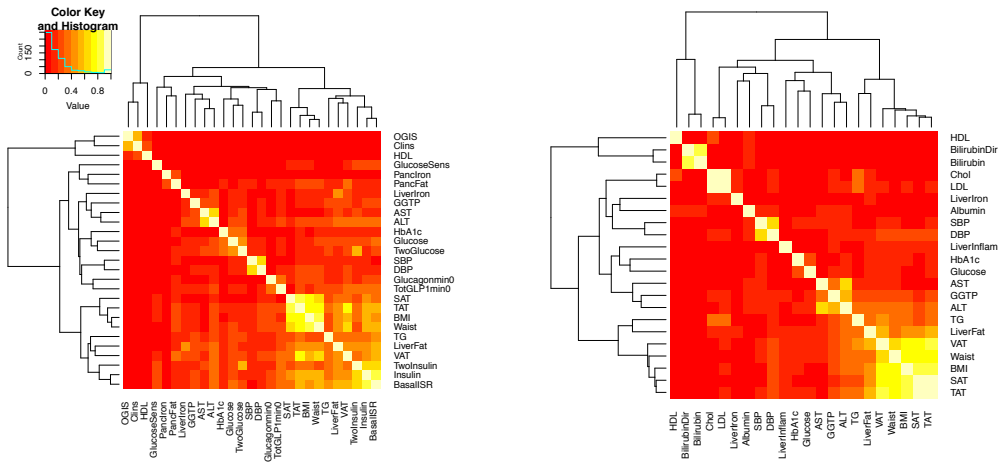

Figure 4.6: The heatmap cluster among the variables of IMI DIRECT combined cohorts and the UK Biobank, presented in the right and left panels, respectively (data are inverse normal transformed).
ALT, alanine transaminase; AST, aspartate transaminase; BasalISR, insulin secretion at the beginning of the oral glucose tolerance test/ mixed-meal tolerance test; BilrubinDir, Direct bilirubin; BMI, body mass index; Chol, Cholesterol; Clins, mean insulin clearance during the oral glucose tolerance test/mixed meal tolerance test, calculated as (mean insulin secretion)/(mean insulin concentration); DBP, mean diastolic blood pressure; 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; Insulin, fasting insulin from venous plasma samples; LDL, fasting low-density lipoprotein cholesterol; LiverInflam, liver inflammation factor; OGIS, oral glucose insulin sensitivity index according to the method of Mari et al.; PancFat, pancreas fat; PancIron, pancreas iron; SAT, subcutaneous adipose tissue; SBP, mean systolic blood pressure; TAT, total adipose tissue; TG, fasting triglycerides; 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; VAT, visceral adipose tissue.

sian BN, we utilized the inverse normalized numeric variables, where the parameters were determined based on their maximum likelihood estimate. The local distribution of each node could then be defined as a linear regression model because a Gaussian BN was used. Our main aim was to study the determinants of elevated liver fat content; as such, we, specifically, focused on the linear regression results of the liver fat node. Table 4.1 summarizes the associations of 1 standard deviation (SD) of changes in liver fat's upstream parental variables with liver fat, discovered in the IMI DIRECT (n=1264), and UKBB (n=3667) BNs.

To determine the most stable BN structure, model averaging was undertaken: data was first bootstrapped and a BN was built for each sample. Thereafter, an averaged model was built with the edges' strengths and directional probabilities estimated from the frequency of the observed arcs(edges) across all the bootstrapped BNs. The left and right panels of Figure 4.7 display the averaged BN of IMI DIRECT and UKBB, respectively, restricted to the strength and directional probabilities of ≥ 0.8. With a focus on liver fat, we also

Table 4.1: The linear regression associations of 1 standard deviation (SD) of changes in liver fat's upstream parental variables with liver fat, discovered in the IMI DIRECT (n=1264) and UKBB (n=3667) BNs. SE: standard error.

	β	βSE	*p-value*
IMI DIRECT			
Intercept	2.955e-06	2.456e-02	0.999904
Visceral adipose tissue, %	1.910e-01	3.185e-02	2.74e-09 ***
BMI, kg/m^2	1.258e-01	3.085e-02	4.89e-05 ***
2-hour insulin, pmol/L	2.252e-01	4.415e-02	3.98e-07 ***
Basal insulin secretion, pmol min^-1m^-2	2.910e-01	3.670e-02	5.51e-15 ***
Insulin sensitivity (2-hour OGIS), ml min^-1m^-2	1.959e-01	5.816e-02	0.000783 ***
UK Biobank			
Intercept	-1.66E-01	8.90E-03	< 2e-16 ***
Visceral adipose tissue, %	8.10E-01	1.75E-02	< 2e-16 ***
Albumin, g/L	1.11E-01	1.49E-02	1.47e-13 ***
Fasting glucose, mmol/L	1.06E-01	3.25E-02	0.00115 **
Glycated hemoglobin (HbA1c), mmol/mol	3.97E-02	1.29E-02	0.00201 **

considered its Markov blanket that includes the nodes with adequate information to stand as a separate BN (parents, children, and other parents of those children). The Markov blanket of the liver fat node, as highlighted in green in Figure 4.7, include visceral fat and basal insulin secretion rate as parents; ALT, liver iron and GGTP as children; OGIS, total concentration of fasting total GLP-1, HbA1c and subcutaneous fat as other parents of children, in the IMI DIRECT BN (left panel). Similarly, liver fat's Markov blanket for the UKBB BN includes visceral fat as a parent, HDL, ALT and TG as children, Waist and GGTP as other parents of children (right panel in Figure 4.7).

Furthermore, separate BN analyses were conducted in diabetes and non-diabetes cohorts of the IMI DIRECT study, with the goal of identifying the similar and different causal patterns in the presence and absence of T2D. The liver fat's Markov blanket includes basal insulin secretion rate, liver iron, BMI, visceral and trunk abdominal fat (obesity-related variables) in the non-diabetes cohort , whereas, in the diabetes cohort, it includes more glycemic measures and liver enzyme variables; ALT, GGTP, AST, glucose sensitivity, HbA1c, fasting glucose, basal insulin secretion rate and liver iron

In conjunction with BN analysis, we conducted a series of bidirectional two-sample MR analysis to study the reciprocal relationships between the BN variables. We utilized the latest summary statistics from both the GWAS catalog (30) and MR-CIEU (31), restricted to genomes of European ancestry and prioritized based on the latest release and the sample size. Genetic variants as IVs were selected at GWAS-threshold p-value $< 5 \times 10^{-8}$, and when needed proxies with linkage disequilibrium at $r^2 \geq 0.8$ were used. The IVW method for multiple IVs and Wald ratio method for a single IV were deployed to run the MR

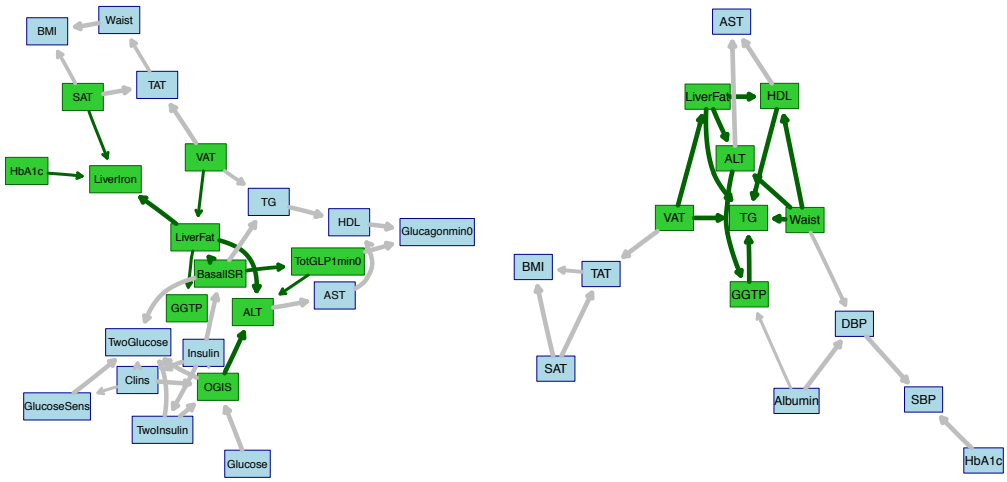

Figure 4.7: Averaged Bayesian network of the bootstrapped samples, with only the arcs with strength and directional probabilities ≥ 0.8, among the variables of the IMI DIRECT combined cohorts and the UK Biobank, presented in the left and right panels, respectively (data are inverse normal transformed). The Markov blanket of the liver fat node, which includes the nodes with adequate information to stand as a separate BN (parents, children, and other parents of those children) is highlighted in green.
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); DBP, mean diastolic blood pressure; 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; Insulin, fasting insulin from venous plasma samples; OGIS, oral glucose insulin sensitivity index according to the method of Mari et al.; PancFat, pancreas fat; PancIron, pancreas iron; SAT, subcutaneous adipose tissue; SBP, mean systolic blood pressure; TAT, total adipose tissue; TG, fasting triglycerides; 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; VAT, visceral adipose tissue.

analyses. The MR results were considered statistically significant if the causal association amongst IVW, weighted median (WM), and MR-Egger were directionally concordant and nominally statistically significant (p<0.05), and that the IVW (main method) passed the Bonferroni corrected threshold (p-value 0.05/23 = $2.2e^{-3}$) with no statistical evidence of heterogeneity and/or pleiotropy. Following these criteria, the conducted MR results suggested several nominal directional associations between hepatic biomarkers, glycemic, and adiposity measures (summarized in Table 4.2).

In summary, our analyses identified basal insulin secretion rate and visceral fat as the two key drivers in liver fat accumulation. In addition, the sensitivity analysis on the IMI DIRECT cohorts identified a network mostly dominated by dysglycemia in the diabetes cohort, whereas, in the non-diabetes cohort it was mainly controlled by excess adiposity. Moreover, the BN findings were mostly validated by the conducted MR analyses, where the genetic instruments were available to test for the directional associations.

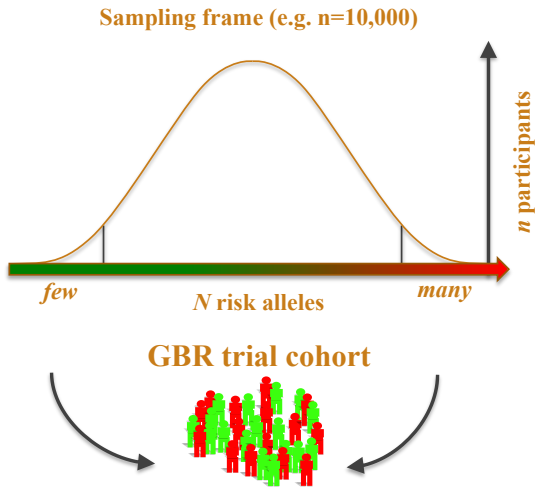

Figure 4.8: Genotype-based recall (GBR) randomized controlled trial approach, in which the genetic burden of individuals is used in recruiting two groups of participants: one group with low GRS and the other with high GRS. By courtesy of Prof. Paul W. Franks.

4 Paper IV

Many complex diseases are assumed to result from gene × environment interactions. Although observational studies have proven effective for the discovery of interactions, those findings require validation in RCTs, if they are to be informative in clinical practice. However, the validation of gene × environment interaction results from epidemiological studies in RCTs is usually hampered by insufficient statistical power. In paper IV, we hypothesized that using the GBR approach to design the clinical trials compared to RCTs can help improve statistical power when testing gene-environment/treatment interactions. As shown in Figure 4.8, GBR refers to the design of studies in which the genetic burden of individuals is used in recruiting two groups of participants: one group with low GRS and the other with high GRS.

To test our hypothesis, we conducted simulation-based power calculations focusing on two different regression models with parameters taken from two different DPP sub-studies:

- power calculation for the interaction between GRS (comprised of 32 risk SNPs) and intensive lifestyle intervention (ILI) in association with 1-year small LDL particle levels as the quantitative outcome variable, in a multivariable linear regression model (per allele interaction effect was estimated as 0.03 nmol/l) [105].
- power calculation for the interaction between variant rs8065082 at SLC47A1 and metformin in association with time to diabetes incidence as the outcome, in a Cox

proportional hazards regression model (per allele interaction effect was estimated as HR = 0.68) [106].

We calculated statistical power given different sample sizes for a GBR trial and compared it with the conventional RCT setting. Mathematically, the multivariable linear regression model can be expressed as:

$$Y = \beta_0 + \sum_{i=1}^{m}(\beta_i \times X_i) + \epsilon, \epsilon \sim N(0, \sigma^2), \tag{4.1}$$

where Y is the outcome/dependent variable described by a combination of predictors/ independent variables (X_i), $\beta 0$ is the intercept and βi denotes the effect size per each predictor, and ϵ is the unexplained error, normally distributed with mean=0 and variance σ^2, measuring the difference between the observed and the predicted Y. Similarly, we modeled the1-year small LDL particle levels with $GRS \times ILI$ interactions as $Y = \beta_0 + (\beta_{ILI} \times X_{ILI}) + (\beta_{SNPs} \times X_{SNPs}) + (\beta_{Interaction} \times X_{SNPs \times ILI}) + \epsilon$. Both *Zero/One* and *standard error* methods (read more on Chapter 3 - Section 3) were utilized in the power calculation of the linear regression model (yielding the same power estimates).

Cox proportional hazards regression models as a class of survival models can be mathematically expressed through the hazard function as follows:

$$h(t|x) = h_0(t) + \times exp(\beta x), \tag{4.2}$$

where hazard rate h(t|x) is the probability that an event will occur at time t given the risk factors and independent variables (x), β is a vector of effect sizes and $h_0(t)$ is the baseline hazard with x equal to zero. For the simulation purpose, the hazard function needs to be converted to its survival time equivalent, which is the time that passes before an event occurs to an individual. We took the Bender et al. [147] approach for simulating the survival times in our analyses, considering the three different baseline hazard distribution models (i.e. Weibull, Gompertz, exponential, see the Equations 4.3-4.5).

$$Weibull: T = -\frac{log(U)}{\lambda \times exp(\beta x)}, \lambda = \frac{1}{\mu} \tag{4.3}$$

$$Exponential: T = -[\frac{log(U)}{\lambda \times exp(\beta x)}]^{\frac{1}{\nu}}, \nu = \frac{1}{\sigma}, \lambda = exp(-\mu\sigma) \tag{4.4}$$

$$Gompertz: T = \frac{1}{\alpha} log[1 - \frac{\alpha log(U)}{\lambda} exp(\beta x)], \alpha = \frac{\pi}{\sqrt{6\sigma}}, \lambda = \alpha exp(-\gamma - \alpha\mu), \tag{4.5}$$

where T represents the survival times, described by a vector of independent variables (x) with β as a vector of their effect sizes, U is the uniform distribution on the interval [0,1], μ is the mean of time values, σ is the standard deviation of the time values, γ is the Euler's constant ≈ 0.5772 and $\pi \approx 3.14159$. Similarly, we modeled the time to diabetes incidence with $rs8065082 \times metformin$ interaction with βx values defined as $(\beta_{metformin} \times X_{metformin}) + (\beta_{rs8065082} \times X_{rs8065082}) + (\beta_{Interaction} \times X_{rs8065082 \times metformin})$ and using the Gompertz model as it yielded in better fitting survival times according to DPP data. Accordingly, the statistical power for the Cox regression models was estimated through the *Zero/One* method.

Through our simulation-based power calculation, both studies showed higher statistical power in GBR setting in comparison with the conventional random sampling. To achieve the 80% statistical power in testing the interaction between the rs8065082 variant and metformin on diabetes incidence, $\sim$ 1700 individuals were required through the GBR approach, whereas, around $\sim$ 3000 individuals were required in the conventional setting. Similarly, to detect the interaction between GRS (comprised of 32 risk SNPs) and ILI, ($\sim$ 400) individuals were required in the GBR approach, whereas, around five times ($\sim$ 1900) individuals were required when participants were randomly selected without considering their genetic background.

Considering that the statistical power to test the gene-environment/treatment interactions depends on factors such as the allele frequency of the genetic variants, the size of their interaction effects, the initial sampling frame, and the error level in the outcome measurement, we defined different scenarios to study the effect of these factors in GBR power. Almost all scenarios showed an increased statistical power when designing the trial via the GBR approach in comparison with the conventional approach with a random sampling of the participants.

Figure 4.9 illustrates the effect of the initial sampling frame (n= 5000, 10000, 50000, 100000 and 500000) on the statistical power of the interaction effect in both GBR and conventional approaches for the Cox regression model (as defined above). As it shows, increasing the initial sampling frame results in power increase when participants are recruited from the extremes of the GRS distribution, for example, to reach the 80% power, n$\sim$ 300 individuals are required with an initial sampling frame of 500,000 individuals, whereas, n$\sim$ 900 individuals are required with an initial sampling frame of 10,000 individuals). However, it has no effect on the interaction power when participants are randomly selected in the conventional setting with no emphasis on their GRS.

In summary, using parameters from DPP's published studies and also several diverse scenarios for time-to-event and quantitative outcome variables, our simulation-based power calculations showed that the GBR approach is, in most cases, a more powerful approach in comparision with the conventional methods when designing the clinical trials to test the gene-treatment/environment interaction effect.

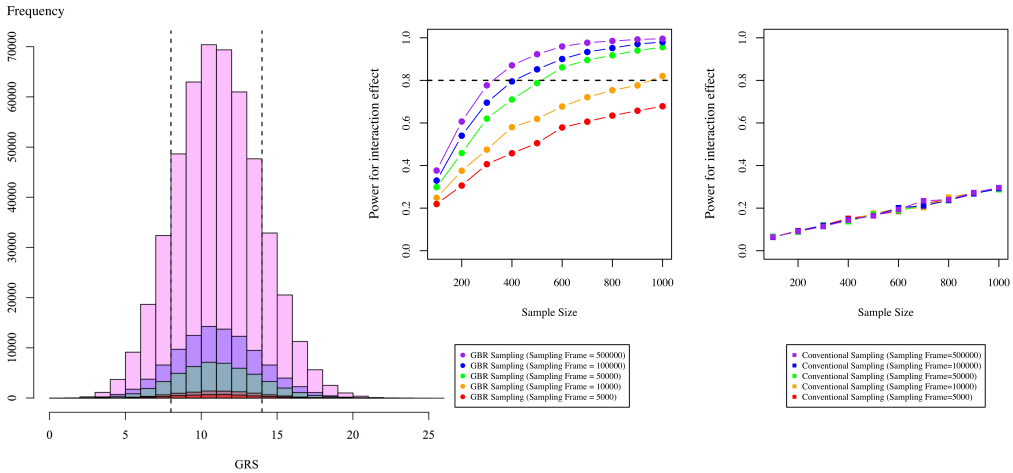

Figure 4.9: The effect of initial sampling frame (n= 5000, 10000, 50000, 100000 and 500000) in genotype-based recall (GBR) and conventional random sampling approaches on statistical power of gene-environment/treatment interaction in a Cox proportional hazards regression model.
Left panel shows the genetic risk score (GRS) histogram for different initial sampling frames (n= 5000, 10000, 50000, 100000 and 500000) of 20 random genetic variants.

Ironically, the last paper of this thesis is the first project of my PhD studies. Indeed, I started working on this project as my master thesis and being less experienced, I received the most supervision of all papers during this project. I did most of the statistical analysis, simulation-based power calculations and I also developed a web interface for power calculation in designing clinical trials to validate the gene- environment/treatment interaction hypotheses from epidemiological studies. It covers both linear and Cox regression models with user-defined SNPs or GRS parameters (`https://gbr-power.crc.med.lu.se/`).

5 Overall summary and conclusions

The ultimate objective of all four papers included in this thesis is improving prediction and causal inference for diseases of metabolic dyshomeostasis. I focused on NAFLD and T2D, two pervasive, complex, and interrelated metabolic abnormalities, that cause serious health complications and suffering for the affected person, and impose substantial economic and clinical burdens for many societies around the world. These conditions are strongly related, as described in thousands of published studies. However, the mechanisms underlying these the interplay between these diseases is poorly understood. Given that understanding the etiology of these diseases might aid their *diagnosis*, *prevention*, and *treatment*, we set about investigating fatty liver and glucose dysregulation, the results of which are described in the papers included in this PhD thesis. Specifically, we attempted to address these goals as follows:

- **Diagnosis:** In **paper II**, a series of machine learning-based prediction models for the diagnosis of fatty liver were developed, using different combinations of complex input data including clinical and multi-omic datasets. The diagnostic models could be potentially used for screening at-risk populations for the diagnosis of NAFLD. Our findings highlight beta-cell function and insulin sensitivity as the most informative predictors in the developed diagnostic models.

- **Prevention:** While it is often not necessary to understand the causal basis of associations that are useful for prediction and diagnosis, intervening to prevent the diseases requires an understanding of causal pathways and mechanisms of action. To address this, **paper III** used causal inference methods to examine a range of putative causal associations underlying the development of fatty liver; the results show that basal insulin secretion and visceral fat accumulation are key drivers. This analytical method was hypothesis-free and data-driven; however, within the same context, **paper I** was built on testing a defined hypothesis (known as the twin-cycle hypothesis), where the effect of physical activity on glucose regulation was studied by testing the mediated pathways. The results showed mediations through basal insulin secretion rate, insulin sensitivity, and liver fat accumulation.

- **Treatment:** RCTs are often used to assess the efficacy of a given treatment, but they can be extremely expensive and time-consuming. Thus, many stakeholders are eager for methods that help optimize the design of clinical trials to reduce costs and time to completion. In **paper IV**, GBR clinical trials, where the participants are selected based on their genetic characteristics, were simulated and compared with RCTs in terms of their statistical power and the required sample sizes. The analysis showed that GBR trials are, under many scenarios, more powerful than conventional trials for testing gene-treatment interactions.

Through IMI DIRECT and UK Biobank, we had access to multi-omics data, extensive environmental exposures, and biological intermediates, within which our projects were conducted. The fundamental challenge was in utilizing these diverse datasets to understand the aforementioned conditions, which we attempted to overcome through cutting-edge machine learning, statistical, and bioinformatics methods.

6 Future perspectives

Extension of the projects presented in this thesis can be considered in a few aspects:

- As in paper I, with the help of SEM models, we can test other proposed hypotheses when data allows. One potential hypothesis is to test the genetically-driven NAFLD

and metabolic NAFLD (proposed by Liu et al., read more on Chapter 1 - Section 2) in causal association with T2D and obesity through SEM models to assess the model's goodness of fit. Considering the availability of data, such analysis is possible in both IMI DIRECT and UKBB that we have access to.

- One of our limitations was that working with SEM models required complete case data, which dropped our sample sizes significantly. In paper I, this was mainly due to the physical activity data in the IMI DIRECT cohorts, and to retain the sample size, one approach to be considered as a future analysis is imputation.

- In paper II, we mainly presented our results by random forest analyses that as a typical machine learning tool, precise effect size values per variable, or a formula for the developed models could not be reported. One approach to mitigate this and unlock the hidden box of random forest can be undertaking the decision tree approach to show how the branches and nodes are defined for different cut-off values per variable. The intention is not for reporting a precise effect estimate, but only to get an idea of how a tree with all the variables would look like in one of the trees of a random forest.

- Another approach to help with simplifying the developed models in paper II, could be limiting the omics models to only the very top subset of the LASSO predictors. These models could then be implemented in the web interface for the researchers' access.

- One of the limitations of paper II was categorizing the continuous liver fat data into two groups as we did not have enough power to develop the prediction models on the continuous data. Indeed we did attempt to develop our models in 4 clinical categories of liver fat percentage (<2%, 2%-5%, >5%-16% and >16%), yet we did not reach the desired prediction power. As a continuum of this project and further study these categories, one can consider clustering analysis (such as heatmap) at each of the liver fat starta to find out the key predictors of each.

- Being completed during the last few months of my PhD studies, Paper III can be further improved prior to publication. In paper III, basal insulin secretion rate and visceral fat appeared to be key determinants of liver fat. One possible addition to the current work could be studying these two factors at different stages of the liver fat progression (<2%, 2%-5%, >5%-16% and >16%).

- In paper III, we only considered the clinical variables in the network and this work can be extended to different layers of omics. Indeed the selected omic features from paper II can be exploited here and further studied in network analyses to find out the interacting factors.

- As another research question, the lean participants with fatty liver could be studied genetically and metabolically and be compared with those non-lean with fatty liver. This can be investigated through both IMI DIRECT and UKBB datasets.

- Through our UKBB application, we did not have access to the full abdominal MRI data ($n \sim 40,000$) and we only had access to $\sim 4,000$ imaging data. Here, the developed models in paper II can be exploited to predict liver fat percentage to be used in paper III and once we have access to the full data, we can validate the models. If we could pass the validation, we can even further consider estimating the liver fat data for the whole UKBB participants ($\sim 500,000$) for those we have genetic, anthropometric and most of the blood biomarkers available.

- Estimating the liver fat data through the developed models in paper II can also be considered for the IMI DIRECT longitudinal data, where we lack liver fat data but we have the key glycemic and anthropometric data measured at different time points. This allows us to further study the liver fat progression at different stages of the condition.

- Logistic regression models are still one of the commonly used regression models in epidemiological studies for studying the binary outcomes. In our simulation analyses of paper IV, we only considered the Cox proportional hazard regression model, as it contains more information (with a time-to-event continuous outcome) than a logistic regression model (with a binary event as outcome). However, considering the popularity of the logistic regression models, this can be added to our online power calculator to study gene-treatment/environment interaction in GBR and conventional settings.

- One of the assumption we made in paper IV, was that all the interacting SNPs were utilized to construct the GRS, however, this is not always the case. This can be considered as another addition to the web power calculator by differentiating the SNPs that are only interacting than the rest. Weighted GRS, where each SNP is multiplied by its effect size prior to the overall summing can also be considered as an extension to the current power calculator.

Table 4.2: The subset of two-sample Mendelian randomization exposure-outcome associations that were considered statistically significant based on the following criteria:
the causal association amongst IVW, weighted median (WM), and MR-Egger were directionally concordant and nominally statistically significant (p<0.05), and that the IVW (main method) passed the Bonferroni corrected threshold (p-value 0.05/23 = 2.2e-03) with no statistical evidence of heterogeneity and/or pleiotropy.
For those with a single genetic instrument (**), the Wald ratio method was deployed to run the MR analyses. Units are 1 standard deviation (SD) unless specified. ALT, alanine transaminase; AST, aspartate transaminase; BMI, body mass index; DBP, mean diastolic blood pressure; GGTP, gamma-glutamyl transpeptidase; HbA1c, hemoglobin A1C; HDL, fasting high-density lipoprotein cholesterol; SBP, mean systolic blood pressure; TG, fasting triglycerides.

Author, year	**Exposure**	**Outcome**	β	*p-value*
Prins BP, 2017	GGTP	AST	0.26	$5.26e^{-42}$
	GGTP	ALT	0.33	$4.22e^{-61}$
	AST	GGTP	0.26	$2.92e^{-11}$
	AST	ALT	0.52	$3.23e^{-67}$
	Serum albumin level	**DBP	-0.06	$1.27e^{-10}$
	Serum albumin level	**Total Bilirubin	0.05	$1.34e^{-05}$
	Serum albumin level	**GGTP	0.04	$2.70e^{-03}$
	Serum albumin level	**AST	0.05	$6.33e^{-05}$
Manning AK,2012	Fasting glucose (mmol/L)	HbA1C	0.37	$5.93e^{-21}$
Warren HR, 2017	SBP	DBP	0.66	$2.12e^{-102}$
	DBP	Waist	0.08	$1.52e^{-04}$
Suhre K, 2017	Glucagon	**DBP	-0.04	$6.17e^{-04}$
Locke, 2015	BMI (kg/m2)	HDL	-0.34	$1.80e^{-73}$
		HbA1c	0.20	$2.54e^{-42}$
		Waist	0.77	$0.00e^{+00}$
		Fasting insulin	0.17	$4.94e^{-18}$
		ALT	0.25	$7.62e^{-73}$
Shungin D, 2015	Waist	HDL	-0.42	$2.62e^{-69}$
		GGTP	0.27	$6.16e^{-45}$
		AST	0.08	$7.20e^{-07}$
		SBP	0.14	$9.81e^{-16}$
		DBP	0.23	$4.66e^{-39}$
		BMI	1.12	$0.00e^{+00}$
		2-hr glucose	0.08	$1.52e^{-04}$
		HbA1c	0.26	$1.82e^{-28}$
		Fasting insulin	0.15	$6.17e^{-06}$
		ALT	0.28	$2.45e^{-36}$
Willer CJ, 2013	HDL	TG	-0.82	$8.01e^{-11}$
	TG	GGTP	0.08	$1.01e^{-03}$

References

[1] Nadhipuram V Bhagavan. *Medical biochemistry*. Academic press, 2002.

[2] Michael Roden and Elisabeth Bernroider. Hepatic glucose metabolism in humans—its role in health and disease. *Best Practice & Research Clinical Endocrinology & Metabolism*, 17(3):365–383, 2003.

[3] Barry A Mizock. Alterations in carbohydrate metabolism during stress: a review of the literature. *The American journal of medicine*, 98(1):75–84, 1995.

[4] Jason Karpac and Heinrich Jasper. Metabolic homeostasis: HDACs take center stage. *Cell*, 145(4):497–499, 2011.

[5] Christopher D Byrne and Giovanni Targher. NAFLD: a multisystem disease. *Journal of hepatology*, 62(1):S47–S64, 2015.

[6] David M Nathan. Long-term complications of diabetes mellitus. *New England Journal of Medicine*, 328(23):1676–1685, 1993.

[7] Giovanni Targher and Christopher D Byrne. Nonalcoholic fatty liver disease: a novel cardiometabolic risk factor for type 2 diabetes and its complications. *The Journal of Clinical Endocrinology & Metabolism*, 98(2):483–495, 2013.

[8] Hannele Yki-Järvinen. Non-alcoholic fatty liver disease as a cause and a consequence of metabolic syndrome. *The lancet Diabetes & endocrinology*, 2(11):901–910, 2014.

[9] Zobair Younossi, Frank Tacke, Marco Arrese, Barjesh Chander Sharma, Ibrahim Mostafa, Elisabetta Bugianesi, Vincent Wai-Sun Wong, Yusuf Yilmaz, Jacob George, Jiangao Fan, et al. Global perspectives on nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. *Hepatology*, 69(6):2672–2682, 2019.

[10] Marcus E Carr. Diabetes mellitus: a hypercoagulable state. *Journal of Diabetes and its Complications*, 15(1):44–54, 2001.

[11] Amalia Gastaldelli, Michaela Kozakova, Kurt Højlund, Allan Flyvbjerg, Angela Favuzzi, Asimina Mitrakou, Beverley Balkau, and RISC investigators. Fatty liver is associated with insulin resistance, risk of coronary heart disease, and early atherosclerosis in a large European population. *Hepatology*, 49(5):1537–1544, 2009.

[12] Anna Kotronen, Leena Juurinen, Mirja Tiikkainen, Satu Vehkavaara, and Hannele Yki-Järvinen. Increased liver fat, impaired insulin clearance, and hepatic and adipose tissue insulin resistance in type 2 diabetes. *Gastroenterology*, 135(1):122–130, 2008.

[13] Mala Dharmalingam and P Ganavi Yamasandhi. Nonalcoholic fatty liver disease and type 2 diabetes mellitus. *Indian journal of endocrinology and metabolism*, 22(3):421, 2018.

[14] Soumya Murag, Aijaz Ahmed, and Donghee Kim. Recent Epidemiology of Nonalcoholic Fatty Liver Disease. *Gut and liver*.

[15] Paul W Franks, Ewan Pearson, and Jose C Florez. Gene-environment and gene-treatment interactions in type 2 diabetes: progress, pitfalls, and prospects. *Diabetes care*, 36(5):1413–1421, 2013.

[16] Paul W Franks and Guillaume Paré. Putting the genome in context: gene-environment interactions in type 2 diabetes. *Current diabetes reports*, 16(7):1–14, 2016.

[17] Alice Emma Taliento, Marcello Dallio, Alessandro Federico, Daniele Prati, and Luca Valenti. Novel insights into the genetic landscape of nonalcoholic fatty liver disease. *International journal of environmental research and public health*, 16(15):2755, 2019.

[18] Quentin M Anstee and Christopher P Day. The genetics of NAFLD. *Nature reviews Gastroenterology & hepatology*, 10(11):645–655, 2013.

[19] Vlad Ratziu, Mary Rinella, Ulrich Beuers, Rohit Loomba, Quentin M Anstee, Stephen Harrison, Sven Francque, Arun Sanyal, Philip N Newsome, and Zobair Younossi. The times they are a-changin'(for NAFLD as well). *Journal of Hepatology*, 2020.

[20] Diabetes Mellitus. Diagnosis and classification of diabetes mellitus. *Diabetes care*, 28 (S37):S5–S10, 2005.

[21] Jose C Florez. Mining the genome for therapeutic targets. *Diabetes*, 66(7):1770–1778, 2017.

[22] Jose C Florez, Miriam S Udler, and Robert L Hanson. Genetics of type 2 diabetes. In *Diabetes in America*. National Institutes of Health, Bethesda, 2016.

[23] Indu G Poornima, Pratik Parikh, and Richard P Shannon. Diabetic cardiomyopathy: the search for a unifying hypothesis. *Circulation research*, 98(5):596–605, 2006.

[24] Laura Wyness. Understanding the role of diet in type 2 diabetes prevention. *British journal of community nursing*, 14(9):374–379, 2009.

[25] Sheri R Colberg, Ronald J Sigal, Bo Fernhall, Judith G Regensteiner, Bryan J Blissmer, Richard R Rubin, Lisa Chasan-Taber, Ann L Albright, and Barry Braun. Exercise and type 2 diabetes: the American College of Sports Medicine and the American Diabetes Association: joint position statement. *Diabetes care*, 33(12):e147–e167, 2010.

[26] Alaitz Poveda, Robert W Koivula, Shafqat Ahmad, Inês Barroso, Göran Hallmans, Ingegerd Johansson, Frida Renström, and Paul W Franks. Innate biology versus lifestyle behaviour in the aetiology of obesity and type 2 diabetes: the GLACIER Study. *Diabetologia*, 59(3):462–471, 2016.

[27] E Lahjibi, Barbara Heude, JM Dekker, Kurt Højlund, Martine Laville, John Nolan, J-M Oppert, Beverley Balkau, RISC Study Group, et al. Impact of objectively measured sedentary behaviour on changes in insulin resistance and secretion over 3 years in the RISC study: interaction with weight gain. *Diabetes & metabolism*, 39(3): 217–225, 2013.

[28] Roy Taylor. Remission of type 2 diabetes by weight loss in a non-white population. *The lancet. Diabetes & Endocrinology*, 8(6):458–459, 2020.

[29] William H Herman. Diabetes epidemiology: guiding clinical and public health practice: the Kelly West Award Lecture, 2006. *Diabetes Care*, 30(7):1912–1919, 2007.

[30] Sherif RZ Abdel-Misih and Mark Bloomston. Liver anatomy. *Surgical Clinics*, 90 (4):643–653, 2010.

[31] Amedeo Lonardo, Simona Leoni, Khalid A Alswat, and Yasser Fouad. History of Nonalcoholic Fatty Liver Disease. *International Journal of Molecular Sciences*, 21(16): 5888, 2020.

[32] Amélio F Godoy-Matos, Wellington S Silva Júnior, and Cynthia M Valerio. NAFLD as a continuum: from obesity to metabolic syndrome and diabetes. *Diabetology & Metabolic Syndrome*, 12(1):1–20, 2020.

[33] Siôn A Parry and Leanne Hodson. Managing NAFLD in Type 2 Diabetes: The Effect of Lifestyle Interventions, a Narrative Review. *Advances in Therapy*, pages 1–26, 2020.

[34] Amedeo Lonardo, Simonetta Lugari, Stefano Ballestri, Fabio Nascimbeni, Enrica Baldelli, and Mauro Maurantonio. A round trip from nonalcoholic fatty liver disease to diabetes: molecular targets to the rescue? *Acta diabetologica*, 56(4):385–396, 2019.

[35] A Katrina Loomis, Shaum Kabadi, David Preiss, Craig Hyde, Vinicius Bonato, Matthew St. Louis, Jigar Desai, Jason MR Gill, Paul Welsh, Dawn Waterworth, et al.

Body mass index and risk of nonalcoholic fatty liver disease: two electronic health record prospective studies. *The Journal of Clinical Endocrinology & Metabolism*, 101 (3):945–952, 2016.

[36] Naveed Sattar and Jason MR Gill. Type 2 diabetes as a disease of ectopic fat? *BMC medicine*, 12(1):1–6, 2014.

[37] Zhipeng Liu, Yang Zhang, Sarah Graham, Xiaokun Wang, Defeng Cai, Menghao Huang, Roger Pique-Regi, Xiaocheng Charlie Dong, Y Eugene Chen, Cristen Willer, et al. Causal relationships between NAFLD, T2D and obesity have implications for disease subphenotyping. *Journal of Hepatology*, 2020.

[38] Roy Taylor. Type 2 diabetes: etiology and reversibility. *Diabetes care*, 36(4):1047–1055, 2013.

[39] R Taylor. Pathogenesis of type 2 diabetes: tracing the reverse route from cure to cause. *Diabetologia*, 51(10):1781–1789, 2008.

[40] Elena Buzzetti, Massimo Pinzani, and Emmanuel A Tsochatzis. The multiple-hit pathogenesis of non-alcoholic fatty liver disease (NAFLD). *Metabolism*, 65(8):1038–1048, 2016.

[41] Lutgarda Bozzetto, Anna Prinster, Marcello Mancini, Rosalba Giacco, Claudia De Natale, Marco Salvatore, Gabriele Riccardi, Angela A Rivellese, and Giovanni Annuzzi. Liver fat in obesity: role of type 2 diabetes mellitus and adipose tissue distribution. *European journal of clinical investigation*, 41(1):39–44, 2011.

[42] Mohammed Eslam, Arun J Sanyal, Jacob George, Arun Sanyal, Brent Neuschwander-Tetri, Claudio Tiribelli, David E Kleiner, Elizabeth Brunt, Elisabetta Bugianesi, Hannele Yki-Järvinen, et al. MAFLD: a consensus-driven proposed nomenclature for metabolic associated fatty liver disease. *Gastroenterology*, 158(7):1999–2014, 2020.

[43] Christopher P Day and Oliver FW James. Steatohepatitis: a tale of two "hits"?, 1998.

[44] Andrew P Morris, Benjamin F Voight, Tanya M Teslovich, Teresa Ferreira, Ayellet V Segre, Valgerdur Steinthorsdottir, Rona J Strawbridge, Hassan Khan, Harald Grallert, Anubha Mahajan, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*, 44(9):981, 2012.

[45] Robert A Scott, Vasiliki Lagou, Ryan P Welch, Eleanor Wheeler, May E Montasser, Reedik Mägi, Rona J Strawbridge, Emil Rehnberg, Stefan Gustafsson, Stavroula Kanoni, et al. Large-scale association analyses identify new loci influencing glycemic traits

and provide insight into the underlying biological pathways. *Nature genetics*, 44(9): 991–1005, 2012.

[46] Quentin M Anstee, Rebecca Darlay, Simon Cockell, Marica Meroni, Olivier Govaere, Dina Tiniakos, Alastair D Burt, Pierre Bedossa, Jeremy Palmer, Yang-Lin Liu, et al. Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a histologically-characterised cohort. *Journal of Hepatology*, 2020.

[47] Kyle J Gaulton, Teresa Ferreira, Yeji Lee, Anne Raimondo, Reedik Mägi, Michael E Reschen, Anubha Mahajan, Adam Locke, N William Rayner, Neil Robertson, et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nature genetics*, 47(12):1415–1425, 2015.

[48] Mohammed Eslam, Luca Valenti, and Stefano Romeo. Genetics and epigenetics of NAFLD and NASH: clinical impact. *Journal of hepatology*, 68(2):268–279, 2018.

[49] Stefano Romeo, Julia Kozlitina, Chao Xing, Alexander Pertsemlidis, David Cox, Len A Pennacchio, Eric Boerwinkle, Jonathan C Cohen, and Helen H Hobbs. Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nature genetics*, 40(12):1461–1465, 2008.

[50] Yuya Seko, Kanji Yamaguchi, and Yoshito Itoh. The genetic backgrounds in nonalcoholic fatty liver disease. *Clinical journal of gastroenterology*, 11(2):97–102, 2018.

[51] Piero Pingitore, Carlo Pirazzi, Rosellina M Mancina, Benedetta M Motta, Cesare Indiveri, Arturo Pujia, Tiziana Montalcini, Kristina Hedfalk, and Stefano Romeo. Recombinant PNPLA3 protein shows triglyceride hydrolase activity and its I148M mutation results in loss of function. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1841(4):574–580, 2014.

[52] Xiaocheng Charlie Dong. PNPLA3—A Potential Therapeutic Target for Personalized Treatment of Chronic Liver Disease. *Frontiers in Medicine*, 6, 2019.

[53] Jinsheng Yu, Sharon Marsh, Junbo Hu, Wenke Feng, and Chaodong Wu. The pathogenesis of nonalcoholic fatty liver disease: interplay between diet, gut microbiota, and genetic background. *Gastroenterology research and practice*, 2016, 2016.

[54] Jeffrey D Browning, Lidia S Szczepaniak, Robert Dobbins, Pamela Nuremberg, Jay D Horton, Jonathan C Cohen, Scott M Grundy, and Helen H Hobbs. Prevalence of hepatic steatosis in an urban population in the United States: impact of ethnicity. *Hepatology*, 40(6):1387–1395, 2004.

[55] Sherif Saadeh, Zobair M Younossi, Erick M Remer, Terry Gramlich, Janus P Ong, Maja Hurley, Kevin D Mullen, James N Cooper, and Michael J Sheridan. The utility

of radiological imaging in nonalcoholic fatty liver disease. *Gastroenterology*, 123(3): 745–750, 2002.

[56] Naga Chalasani, Zobair Younossi, Joel E Lavine, Michael Charlton, Kenneth Cusi, Mary Rinella, Stephen A Harrison, Elizabeth M Brunt, and Arun J Sanyal. The diagnosis and management of nonalcoholic fatty liver disease: practice guidance from the American Association for the Study of Liver Diseases. *Hepatology*, 67(1):328–357, 2018.

[57] Laurent Castera, Mireen Friedrich-Rust, and Rohit Loomba. Noninvasive assessment of liver disease in patients with nonalcoholic fatty liver disease. *Gastroenterology*, 156(5):1264–1281, 2019.

[58] Giorgio Bedogni, Stefano Bellentani, Lucia Miglioli, Flora Masutti, Marilena Passalacqua, Anna Castiglione, and Claudio Tiribelli. The Fatty Liver Index: a simple and accurate predictor of hepatic steatosis in the general population. *BMC gastroenterology*, 6(1):33, 2006.

[59] Jeong-Hoon Lee, Donghee Kim, Hwa Jung Kim, Chang-Hoon Lee, Jong In Yang, Won Kim, Yoon Jun Kim, Jung-Hwan Yoon, Sang-Heon Cho, Myung-Whun Sung, et al. Hepatic steatosis index: a simple screening tool reflecting nonalcoholic fatty liver disease. *Digestive and Liver Disease*, 42(7):503–508, 2010.

[60] L Fedchuk, F Nascimbeni, R Pais, F Charlotte, C Housset, V Ratziu, and LIDO Study Group. Performance and limitations of steatosis biomarkers in patients with nonalcoholic fatty liver disease. *Alimentary pharmacology & therapeutics*, 40(10):1209–1222, 2014.

[61] Laurent Castera. Diagnosis of non-alcoholic fatty liver disease/non-alcoholic steatohepatitis: non-invasive tests are enough. *Liver International*, 38:67–70, 2018.

[62] Frank Hu. *Obesity epidemiology*. Oxford University Press, 2008.

[63] Amalia Gastaldelli, Kenneth Cusi, Maura Pettiti, Jean Hardies, Yoshinori Miyazaki, Rachele Berria, Emma Buzzigoli, Anna Maria Sironi, Eugenio Cersosimo, Ele Ferrannini, et al. Relationship between hepatic/visceral fat and hepatic insulin resistance in nondiabetic and type 2 diabetic subjects. *Gastroenterology*, 133(2):496–506, 2007.

[64] Carolina Ortiz-Lopez, Romina Lomonaco, Beverly Orsak, Joan Finch, Zhi Chang, Valeria G Kochunov, Jean Hardies, and Kenneth Cusi. Prevalence of prediabetes and diabetes and metabolic profile of patients with nonalcoholic fatty liver disease (NAFLD). *Diabetes care*, 35(4):873–878, 2012.

[65] Mariana Machado, Pedro Marques-Vidal, and Helena Cortez-Pinto. Hepatic histology in obese patients undergoing bariatric surgery. *Journal of hepatology*, 45(4): 600–606, 2006.

[66] Onpan Cheung, Ashwani Kapoor, Puneet Puri, Sakita Sistrun, Velimir A Luketic, Carol C Sargeant, Melissa J Contos, Mitchell L Shiffman, Richard T Stravitz, Richard K Sterling, et al. The impact of fat distribution on the severity of nonalcoholic fatty liver disease and metabolic syndrome. *Hepatology*, 46(4):1091–1100, 2007.

[67] Giovanni Targher, Lorenzo Bertolini, Roberto Padovani, Stefano Rodella, Roberto Tessari, Luciano Zenari, Christopher Day, and Guido Arcaro. Prevalence of nonalcoholic fatty liver disease and its association with cardiovascular disease among type 2 diabetic patients. *Diabetes care*, 30(5):1212–1218, 2007.

[68] Ji Cheol Bae, Eun Jung Rhee, Won Young Lee, Se Eun Park, Cheol Young Park, Ki Won Oh, Sung Woo Park, and Sun Woo Kim. Combined effect of nonalcoholic fatty liver disease and impaired fasting glucose on the development of type 2 diabetes: a 4-year retrospective longitudinal study. *Diabetes care*, 34(3):727–729, 2011.

[69] Jaana Lindström, Markku Peltonen, JG Eriksson, Pirjo Ilanne-Parikka, Sirkka Aunola, Sirkka Keinänen-Kiukaanniemi, Matti Uusitupa, Jaakko Tuomilehto, Finnish Diabetes Prevention Study (DPS, et al. Improved lifestyle and decreased diabetes risk over 13 years: long-term follow-up of the randomised Finnish Diabetes Prevention Study (DPS). *Diabetologia*, 56(2):284–293, 2013.

[70] Jaana Lindström, Pirjo Ilanne-Parikka, Markku Peltonen, Sirkka Aunola, Johan G Eriksson, Katri Hemiö, Helena Hämäläinen, Pirjo Härkönen, Sirkka Keinänen-Kiukaanniemi, Mauri Laakso, et al. Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention: follow-up of the Finnish Diabetes Prevention Study. *The Lancet*, 368(9548):1673–1679, 2006.

[71] Peter Jepsen, Søren Paaske Johnsen, MW Gillman, and Henrik Toft Sørensen. Interpretation of observational studies. *Heart*, 90(8):956–960, 2004.

[72] Paul W Franks and N Atabaki-Pasdar. Causal inference in obesity research. *Journal of internal medicine*, 281(3):222–232, 2017.

[73] Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, 2008.

[74] Shuai Yuan and Susanna C Larsson. An atlas on risk factors for type 2 diabetes: a wide-angled Mendelian randomisation study. *Diabetologia*, pages 1–13, 2020.

[75] Yongyuan Zhang, Tao Zhang, Chengqi Zhang, Fang Tang, Nvjuan Zhong, Hongkai Li, Xinhong Song, Haiyan Lin, Yanxun Liu, and Fuzhong Xue. Identification of reciprocal causality between non-alcoholic fatty liver disease and metabolic syndrome by a simplified Bayesian network in a Chinese population. *BMJ open*, 5(9), 2015.

[76] Barbara H Bardenheier, Kai McKeever Bullard, Carl J Caspersen, Yiling J Cheng, Edward W Gregg, and Linda S Geiss. A novel use of structural equation models to examine factors associated with prediabetes among adults aged 50 years and older: National Health and Nutrition Examination Survey 2001–2006. *Diabetes care*, 36 (9):2655–2662, 2013.

[77] P Dongiovanni, S Stender, A Pietrelli, RM Mancina, A Cespiati, S Petta, S Pelusi, P Pingitore, S Badiali, M Maggioni, et al. Causal relationship of hepatic fat with liver damage and insulin resistance in nonalcoholic fatty liver. *Journal of internal medicine*, 283(4):356–370, 2018.

[78] N Maneka G De Silva, Maria Carolina Borges, Aroon D Hingorani, Jorgen Engmann, Tina Shah, Xiaoshuai Zhang, Claudia Langenberg, Andrew Wong, Diana Kuh, John C Chambers, et al. Liver function and risk of type 2 diabetes: bidirectional mendelian randomization study. *Diabetes*, 68(8):1681–1691, 2019.

[79] Stefan Stender, Julia Kozlitina, Børge G Nordestgaard, Anne Tybjærg-Hansen, Helen H Hobbs, and Jonathan C Cohen. Adiposity amplifies the genetic risk of fatty liver disease conferred by multiple loci. *Nature genetics*, 49(6):842–847, 2017.

[80] Robert W Koivula, Alison Heggie, Anna Barnett, Henna Cederberg, Tue H Hansen, Anitra D Koopman, Martin Ridderstråle, Femke Rutters, Henrik Vestergaard, Ramneek Gupta, 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(6):1132–1142, 2014.

[81] Robert W Koivula, Ian M Forgie, Azra Kurbasic, Ana Viñuela, Alison Heggie, Giuseppe N Giordano, Tue H Hansen, Michelle Hudson, Anitra DM Koopman, Femke Rutters, 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(9):1601–1615, 2019.

[82] A Amer Diabet. Standards of medical care in diabetes-2011 American Diabetes Association. *Diabetes Care*, 34:S11–S61, 2011.

[83] Andrea Mari, Andrea Tura, Amalia Gastaldelli, and Ele Ferrannini. Assessing insulin secretion by modeling in multiple-meal tests: role of potentiation. *Diabetes*, 51(suppl 1):S221–S226, 2002.

[84] Tom White, Kate Westgate, Stefanie Hollidge, Michelle Venables, Patrick Olivier, Nick Wareham, and Soren Brage. Estimating energy expenditure from wrist and thigh accelerometry in free-living adults: a doubly labelled water study. *International Journal of Obesity*, 43(11):2333–2342, 2019.

[85] DA Wood, K Kotseva, S Connolly, C Jennings, A Mead, J Jones, A Holden, Dirk De Bacquer, T Collier, Gui De Backer, et al. Nurse-coordinated multidisciplinary, family-based cardiovascular disease prevention programme (EUROACTION) for patients with coronary heart disease and asymptomatic individuals at high risk of cardiovascular disease: a paired, cluster-randomised controlled trial. *The Lancet*, 371 (9629):1999–2012, 2008.

[86] Alison E Black. Critical evaluation of energy intake using the Goldberg cut-off for energy intake: basal metabolic rate. A practical guide to its calculation, use and limitations. *International journal of obesity*, 24(9):1119–1130, 2000.

[87] E Louise Thomas, JA Fitzpatrick, SJ Malik, Simon D Taylor-Robinson, and Jimmy D Bell. Whole body fat: content and distribution. *Progress in nuclear magnetic resonance spectroscopy*, 73:56–80, 2013.

[88] Declan P O'Regan, Martina F Callaghan, Marzena Wylezinska-Arridge, Julie Fitzpatrick, Rossi P Naoumova, Joseph V Hajnal, and Stephan A Schmitz. Liver fat content and T2*: simultaneous measurement by using breath-hold multiecho MR imaging at 3.0 T—feasibility. *Radiology*, 247(2):550–557, 2008.

[89] TIMOTHY G ST. PIERRE, PAUL R CLARK, and WANIDA Chua-Anusorn. Measurement and mapping of liver iron concentrations using magnetic resonance imaging. *Annals of the New York Academy of Sciences*, 1054(1):379–385, 2005.

[90] Caroline A Schneider, Wayne S Rasband, and Kevin W Eliceiri. NIH Image to ImageJ: 25 years of image analysis. *Nature methods*, 9(7):671–675, 2012.

[91] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88 (1):76–82, 2011.

[92] Santiago Marco-Sola, Michael Sammeth, Roderic Guigó, and Paolo Ribeca. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature methods*, 9 (12):1185, 2012.

[93] Kimi Drobin, Peter Nilsson, and Jochen M Schwenk. Highly multiplexed antibody suspension bead arrays for plasma protein profiling. In *The Low Molecular Weight Proteome*, pages 137–145. Springer, 2013.

[94] Erika Assarsson, Martin Lundberg, Göran Holmquist, Johan Björkesten, Stine Bucht Thorsen, Daniel Ekman, Anna Eriksson, Emma Rennel Dickens, Sandra Ohlsson, Gabriella Edfeldt, et al. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PloS one*, 9(4):e95192, 2014.

[95] Paulomi Aldo, Gregory Marusov, Danielle Svancara, James David, and Gil Mor. Simple Plex™: A Novel Multi-Analyte, Automated Microfluidic Immunoassay Platform for the Detection of Human and Mouse Cytokines and Chemokines. *American journal of reproductive immunology*, 75(6):678–693, 2016.

[96] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos med*, 12(3):e1001779, 2015.

[97] Rory Collins. What makes UK Biobank special? *The Lancet*, 9822(379):1173–1174, 2012.

[98] Robert Clarke, Martin Shipley, Sarah Lewington, Linda Youngman, Rory Collins, Michael Marmot, and Richard Peto. Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies. *American journal of epidemiology*, 150(4):341–353, 1999.

[99] Henry R Wilman, Matt Kelly, Steve Garratt, Paul M Matthews, Matteo Milanesi, Amy Herlihy, Micheal Gyngell, Stefan Neubauer, Jimmy D Bell, Rajarshi Banerjee, et al. Characterisation of liver fat in the UK Biobank cohort. *PloS one*, 12(2):e0172921, 2017.

[100] Diabetes Prevention Program Research Group et al. The Diabetes Prevention Program: baseline characteristics of the randomized cohort. *Diabetes care*, 23(11):1619, 2000.

[101] American Diabetes Association et al. The Diabetes Prevention Program. Design and methods for a clinical trial in the prevention of type 2 diabetes. *Diabetes care*, 22(4): 623–634, 1999.

[102] Diabetes Prevention Program (DPP) Research Group et al. The Diabetes Prevention Program (DPP): description of lifestyle intervention. *Diabetes care*, 25(12):2165–2171, 2002.

[103] Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *New England journal of medicine*, 346(6):393–403, 2002.

[104] DPP Research Group et al. The Diabetes Prevention Program-recruitment methods and results. *Controlled clinical trials*, 2(23):157–171, 2002.

[105] Kathleen A Jablonski, Jarred B McAteer, Paul IW de Bakker, Paul W Franks, Toni I Pollin, Robert L Hanson, Richa Saxena, Sarah Fowler, Alan R Shuldiner, William C Knowler, et al. Common variants in 40 genes assessed for diabetes incidence and response to metformin and lifestyle intervention in the diabetes prevention program. *Diabetes*, 59(10):2672–2681, 2010.

[106] Toni I Pollin, Tamara Isakova, Kathleen A Jablonski, Paul IW De Bakker, Andrew Taylor, Jarred McAteer, Qing Pan, Edward S Horton, Linda M Delahanty, David Altshuler, et al. Genetic modulation of lipid profiles following lifestyle modification or metformin treatment: the Diabetes Prevention Program. *PLoS Genet*, 8(8): e1002895, 2012.

[107] Carolyn Mair, Gada Kadoda, Martin Lefley, Keith Phalp, Chris Schofield, Martin Shepperd, and Steve Webster. An investigation of machine learning based prediction systems. *Journal of systems and software*, 53(1):23–29, 2000.

[108] Valeria Fonti and Eduard Belitser. Feature selection using lasso. *VU Amsterdam Research Paper in Business Analytics*, 30:1–25, 2017.

[109] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[110] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, 2006.

[111] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24, 2007.

[112] Cha Zhang and Yunqian Ma. *Ensemble machine learning: methods and applications*. Springer, 2012.

[113] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[114] Gil McVean. A genealogical interpretation of principal components analysis. *PLoS Genet*, 5(10):e1000686, 2009.

[115] Marc Höfler. Causal inference based on counterfactuals. *BMC medical research methodology*, 5(1):28, 2005.

[116] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

[117] Piotr Tarka. An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & quantity*, 52(1):313–354, 2018.

[118] Peter M Bentler. Comparative fit indexes in structural models. *Psychological bulletin*, 107(2):238, 1990.

[119] Ledyard R Tucker and Charles Lewis. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1):1–10, 1973.

[120] David A Kenny. Measuring model fit, 2015.

[121] Stephen G West, Aaron B Taylor, Wei Wu, et al. Model fit and model selection in structural equation modeling. *Handbook of structural equation modeling*, 1:209–231, 2012.

[122] Edward E Rigdon. CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 3(4): 369–379, 1996.

[123] Chengwei Su, Angeline Andrew, Margaret R Karagas, and Mark E Borsuk. Using bayesian networks to discover relations between genes, environment, and disease. *BioData mining*, 6(1):6, 2013.

[124] Byron Ellis and Wing Hung Wong. Learning causal bayesian network structures from experimental data. *Journal of the American Statistical Association*, 103(482):778–789, 2008.

[125] David Heckerman and John S Breese. Causal independence for probability assessment and inference using bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 26(6):826–831, 1996.

[126] Marco Scutari, Catharina Elisabeth Graafland, and José Manuel Gutiérrez. Who learns better bayesian network structures: Constraint-based, score-based or hybrid algorithms? In *International Conference on Probabilistic Graphical Models*, pages 416–427, 2018.

[127] Sander Greenland. An introduction to instrumental variables for epidemiologists. *International journal of epidemiology*, 29(4):722–729, 2000.

[128] Stephen Burgess, Dylan S Small, and Simon G Thompson. A review of instrumental variable estimators for mendelian randomization. *Statistical methods in medical research*, 26(5):2333–2355, 2017.

[129] David M Evans and George Davey Smith. Mendelian randomization: new applications in the coming age of hypothesis-free causality. *Annual review of genomics and human genetics*, 16:327–350, 2015.

[130] Jack Bowden, George Davey Smith, Philip C Haycock, and Stephen Burgess. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology*, 40(4):304–314, 2016.

[131] Stephen Burgess, Jack Bowden, Tove Fall, Erik Ingelsson, and Simon G Thompson. Sensitivity analyses for robust causal inference from mendelian randomization analyses with multiple genetic variants. *Epidemiology (Cambridge, Mass.)*, 28(1):30, 2017.

[132] Stephen Burgess and Simon G Thompson. Interpreting findings from Mendelian randomization using the MR-Egger method. *European journal of epidemiology*, 32 (5):377–389, 2017.

[133] Alexander Teumer. Common methods for performing mendelian randomization. *Frontiers in cardiovascular medicine*, 5:51, 2018.

[134] Jack Bowden and Michael V Holmes. Meta-analysis and Mendelian randomization: A review. *Research synthesis methods*, 10(4):486–496, 2019.

[135] J Cohen. Statistical power analysis for the behavioral sciences, 2nd edn. Á/L, 1988.

[136] Ruth Tsang, Lindsey Colley, and Larry D Lynd. Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. *Journal of clinical epidemiology*, 62(6):609–616, 2009.

[137] UNDERPOWERED STATISTICS. STATISTICAL POWER AND UNDERPOWERED STATISTICS. 2015.

[138] William J Browne, Mousa Golalizadeh Lahi, and Richard MA Parker. A guide to sample size calculations for random effect models via simulation and the MLPowSim software package. *Bristol, United Kingdom: University of Bristol*, 2009.

[139] Len Thomas. Retrospective power analysis. *Conservation Biology*, 11(1):276–280, 1997.

[140] John M Hoenig and Dennis M Heisey. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1):19–24, 2001.

[141] Annelotte Philipsen, ANNE-LOUISE HANSEN, MARIT JØRGENSEN, Søren Brage, Bendix Carstensen, Annelli Sandbaek, THOMAS ALMDAL, Jeppe Gram, ERLING PEDERSEN, Torsten Lauritzen, et al. Associations of objectively measured physical activity and abdominal fat distribution. *Medicine & Science in Sports & Exercise*, 47(5):983–989, 2015.

[142] Robert W Koivula, Naeimeh Atabaki-Pasdar, Giuseppe N Giordano, Tom White, Jerzy Adamski, Jimmy D Bell, Joline Beulens, Søren Brage, Søren Brunak, Federico De Masi, et al. The role of physical activity in metabolic homeostasis before and after the onset of type 2 diabetes: an IMI DIRECT study. *Diabetologia*, pages 1–13, 2020.

[143] Michael E Sobel. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology*, 13:290–312, 1982.

[144] Xiaowei Zhan, Youna Hu, Bingshan Li, Goncalo R Abecasis, and Dajiang J Liu. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics*, 32(9):1423–1426, 2016.

[145] Ursula Neumann, Nikita Genze, and Dominik Heider. EFS: an ensemble feature selection tool implemented as R-package and web-application. *BioData mining*, 10 (1):1–9, 2017.

[146] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning*, 65(1): 31–78, 2006.

[147] Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723, 2005.