NUR 883

9 789493 289222

Multi-Omics in Research: Epidemiology, Methodology, and Advanced Data Analysis

TARIQ OSAMA FAQUIH

Multi-Omics in Research

Epidemiology, Methodology, and Advanced Data Analysis

TARIQ OSAMA FAQUIH

Multi-Omics in Research:
Epidemiology, Methodology, and Advanced Data Analysis

Proefschrift

ter verkrijging van

de graad van doctor aan de Universiteit Leiden,

op gezag van rector magnificus prof.dr.ir. H. Bijl,

volgens besluit van het college voor promoties

te verdedigen op dinsdag 28 maart 2023

klokke 16:15 uur

door

Tariq Osama Faquih

geboren te Riyadh, Saudi Arabia

in 1989

UITNODIGING

voor het bijwonen
van de openbare verdediging
van het proefschrift

Multi-Omics in Research
Epidemiology, Methodology, and Advanced Data Analysis

door Tariq O. Faquih

E-mail:
t.o.faquih@lumc.nl
tfaquih@bwh.harvard.edu

op dinsdag 28 maart 2023 om
16:15 uur in het
Academiegebouw
Rapenburg 73, Leiden

Actuele informatie
over het bijwonen van de
verdediging en receptie volgt

Livestream
https://tinyurl.com/33pejvfb

Paranimfen
Eleonora Camilleri
Jiao Luo

STELLINGEN

behorende bij het proefschrift

Multi-Omics in Research: Epidemiology, Methodology, and Advanced Data Analysis

1. Choosing the method to impute missing metabolite data, or not to impute at all, depends on the study characteristics and the nature of measured metabolites – this thesis (chapter 3)
2. The alluring quantity of big data in OMICs research does not necessarily reflect its quality. – this thesis (chapter 2)
3. Network analysis methods in OMICs can be used to identify novel biological pathways, enrich insight in disease etiology, and reveal potential drug targets. – this thesis (chapter 5)
4. Genes associated with a reduction of body weight do not necessarily associate with a favorable metabolomic profile. – this thesis (chapter 7)
5. "The metabolic patterns of individuals are crucial in connection with their susceptibility to disease" (Roger J. Williams, Biochemical Institute Studies IV, 1951) but their understanding requires advanced analytical, epidemiological, and biological approaches.
6. Although copy-number variations are ignored in most genome wide association studies, they are an important type of genomic variation that can help resolve the infamous "missing heritability problem"– (Génin, E., Hum Genet, 2020).
7. Minimum risk levels (MRL) of PFAS in drinking water and in the environment are claimed to be safe for humans (Sunderland, EM. et al., J Expo Sci Environ Epidemiol., 2019). Constantly adjusting the MRL guidelines only facilitates opportunistic interpretation by the manufacturers. The ideal solution is to ban and remove all forms of PFAS from our environment.
8. Despite the strong associations between metabolites and some diseases, based on genetic causal inference studies, these associations do not seem to be causal. This may be explained by "common antecedents" that result both in altered metabolite levels as well as disease (Surendran, P. et al. Nat Med., 2022), but also by the pleiotropic nature of many metabolites.
9. "Nature considered rationally, that is to say, submitted to the process of thought, is a unity in diversity of phenomena; a harmony, blending together all created things, however dissimilar in form and attributes; one great whole animated by the breath of life." The recent technological developments in multi-OMICs research make it realistic to address nature as this great whole (Alexander von Humboldt, Cosmos, 1845).
10. Let's embrace Carl Sagan's words: "Like it or not, we are stuck with science. We had better make the best of it. When we finally come to terms with it and fully recognize its beauty and its power, we will find, in spiritual as well as in practical matters, that we have made a bargain strongly in our favor". (The Demon-Haunted World, 1997).
11. "All we have to decide is what to do with the time that is given us." (J.R.R. Tolkien, The Fellowship of the Ring, 1954). We should pursue what is important and meaningful to us, even if it is a hard a path to take.

Multi-Omics in Research: Epidemiology, Methodology, and Advanced Data Analysis

Tariq Osama Faquih

Multi-Omics in Research: Epidemiology, Methodology, and Advanced Data Analysis
ISBN: 978-94-93289-22-2
NUR: 883

Cover design: Tariq Faquih
Design and Layout: Tariq Faquih & Vincent van Zandvoord (BuroVormvast.nl)
Printing: printsupport4u | www. printsupport4u.nl
Financial support:
Financial support King Abdullah Scholarship Program and King Faisal Specialist Hospital & Research Center for the publication of this thesis also gratefully acknowledged.

Promotor

Prof.dr.ir. J.A.P. Willems van Dijk

Co-promotors
Dr. A. van Hylckama Vlieg
Dr. D.O. Mook-Kanamori

Leden promotiecommissie
Prof.dr. M. Wuhrer
Prof.dr. R.H.H. Groenwold
Dr. J.F. Felix (Erasmus MC, University Medical Center Rotterdam)
Dr. A.N. Aziz (University of Bonn, Germany)

To my Mother and Father,

Table of Contents

Chapter 1
General introduction

1 GENERAL INTRODUCTION: THE RISE OF OMICS

The human genome project was a scientific milestone for human biological understanding (1, 2). This achievement started an era of large genetic analyses of an assortment of diseases. Genetic studies, such as genome-wide association studies (GWAS), have since revolutionized our understanding of disease etiology, prognosis, and diagnosis, and have contributed to public health (3, 4). Cardiometabolic diseases, including obesity, cardiovascular disease (CVD), type 2 diabetes (T2D), hypertension, and liver disease such as non-alcoholic fatty liver disease (NAFLD), are prevalent diseases that have benefited from genomic studies. GWAS have resulted in the identification of thousands of single nucleotide polymorphisms (SNPs) associated with these diseases (5). Furthermore, these associated SNPs enabled genetic epidemiological studies that identified causal associations, expanded our understanding of the pathophysiology, and improved the prediction of these diseases. In the wake of rapid technological advancements, it became possible to perform extremely large genomic studies in millions of individuals.

2 PROTEOMICS AND METABOLOMICS

Technological developments have enabled the study of genome-wide gene expression (6), whole genome DNA modifications in various body tissues (7), and the large scale measurement of proteins and metabolites downstream of the genome (8). The large-scale study of biological measures is generally referred to as OMICs. When referring to proteomics and metabolomics, we are referring to modern methods of mass measurements of hundreds to thousands of proteins and metabolites from a single sample. These types of studies usually include a large number of individuals and therefore often require the collaboration of several cohorts.

2.1 Proteomics

Measurement and analysis of proteins has been possible for over 200 years (9). The start of modern proteomics can be dated back to the 1960s and 70s with the advent of two-dimensional gel electrophoresis (10) and the creation of protein databases (11). However, this process was slow, had low throughput, and was not easily reproducible. Advances in mass spectrometry (MS) in the 1990s provided a powerful tool for the identification of proteins that bypassed the limitations of previous methods (10). After the completion of the human genome project, the proteome became a new focus to complement the newly sequenced genome (10, 12).

Proteomics and proteome research aims to achieve several goals. First, to use high throughput technologies to enable the identification and quantification of all human proteins. The number of proteins that can be produced from genes is amplified by alternative RNA splicing and post-translational modifications. Whereas the genome is nearly identical in every cell of the body, the proteome can differ substantially. The variable expression of genes in different cell types as well as environmental influences determine when and in which cells proteins are produced (13). These facts make the proteome more dynamic than the genome. Moreover, it makes it difficult to pinpoint the total number of proteins in humans. Currently, the estimated number of human proteins varies from 10,000 to billions (13). Second, in addition to the quantification of proteins, proteomic research enables studying the functionality of proteins. This includes the association between protein function and disease. Assessing functionality is complicated by protein-protein interactions and protein-DNA interactions (14). Currently, proteomics research has yielded large protein atlases publicly available online, such as the human protein atlas (15). Diseases that have been studied using proteomics include T2D and CVD (16), Alzheimer's disease (17), NAFLD (18), osteoarthritis (19), and venous thrombosis (20), to name a few.

2.2 Metabolomics

Metabolomic research measures small biomolecules in biofluids that are often substrates and products of metabolism (21, 22). Metabolites can thus be consumed and produced by endogenous metabolic processes or acquired from external sources and subsequently modified. Metabolites include amino acids, fatty acids, cholesterols, nucleotides, triglycerides, lipids, lipoproteins, and externally acquired compounds such nicotine and its metabolites from smoking (22, 23). One of the earliest studies on metabolite measurements was reported in the 1940's by Willems et al. (24-26). This work was included in their pivotal publication in 1951 where they coined the concept of "metabolic profiles" (24, 26). In this work, the authors demonstrated the methodology of quantifying and estimating different metabolites from urine and saliva samples using paper chromatography. They further reported their extensive work on different metabolic profiles of alcoholics, schizophrenics, mentally deficient children, overweight and underweight individuals, and the metabolic profile of different diets (24).

In parallel with the aforementioned completion of the human genome project, advances in MS and nuclear magnetic resonance (NMR) technology and the expansion of proteomics in the early 2000s, metabolomics has followed suit and gained momentum. This advancement has been driven by the development of commercial and non-commercial resources for metabolomic measurements that have made this more accessible and affordable for researchers (27). Metabolomics data have been used to provide insight into the pathophysiology of several diseases. Examples include the lipoprotein and metabolic profile that have been associated with the risk of coronary artery disease (28), the metabolomic profile of healthy and unhealthy body weight and their associations with disease outcomes (29), and the metabolomic profile for depression (30) as well as numerous other diseases. Furthermore, metabolomics has been used for discovering disease biomarkers such as neurodegenerative diseases, NAFLD (31), and colorectal cancer (32).

3 A SELECTION OF CURRENT PROTEOMICS AND METABOLOMICS MEASUREMENT PLATFORMS

A myriad of protein detection and quantification technologies have been developed over the last decades. These technologies include enzyme-linked immunosorbent assay (ELISA) and western blotting, two-dimensional gel electrophoresis, gas- and liquid chromatography, MS, NMR, and aptamer-based proteomics. Most metabolomics platforms use NMR or MS for metabolite detection and quantification. In this thesis we have used three techniques: for proteomics we have used the aptamer-based platform SomaScan, for metabolomics we have used the NMR based Nightingale Inc. platform and the ultra-high performance liquid chromatography - tandem mass spectrometry (UHPLC-MS/MS) based platform used by Metabolon Inc.

3.1 SomaScan for Proteomics

New techniques have been developed for high throughput protein measurements in recent years. One such method makes use of nucleotide based "aptamers" to identify and quantify proteins. A leading platform that uses aptamers for proteomic measurements is SomaScan. SomaScan (SomaLogic, Inc. Boulder, CO, USA) is a high throughput proteomics platform capable of simultaneous measurement of over a thousand proteins. Unlike traditional immunoassay instruments, SomaScan utilizes "Systematic Evolution of Ligands by Exponential enrichment" (SELEX), a biochemical technique used to create a library with a wide range of modified synthetic oligonucleotide ligands (i.e., the aptamers) bioengineered to bind to their respective protein targets. These aptamers are designed to emit a fluorescence signal only when they are bound to

their target protein. This fluorescence signature is then used to measure the relative concentrations of the target protein. This method provides several benefits over traditional immunoassay methods; Aptamers are inexpensive to produce, highly modifiable and are chemically stable. SomaLogic has developed a vast library of thousands of unique aptamers to detect proteins from a single biological sample (33). However, this technique is not perfect and some studies have noted that aptamers binding affinity can be affected by cross reactivity issues, genetic variations altering the protein structure, post-translational modifications, and the effects of the complexity and stability of the target protein structure (34). Nevertheless, SomaScan has been used for connecting genetics with the proteome profile of different diseases (35), creating a genomic atlas of the human proteome (36) and predicting coronary heart disease (37, 38) to name a few.

3.2 Metabolomics Measurement Techniques and Platforms

Metabolomic platforms are often divided into two categories: targeted and untargeted (also referred to as non-targeted). The targeted approach focuses on detection of predefined target metabolites (39). A benefit of using targeted metabolomics is its consistency and the possibility of generating absolute metabolite concentrations. Targeted metabolomic platforms, such as the Nightingale platform, currently measure several hundred metabolites (39, 40). Untargeted platforms, on the other hand do not fully target specific metabolites before the measurements. Instead, the platforms aim to detect and quantify as many metabolites as possible and subsequently identify them by cross referencing in large libraries of metabolites. The Metabolon platform is an example of untargeted platform that utilizes UHPLC-MS/MS technologies (41) and is capable of measuring thousands of metabolites including known and unknown metabolites.

3.2.1 Nightingale: Proton Nuclear Magnetic Resonance

The Nightingale metabolomics platform (Nightingale Health Plc., Helsinki, Finland) is a targeted platform that utilizes proton NMR (1H NMR) (39, 42). 1H NMR detects the hydrogen atoms of a prespecified selection of metabolites or macromolecules—that must be typically found in high concentrations in the samples—to capture their spectral characteristics. Due to these requirements 1H NMR is not as sensitive as other methods, such as MS (41). However, the properties of 1H NMR make it ideal for in depth quantitative measurement of lipoprotein particles (39). Moreover, it enables the reproducible quantification of absolute concentrations. The Nightingale platform utilizes three "molecular windows" for detecting different groups of metabolites. The lipoprotein (LIPO) window mainly detects the strong signal of lipoprotein subclasses and their lipid content. The low molecular weight molecules (LMWM) window filters out the signals from the LIPO window to detect the molecules with low molecular weight such as amino acids and glucose metabolites (42). Finally, the window for the lipids and lipids related molecules (LIPID) is used to specifically measure the saturated and unsaturated fatty acids, free and esterified cholesterol, sphingolipids, and phosphoglycerides (43). By using these three windows, 1H NMR Nightingale enables both in depth lipoprotein quantification and detection of several low molecular weight metabolites that are typically not easily measured using NMR. The Nightingale platforms provides approximately 230-245 metabolite measures including ratios and measurements of subfractions of lipoproteins (39).

3.2.2 Metabolon: Liquid Chromatography Tandem Mass Spectrometry

Chromatography and Mass Spectrometry

Chromatography is an important step for the separation of the biological molecules to enable the use of MS, particularly for metabolite and protein quantification (21). Chromatography is a

technique that applies high pressures to push the components of a biological sample through columns of silica, thus separating them. This requires the samples to be properly prepared and ionized before the procedure in order for the components to travel in the column. Moreover, the high pressure applied into the column must be consistent. This pressure can be supplied by either liquid or gas, respectively referred to as liquid chromatography (LC) and gas chromatography (GC). Exceptionally high pressure is in an advanced variant technique known as ultra-high performance liquid chromatography. The silica columns are designed to exhibit unique chemicals properties that have binding affinity with specific types of biological molecules. This interaction between the molecules and the column causes them to travel slower than those that do not bind to the column. In addition, the travel times differ between the different molecules in the column due to their unique affinities. The time it takes the molecules to pass through the column (referred to as the retention time) is key for inferring the molecular characteristics of the reacting molecules. Following this separation, mass spectrometry can be used to identify the molecules (21).

Fragmentation of proteins and metabolites is an essential step preceding mass spectrometry. Fragmentation is commonly achieved by shooting beams of electrons that break the biological molecules. These fragments are subsequently detected by their electronic charge (z) and their mass (m). By combining these two values as the mass to charge (m/z) ratio, a relatively unique signature is assigned to the fragments. The collected data of m/z ratios of the fragments are then represented as "spectral data". For metabolomics, a second MS step is frequently applied. This technique of coupling a chromatographer with two mass spectrometers in tandem is known as chromatography MS/MS. Tandem MS are used in metabolomics to measure metabolites in the first MS and subsequently fragment them into smaller particles to be measured in the second MS. This substantially increases the sensitivity of the measured m/z ratios of metabolites (44). Subsequently, the retention time from the chromatography step and the spectral data of m/z ratios from both mass spectrometers are combined (21, 41). For final identification, these signatures are compared to a reference library containing a vast number of retention times and m/z ratios corresponding to annotated metabolites or proteins.

Ultra High-Performance Liquid Chromatography and Tandem Mass Spectrometry

Metabolon™ Discovery HD4 platform is an untargeted metabolomic platform at Metabolon Inc. (Durham, North Carolina, USA) that utilizes ultra high-performance liquid chromatography and tandem mass spectrometry (UHPLC-MS/MS). This platform uses four independent UHPLC-MS/MS platforms with different LC columns (41, 45). Two platforms use positive ionization reverse phase chromatography, one uses negative ionization reverse phase chromatography, and one uses hydrophilic interaction liquid chromatography negative ionization (45). Thus, the platform can measure a wide range of metabolites with different chemical properties and affinities with high sensitivity. The signal from the m/z ratio and retention time of the measured ionized molecules are subsequently cross referenced with an in-house large library of metabolite molecules. If the metabolite is known, it will be assigned the annotation from the library. Metabolites found in the library but without a full annotation will also be reported as novel unnamed metabolites (41, 46). The Metabolon™ HD4 platform currently measures up to 1400 metabolites from a single sample.

Xenobiotics

A feature of the Metabolon platforms is its ability to measure not only endogenous metabolites produced by the body but also externally acquired "xenobiotic", or exogenous metabolites in an untargeted fashion. Essential metabolites for the human body that are acquired from the diet are usually considered part of the endogenous metabolite group. Xenobiotics on the other

hand include nonessential metabolites from food consumption (such as caffeine) (8) Moreover, xenobiotics comprise metabolites and chemicals derived from environmental exposures (e.g., pollution or chemical contamination), nutrition and diet, lifestyle choices (e.g., smoking, applying cosmetics), and medication use. The study of the effect of these environmental and external exposures on individuals is also known as the study of the exposome. In addition to the factors above, the exposome also includes factors such as the general environment, social economic status, and climate related factors. Exposome research aims to elucidate the "external" causes of disease and improve their prevention by examining the environment of patients (47). Indeed, environmental exposures are suggested to have a stronger impact on health outcomes than genetic factors (48). Therefore, the Metabolon platform provides a glimpse of an individual's exposome in addition to the broad array of endogenous metabolites. In addition, novel "unnamed" metabolites are also measured which may belong to endogenous or xenobiotic sources. This enables the simultaneous study of both the internal and external metabolomic factors involved in disease etiology and clinical outcomes (8) and the identification of associations with novel, unknown metabolites.

4 GROWING PAINS: CHALLENGES IN EPIDEMIOLOGICAL STUDIES USING OMICS

4.1 Genomic Challenges: capturing structural variation and the missing heritability problem

Genomic studies were the first and are the most established of the contemporary OMICs fields. However, genomics studies have their shortcomings and limitations (49). One of which is their inherent inability to capture any effects from the exposome. Another issue with genomic studies is the observation that a very minor proportion, often less than 5%, of the heritability of many traits is explained by the genetic variants tested (50, 51). This is referred to as the missing heritability problem (52, 53). Despite the increase in the number of genomic studies and the sample sizes of the cohorts in these studies (often hundreds of thousands of subjects), the explained variance from these studies remains low (50). Basic genetic analyses can estimate the heredita-bility of diseases, especially in familial and twin studies that share large portions of the genome (54, 55). When GWAS became possible and widely available, it was expected that the identified loci and SNPs would fully account for the known hereditability estimates. Surprisingly, this was not the case. Even extensive GWAS with thousands of individuals still reported loci that combined explain a sub-portion of the hereditability of several phenotypes, such as T2D and height (53). It has been suggested that one of the reasons for the missing hereditability is that genetics alone do not capture the complete heritability of complex diseases, such as cardiometabolic diseases (52). The influence of environmental factors and their interaction with genetics could account for the missing heritability (52). Another potential reason for this problem is that GWAS do not usually include genetic variations outside of SNPs (52). SNPs are the most common type of variation in the genome and are defined as a germline substitution of a single nucleotide at a specific position in the genome. However, other mutation types exist such as copy number variations (CNV) or repeat expansions, in which large number of nucleotides or patterns of nucleotides are repeated in the genome. The intrinsic properties of SNPs make their detection using DNA sequencing techniques much easier then CNVs and other types of genetic variation. This has contributed to the focus on SNPs in GWAS. However, due to the rapid advances in genomic technology, it has become possible to readily detect CNVs and other structural variations in the genome (56). Indeed, these recent advances have enabled genomic research in examining CNVs

in large studies. These studies have found that CNVs strongly contributed to the hereditability of various traits such as height and weight (52).

4.2 Challenges in Proteomics and Metabolomics

Many challenges may be encountered when using large metabolomics and proteomics data sets. Technical issues can occur during the preparation, processing, and quantification phase of metabolites and proteins. These issues can either be specific to the type of platform and technique used or can be general biological or chemical problems associated with the compound to be measured. For instance, if the biological sample and platform preparation is not performed properly, then the measurement and detection of the metabolites or proteins would be affected. This can occur if, for example, aptamers are not prepared properly or if the chromatography columns are contaminated due to overuse. Moreover, the efficiency of the measurement technique differs depending on the targeted biomarkers. Chemically complex biomarkers can be more difficult to quantify than simpler ones (57).

A more general issue that occurs in OMIC research is the batch and run day effects. OMIC platforms usually use the standard 96 well plate format to store and measure the samples. Therefore, they are limited in the number of samples that can be simultaneously quantified. Studies of hundreds or thousands of individuals requires sending the samples in batches at different time points. Additionally, each batch needs to be split to smaller groups to be measured by the platform over several days. The variation of how well each batch is stored and handled can affect the level of contamination and degradation in the samples. Moreover, the level of contamination may accumulate in the measurement platform itself after each batch run. Thus, the sensitivity of the platform and the samples quality can differ per run day and per batch, resulting in potential batch effects and measurement errors.

Another common problem during quantification occurs when some metabolites or proteins are in low concentrations in the sample. A limitation of most OMICs platforms is their inability to distinguish between different metabolites or proteins if their concentrations are below a certain cutoff range. This cutoff is referred to as the limit of detection (LoD). Metabolites or proteins below the LoD cannot be quantified and instead their concentrations are set to 0 or left blank (46). Run day and batch effects also contribute to the sensitivity of the platform and, in turn, the range of the LoD limits for the platforms (58).

After the physical quantification of the biological samples, computational post-processing steps are required. In these steps, further issues can occur. For example, correct matching of the m/z ratio and retention time from the UHPLC-MS/MS based Metabolon platform to the correct entry in the reference library is prone to machine and human errors. The signal matching procedure is usually automated by a software tool and then double checked manually. However, due to software errors or human errors, it is possible that a metabolite signal is not matched correctly. Similarity between metabolites or poor calibration of the platform during measurement also contribute to likelihood of these issues. Thus, a valid metabolite signature could be unmatched and, in worst case scenarios, be incorrectly matched with a completely different metabolite signature (46).

Once the quantification and postprocessing steps are complete, the generated data is used for statistical analysis. Here as well complications can occur. OMICs approaches have the benefit of measuring hundred to thousands of biochemicals from a single sample leading to high dimensional data. Often, the number of measurements is larger than the number of individuals in the study (N<P). If the sample size is too small, then this high dimensionality can decrease

the power of the study and lead to aberrant results. Furthermore, even if the sample size is sufficient, high dimensional data requires extra analytical procedures such as adjustment for multiple testing (59). Furthermore, beforementioned issues that may arise during quantification and postprocessing must be addressed during the statistical analysis. Indeed, transformation methods must be used to treat the variations between the batches and the run days. Likewise, missing measurements from LoD or other technical difficulties must also be addressed.

4.2.1 Missing Data in Metabolomics

Treating missing data is a common issue in epidemiological research. In metabolomics, handling and imputing missing values in metabolite measurements is a particularly important and challenging issue. As mentioned, missing values can occur due to contaminations in the platforms which in turn affects the sensitivity of the measurements. This issue is more common in untargeted metabolomics platforms (58). The reason for this is that the nature of untargeted metabolomics is to detect a large number of metabolites without prior selection. This approach may suffer from errors during the signal identification and signal matching steps (58). Naturally, the odds of these mistakes occurring increases as the number of metabolites measured expands, such as the case in untargeted metabolomics, and as the number of samples increases. Missing values can also occur due to the beforementioned technical difficulties, such as LoD (60), batch and run day effects, and mismatching issues (58). In addition, missing values could be truly missing and should not be imputed at all. This is the case with xenobiotic metabolites that are expected to be measured in specific individuals only. For example, imputing missing values for metabolites related to the metformin would imply that all participants are diabetics. These different issues and the high dimensionality of the data makes it statistically challenging to apply appropriate imputation methods (58).

5 STUDY POPULATIONS

As aforementioned, OMIC research is typically performed with large sample sizes in large population-based studies to accommodate the large number of OMIC variables. The work in this thesis involved several population-based studies and collaborations with Dutch and international research groups.

5.1 NEO

The Netherlands Epidemiology of Obesity (NEO) study is an ongoing population-based, prospective cohort study of individuals aged 45–65 years, with an oversampling of individuals with overweight or obesity. Men and women aged between 45 and 65 years with a self-reported BMI of 27 kg/m^2 or higher, living in the greater area of Leiden (in the West of the Netherlands) were eligible to participate in the NEO study. In addition, all inhabitants aged between 45 and 65 years from one municipality (Leiderdorp) were invited, irrespective of their BMI. Recruitment of participants started in September 2008 and completed at the end of September 2012. In total, 6,671 participants have been included, of whom 5,217 with a BMI of 27 kg/m^2 or higher. Participants were invited to come to the NEO study center of the Leiden University Medical Center for one baseline study visit after an overnight fast of at least 10 hours. During the visit a blood sample of 108 mL was taken from the participants (61). The study was approved by the medical ethical committee of the Leiden University Medical Center.

5.2 THE-VTE

The Thrombophilia, Hypercoagulability and Environmental Risks in Venous Thromboembolism (THE-VTE) study — a multicenter case control study from Leiden (The Netherlands) and Cambridge (UK) (62). Inclusion took place between March 2003 and December 2008. In total, 626 patients were included, aged 18-75, with a first DVT or PE. Partners of the patients were invited as controls. Subsequent follow-up of the cases was performed to assess recurrence risk. The mean follow-up duration was 4.8 years after discontinuation of oral anticoagulant therapy. Blood samples were taken 2-3 months after discontinuation of anticoagulants. The study was approved by the Medical Ethics Committee of the Leiden University Medical Centre (Leiden, Netherlands) and the NHS Research Ethics Committee in Cambridge, UK.

5.3 PROSPER

The Prospective Study of Pravastatin in the Elderly at Risk (PROSPER) was a randomized, double-blind, placebo-controlled trial among 5,786 men and women between 70-82 years old with a pre-existing vascular disease or a raised risk for such a disease. The aim of the trial was to test the benefits of Pravastatin. Participants were recruited from three countries with 2,517 individuals from Scotland, 2,173 individuals from Ireland and 1,096 individuals from the Netherlands. Fasting blood sample were collected and stored at -80 degrees for later NMR metabolomics analysis (63). The study was approved by the institutional ethics review boards of all centers and written informed consent was obtained from all participants (64).

5.4 NESDA

The Netherlands Study of Depression and Anxiety (NESDA) is an ongoing longitudinal cohort study into the long-term course and consequences of depressive and anxiety disorders. The sample consists of 2,981 participants with depressive/anxiety disorders and healthy controls recruited from the general population, general practices, and secondary mental health centers (65). Blood samples were collected after an overnight fast at the baseline visit (2004-2007). The Ethical Committees of all participating universities approved the NESDA project, and all participants provided written informed consent (66)

5.5 The Rhineland Study

The Rhineland Study is an ongoing prospective population-based cohort study based in two geographically defined areas In Bonn, Germany. Participants were recruited via invitation beginning in 2016. The primary focus of the study is on aging and age-related brain disorders in adult life. The source population consists of all inhabitants aged 30 years or older in the specified Bonn area. Participation was only possible upon invitation and regardless of health status provided they had sufficient command in the German language to provide an informed consent. The ethics committee of the medical faculty of the University of Bonn approved the undertaking of the study and it was carried out according to the recommendations of the International Council for Harmonisation Good Clinical Practice standards. Written informed consent was acquired from all participants per the Declaration of Helsinki.

5.6 INTERVAL

INTERVAL is a prospective cohort study nested within a pragmatic randomized trial of blood donors enrolled from 25 static centers of NHS Blood and Transplant (67). Recruitment of about 50,000 male and female donors started in June 2012 and was completed in June 2014. Blood

donors 18 years and older were consented and recruited from 25 National Health Service Blood and Transplant (NHSBT) static donor centers across England. All participants fulfilled all normal criteria for blood donation (68). Therefore, participants included in the study were predominantly healthy. The INTERVAL study was approved by the Cambridge (East) Research Ethics Committee. Written informed consent was obtained from all participants.

6 OUTLINE

In this thesis, we will look at the methodological challenges and epidemiological applications of genomics, proteomics, and in particular metabolomics. Part I focuses on methodological challenges and applications of proteomics and metabolomics research. In Part I, **Chapter 2**, we demonstrate the application of measures of agreement to compare contemporary large-scale aptamer-based proteomics with standardized clinical measurements in the THE-VTE study. Part I, **Chapter 3** explores the challenges of treating missing data in metabolomics and describes a workflow for imputing the different types of missing data. In Part I, **Chapter 4**, we present a metabolomic age prediction model based on 826 UHPLC-MS/MS measured metabolites from the INTERVAL study. We also report our evaluation of the model using several comorbidities in the NEO study. Part II of this of this thesis focuses on the etiological applications of metabolomics and genomics. In Part II, **Chapter 5**, we combine regression analysis and network analysis to investigate the association between UHPLC-MS/MS metabolite measurements with hepatic triglyceride content in the NEO study. In addition, we illustrate the results as an interactive online atlas. Part II, **Chapter 6**, focuses on the effects of genetic tandem repeat mutations in the *HTT* gene on NMR metabolite measurements in the NEO, PROSPER, and NESDA studies. We further explore the role of BMI mediation on the associations. In Part II, **Chapter 7**, we present the results for the effect of the environmental contaminant Per- and polyfluoroalkyl substances (PFAS) on the metabolic and lipoprotein profile of the general populations in Germany (The Rhineland Study) and the Netherlands (NEO study). In Part III, **Chapter 8**, we discuss our findings and offer our thoughts on the future evolution of multi-OMIC in research.

7 REFERENCES

1. Rood JE, Regev A. The legacy of the Human Genome Project. 2021;373(6562):1442-3.
2. Gibbs RA. The Human Genome Project changed everything. Nature Reviews Genetics. 2020;21(10):575-6.
3. Khoury MJ, Bowen MS, Clyne M, Dotson WD, Gwinn ML, Green RF, et al. From public health genomics to precision public health: a 20-year journey. Genetics in Medicine. 2018;20(6):574-82.
4. Molster CM, Bowman FL, Bilkey GA, Cho AS, Burns BL, Nowak KJ, et al. The Evolution of Public Health Genomics: Exploring Its Past, Present, and Future. Frontiers in public health. 2018;6:247.
5. Atanasovska B, Kumar V, Fu J, Wijmenga C, Hofker MH. GWAS as a Driver of Gene Discovery in Cardiometabolic Diseases. Trends in Endocrinology & Metabolism. 2015;26(12):722-32.
6. PHG Foundation. What is transcriptomics? : PHG Foundation; 2022 [updated 2022/08/16/. Available from: https://www.phgfoundation.org/blog/what-is-transcriptomics.
7. CDC. What is Epigenetics? | CDC Centers for Disease Control and Prevention 2022 [updated 2022/08/15/. Available from: https://www.cdc.gov/genomics/disease/epigenetics.htm.
8. Rattray NJW, Deziel NC, Wallach JD, Khan SA, Vasiliou V, Ioannidis JPA, et al. Beyond genomics: understanding exposotypes through metabolomics. Hum Genomics. 2018;12(1):4.
9. Hartley H. Origin of the Word 'Protein'. Nature. 1951;168(4267):244-.
10. Pandey A, Mann M. Proteomics to study genes and genomes. Nature. 2000;405(6788):837-46.
11. Strasser BJ. Collecting, comparing, and computing sequences: the making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954-1965. Journal of the history of biology. 2010;43(4):623-60.
12. Cox J, Mann M. Is Proteomics the New Genomics? Cell. 2007;130(3):395-8.
13. Ponomarenko EA, Poverennaya EV, Ilgisonis EV, Pyatnitskiy MA, Kopylov AT, Zgoda VG, et al. The Size of the Human Proteome: The Width and Depth. International journal of analytical chemistry. 2016;2016:7436849.
14. Gonzalez MW, Kann MG. Chapter 4: Protein interactions and disease. PLoS computational biology. 2012;8(12):e1002819.
15. Sjöstedt E, Zhong W, Fagerberg L, Karlsson M, Mitsios N, Adori C, et al. An atlas of the protein-coding genes in the human, pig, and mouse brain. 2020;367(6482):eaay5947.
16. Ferrannini G, Manca ML, Magnoni M, Andreotti F, Andreini D, Latini R, et al. Coronary Artery Disease and Type 2 Diabetes: A Proteomic Study. Diabetes Care. 2020;43(4):843-51.
17. Whelan CD, Mattsson N, Nagle MW, Vijayaraghavan S, Hyde C, Janelidze S, et al. Multiplex proteomics identifies novel CSF and plasma biomarkers of early Alzheimer's disease. Acta neuropathologica communications. 2019;7(1):169.
18. Niu L, Geyer PE, Wewer Albrechtsen NJ, Gluud LL, Santos A, Doll S, et al. Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease. Molecular systems biology. 2019;15(3):e8793.
19. Tardif G, Paré F, Gotti C, Roux-Dalvai F, Droit A, Zhai G, et al. Mass spectrometry-based proteomics identify novel serum osteoarthritis biomarkers. Arthritis research & therapy. 2022;24(1):120.
20. Edfors F, Iglesias MJ, Butler LM, Odeberg J. Proteomics in thrombosis research. Research and practice in thrombosis and haemostasis. 2022;6(3):e12706.
21. Alonso A, Marsal S, Julia A. Analytical methods in untargeted metabolomics: state of the art in 2015. Front Bioeng Biotechnol. 2015;3:23.
22. Ryals J, Lawton K, Stevens D, Milburn M. Metabolon, Inc. Pharmacogenomics. 2007;8(7):863-6.
23. Roethig HJ, Munjal S, Feng S, Liang Q, Sarkar M, Walk RA, et al. Population estimates for biomarkers of exposure to cigarette smoke in adult U.S. cigarette smokers. Nicotine & tobacco research : official journal of the Society for Research on Nicotine and Tobacco. 2009;11(10):1216-25.
24. Willems RJ. Individual metabolic patterns and human disease : an exploratory study utilizing predominantly paper chromatographic methods. The University of Texas Publication. 1951;BIOCHEMICAL INSTITUTE STUDIES IV

25. Williams RJ, Berry LJ, Beerstecher E. Individual Metabolic Patterns, Alcoholism, Genetotrophic Diseases*. 1949;35(6):265-71.
26. Gates SC, Sweeley CC. Quantitative metabolic profiling based on gas chromatography. Clinical chemistry. 1978;24(10):1663-73.
27. Miggiels P, Wouters B, van Westen GJP, Dubbelman A-C, Hankemeier T. Novel technologies for metabolomics: More for less. TrAC Trends in Analytical Chemistry. 2019;120:115323.
28. Tikkanen E, Jägerroos V, Holmes MV, Sattar N, Ala-Korpela M, Jousilahti P, et al. Metabolic Biomarker Discovery for Risk of Peripheral Artery Disease Compared With Coronary Artery Disease: Lipoprotein and Metabolite Profiling of 31 657 Individuals From 5 Prospective Cohorts. 2021;10(23):e021995.
29. Cirulli ET, Guo L, Leon Swisher C, Shah N, Huang L, Napier LA, et al. Profound Perturbation of the Metabolome in Obesity Is Associated with Health Risk. Cell Metabolism. 2019;29(2):488-500.e2.
30. Bot M, Milaneschi Y, Al-Shehri T, Amin N, Garmaeva S, Onderwater GLJ, et al. Metabolomics Profile in Depression: A Pooled Analysis of 230 Metabolic Markers in 5283 Cases With Depression and 10,145 Controls. Biological psychiatry. 2020;87(5):409-18.
31. Masoodi M, Gastaldelli A, Hyötyläinen T, Arretxe E, Alonso C, Gaggini M, et al. Metabolomics and lipidomics in NAFLD: biomarkers and non-invasive diagnostic tests. Nature reviews Gastroenterology & hepatology. 2021;18(12):835-56.
32. Zhang A, Sun H, Yan G, Wang P, Han Y, Wang X. Metabolomics in diagnosis and biomarker discovery of colorectal cancer. Cancer Letters. 2014;345(1):17-20.
33. Gold L, Ayers D, Bertino J, Bock C, Bock A, Brody EN, et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. PLoS One. 2010;5(12):e15004.
34. Joshi A, Mayr M. In Aptamers They Trust: The Caveats of the SOMAscan Biomarker Discovery Platform from SomaLogic. Circulation. 2018;138(22):2482-5.
35. Suhre K, Arnold M, Bhagwat AM, Cotton RJ, Engelke R, Raffler J, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nature Communications. 2017;8(1):14357.
36. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. Nature. 2018;558(7708):73-9.
37. Ganz P, Heidecker B, Hveem K, Jonasson C, Kato S, Segal MR, et al. Development and Validation of a Protein-Based Risk Score for Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease. Jama. 2016;315(23):2532-41.
38. Cuvelliez M, Vandewalle V, Brunin M, Beseme O, Hulot A, de Groote P, et al. Circulating proteomic signature of early death in heart failure patients with reduced ejection fraction. Scientific Reports. 2019;9(1):19202.
39. Soininen P, Kangas AJ, Wurtz P, Suna T, Ala-Korpela M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. Circ Cardiovasc Genet. 2015;8(1):192-206.
40. Emwas A-HM, Salek RM, Griffin JL, Merzaban J. NMR-based metabolomics in human disease diagnosis: applications, limitations, and recommendations. Metabolomics. 2013;9(5):1048-72.
41. Evans A, Bridgewater B, Liu Q, Mitchell M, Robinson R, Dai H, et al. High resolution mass spectrometry improves data quantity and quality as compared to unit mass resolution mass spectrometry in high-throughput profiling metabolomics. Metabolomics. 2014;4(2):1.
42. Soininen P, Kangas AJ, Würtz P, Tukiainen T, Tynkkynen T, Laatikainen R, et al. High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. The Analyst. 2009;134(9):1781.
43. Fuertes-Martín R, Correig X, Vallvé JC, Amigó N. Human Serum/Plasma Glycoprotein Analysis by (1) H-NMR, an Emerging Method of Inflammatory Assessment. Journal of clinical medicine. 2020;9(2).
44. Heiles S. Advanced tandem mass spectrometry in metabolomics and lipidomics—methods and applications. Analytical and Bioanalytical Chemistry. 2021;413(24):5927-48.
45. Rhee EP, Waikar SS, Rebholz CM, Zheng Z, Perichon R, Clish CB, et al. Variability of Two Metabolomic Platforms in CKD. Clinical Journal of the American Society of Nephrology. 2019;14(1):40.

46. Dehaven C, Evans A, Dai H, Lawton K. Software Techniques for Enabling High-Throughput Analysis of Metabolomic Datasets. 2012.
47. Vrijheid M. The exposome: a new paradigm to study the impact of environment on health. Thorax. 2014;69(9):876-8.
48. DeBord DG, Carreón T, Lentz TJ, Middendorf PJ, Hoover MD, Schulte PA. Use of the "Exposome" in the Practice of Epidemiology: A Primer on -Omic Technologies. American journal of epidemiology. 2016;184(4):302-14.
49. Rochfort S. Metabolomics reviewed: a new "omics" platform technology for systems biology and implications for natural products research. Journal of natural products. 2005;68(12):1813-20.
50. Matthews LJ, Turkheimer E. Three legs of the missing heritability problem. Studies in history and philosophy of science. 2022;93:183-91.
51. Altmüller J, Palmer LJ, Fischer G, Scherb H, Wjst M. Genomewide scans of complex human diseases: true linkage is hard to find. American journal of human genetics. 2001;69(5):936-50.
52. Génin E. Missing heritability of complex diseases: case solved? Human Genetics. 2020;139(1):103-13.
53. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.
54. Falconer DS. The inheritance of liability to certain diseases, estimated from the incidence among relatives. 1965;29(1):51-76.
55. Young AI. Solving the missing heritability problem. PLOS Genetics. 2019;15(6):e1008222.
56. Gordeeva V, Sharova E, Arapidi G. Progress in Methods for Copy Number Variation Profiling. International journal of molecular sciences. 2022;23(4).
57. Joshi R, Wannamethee G, Engmann J, Gaunt T, Lawlor DA, Price J, et al. Establishing reference intervals for triglyceride-containing lipoprotein subfraction metabolites measured using nuclear magnetic resonance spectroscopy in a UK population. Annals of clinical biochemistry. 2021;58(1):47-53.
58. Do KT, Wahl S, Raffler J, Molnos S, Laimighofer M, Adamski J, et al. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. Metabolomics : Official journal of the Metabolomic Society. 2018;14(10):128-.
59. Tzoulaki I, Ebbels TMD, Valdes A, Elliott P, Ioannidis JPA. Design and Analysis of Metabolomics Studies in Epidemiologic Research: A Primer on -Omic Technologies. American journal of epidemiology. 2014;180(2):129-39.
60. Redestig H, Kobayashi M, Saito K, Kusano M. Exploring matrix effects and quantification performance in metabolomics experiments using artificial biological gradients. Anal Chem. 2011;83(14):5645-51.
61. de Mutsert R, den Heijer M, Rabelink TJ, Smit JW, Romijn JA, Jukema JW, et al. The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection. Eur J Epidemiol. 2013;28(6):513-23.
62. van Hylckama Vlieg A, Baglin CA, Luddington R, MacDonald S, Rosendaal FR, Baglin TP. The risk of a first and a recurrent venous thrombosis associated with an elevated D-dimer level and an elevated thrombin potential: results of the THE-VTE study. Journal of thrombosis and haemostasis : JTH. 2015;13(9):1642-52.
63. Delles C, Rankin NJ, Boachie C, McConnachie A, Ford I, Kangas A, et al. Nuclear magnetic resonance-based metabolomics identifies phenylalanine as a novel predictor of incident heart failure hospitalisation: results from PROSPER and FINRISK 1997. Eur J Heart Fail. 2018;20(4):663-73.
64. Shepherd J, Blauw GJ, Murphy MB, Bollen EL, Buckley BM, Cobbe SM, et al. Pravastatin in elderly individuals at risk of vascular disease (PROSPER): a randomised controlled trial. Lancet. 2002;360(9346):1623-30.
65. Penninx BW, Beekman AT, Smit JH, Zitman FG, Nolen WA, Spinhoven P, et al. The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. Int J Methods Psychiatr Res. 2008;17(3):121-40.

66. de Kluiver H, Jansen R, Milaneschi Y, Bot M, Giltay EJ, Schoevers R, et al. Metabolomic profiles discriminating anxiety from depression. Acta Psychiatr Scand. 2021;144(2):178-93.
67. Moore C, Sambrook J, Walker M, Tolkien Z, Kaptoge S, Allen D, et al. The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. Trials. 2014;15:363-.
68. NHS. NHS Blood and Transplant criteria for giving blood 2014 [updated 2022/08/11/. Available from: https://www.blood.co.uk/who-can-give-blood.

Part I

Methodological Challenges in Proteomics and Metabolomics

Chapter 2

Agreement of aptamer proteomics with standard methods for measuring venous thrombosis biomarkers

Authors: Tariq Faquih, M.S.[1], Dennis Mook-Kanamori, MD PhD.[2,3], Frits Rosendaal, MD PhD.[1], Trevor Baglin, MD PhD.[4], Ko Willems van Dijk, MD PhD.[5,6,7], Astrid van Hylckama Vlieg, PhD.[2]

1. Department of Clinical Epidemiology, Leiden University Medical Center, Postal Zone C7-P, PO Box 9600, 2300 RC, Leiden, The Netherlands. Email: t.o.faquih@lumc.nl, Phone Number: +31 626 696 712
2. Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands.
3. Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands.
4. Medicxi Ventures LLP, 25 Great Pulteney Street, W1F 9LT, London, UK
5. Department of Internal Medicine, Division of Endocrinology, Leiden University Medical Center, Leiden, The Netherlands.
6. Einthoven Laboratory for Experimental Vascular Medicine, Leiden University Medical Center, Leiden, The Netherlands.
7. Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands.

Authors' contributions:
T. Faquih designed the study, analysed and drafted the manuscript. A. van Hylckama Vlieg conceived and designed the study, acquired the data, obtained funding, supervised the study. K. Willems van Dijk and D. O. Mook-Kanamori conceived, supervised, and designed the study. T. P. Baglin acquired the data supervised the study. F. R. Rosendaal supervised the study. All authors reviewed the manuscript.

Acknowledgements:
The authors wish to thank the director of the anticoagulation clinic of Leiden, the Netherlands (F. J. M. van der Meer), who made the recruitment of patients in Leiden possible. L. Willems of Brilman is thanked for performing interviews and blood draws. I. de Jonge is thanked for her administrative support and data management. We are grateful to P. Noordijk, L. Mahic and A. Hoenderdos for sample processing and laboratory analysis.

Conflict of Interests:
T. Baglin has received honoraria for consultancy from Organon Teknika and bioMerieux. D. O. Mook-Kanamori is a part-time clinical research consultant for Metabolon, Inc. All other authors have nothing to disclose.
Summary word count = 220/250; Body word count = 2293/2000; references = 24/30; 1 Table; 1 Figure
Keywords: Venous Thrombosis; Proteomics; Biomarkers; Aptamers, Nucleotide; Immunoassay; Blood Coagulation Factors

Res Pract Thromb Haemost. **2021.** 5:e12526 doi:10.1002/rth2.12526, PMID: 34013156

Essentials

- Measurement agreement of aptamer proteomics for venous thromboembolism (VTE) markers is unknown.
- We selected 27 cases with unprovoked VTE and 27 controls from the THE-VTE study.
- Agreement between the aptamers and the laboratory methods for the VTE biomarkers was poor.
- Currently the usage of aptamer proteomics for VTE biomarkers should be considered with caution.

1 ABSTRACT

Background

Venous thromboembolism (VTE) is a complex disease with an incidence rate of about 1/1000 per year. Despite the availability of validated biomarkers for VTE, unprovoked events account for 50% of first events. Therefore, emerging high-throughput proteomics are promising methods for the expansion of VTE biomarkers. One such promising high-throughput platform is SomaScan, which utilizes a large library of synthetic oligonucleotide ligands known as aptamers to measure thousands of proteins.

Objective

The aim of this study was to evaluate the viability of the aptamer-based SomaScan platform for VTE studies by examining its agreement with standard laboratory methods.

Methods

We examined the agreement between eight established VTE biomarkers measured by SomaScan and standard laboratory immunoassay and viscosity-based instruments in 54 individuals (27 cases and 27 controls) from the THE-VTE study. We preformed the agreement analysis by using a regression model and predicting the estimates and the 95% prediction interval (PI) of the laboratory instruments values using SomaScan values.

Results

SomaScan measurements exhibited overall poor agreement, particularly for D-dimer (average fit [95% PI]: 492.7 ng/mL [110.0-1998.2]) and fibrinogen (average fit [95% PI]: 3.3g/L [2.0-4.7]).

Conclusion

Our results indicate that SomaScan measurement had poor agreement with the standard laboratory measurements. These results may explain why some genome wide association studies with VTE proteins measured by SomaScan did not confirm previously identified loci. Therefore, SomaScan should be considered with caution in VTE studies.

2 INTRODUCTION

Venous thromboembolism (VTE) is a complex disease caused by an imbalance in the coagulation and fibrinolysis pathways. VTE, which encompasses deep vein thrombosis (DVT) and pulmonary embolism (PE), has an incidence rate of 1/1000 per year[1]. Venous thromboembolism is assumed to be caused by both acquired and genetic risk factors[1], but the mechanisms that evoke VTE involve complex interactions of pathways that are still not fully understood[1, 2]. Several genetic variants and proteins and have been identified as risk factors for VTE, such as factor V Leiden (*F5* rs6025), prothrombin 20210A (*FII* rs1799963), low levels of antithrombin, protein S and protein C, and high levels of factor VIII, IX, and XI[1].

The current standard methods to quantify coagulation factors are by viscosity or optical detection based coagulation analysers and immunoassay-based laboratory instruments targeting one or more coagulation factors. For example, elevated levels of D-Dimer are associated with increased risk of DVT, recurrent DVT and mortality[3]. Moreover, D-dimer is measured with >95% diagnostic sensitivity by immunoassay-based instruments and hence broadly used by clinicians for the exclusion of VTE in patients with low or intermediate risk[4].

Further expansion of the number of biomarkers is imperative for studying the aetiology and improving prediction of VTE. Emerging high throughput proteomic platforms are promising tools to identify such novel biomarkers as these platforms are capable of quantifying large numbers of proteins simultaneously from a single sample[5, 6]. The aim of our study was to assess the measurement agreement of one such platform, the aptamer-based SomaScan platform, by comparing its measurements with the current established laboratory methods for eight VTE biomarkers in THE-VTE study.

2.1 SomaScan

SomaScan (SomaLogic, Inc. Boulder, CO, USA) is a high throughput proteomics platform capable of simultaneous measurement of thousands of proteins. Unlike traditional immunoassay instruments, SomaScan utilizes Systematic Evolution of Ligands by Exponential enrichment (SELEX), a biochemical technique used to create a library with a wide range of modified synthetic oligonucleotide ligands (known as aptamers) designed to bind to their respective protein targets. Aptamers provide several benefits over immunoassay methods; They are inexpensive to produce, highly modifiable and are chemically stable. SomaScan has developed a vast library of unique aptamers (SOMAmers) to detect thousands of proteins[7]. This makes the SomaScan platform appealing for researchers and, indeed, several large studies have utilized the platform for various purposes including the identification novel protein markers[8-10].

3 METHODS

3.1 THE-VTE study

For our cases and controls, we used samples from the Thrombophilia, Hypercoagulability and Environmental Risks in Venous Thromboembolism (THE-VTE) study — a multicentre case control study from Leiden (The Netherlands) and Cambridge (UK)[11]. Inclusion took place between March 2003 and December 2008. In total, 626 patients were included, aged 18-75, with a first DVT or PE. Partners of the patients were invited as controls. Subsequent follow-up of the cases was performed to assess recurrence risk. The mean follow-up duration was 4.8 years after discontinuation of oral anticoagulant therapy. Blood samples were taken 2-3 months after discontinuation of anticoagulants. The blood samples were collected into Sarstedt Monovette tubes, in a 0.1 volume of 0.106 m trisodium citrate and separated into plasma by centrifugation then stored at -80 °C. All participants gave written informed consent. The study was approved by the Medical Ethics Committee of the Leiden University Medical Centre (Leiden, Netherlands) and the NHS Research Ethics Committee in Cambridge, UK.

The current pilot study was originally designed to explore the biomarker measurements differences between unprovoked VTE and controls in a case-control design using SomaScan. Before proceeding, we checked the general measurement agreement between SomaScan and standard laboratory immunoassay and viscosity-based instruments in both cases and controls to assess the agreement over the whole range of coagulation factors levels. Unprovoked events were defined as individuals who did not have surgery, trauma or long-term immobilisation 3 months prior to the event. Moreover, patients were excluded if they had an active malignancy, abnormal levels of proteins C, protein S and antithrombin (<80 U/dL), used hormone replacement therapy or hormonal contraceptives at the time of the event. Patients with factor V Leiden or prothrombin (PT20210A) mutations were also excluded. We selected a sample of 16 cases with unprovoked VTE and further added eleven patients who experienced a recurrent venous thrombosis during follow-up, resulting in the inclusion of 27 VTE cases. By including unprovoked VTE cases as well as controls we covered a wide range of VTE biomarker values. Finally, we randomly selected 27 participants without VTE as controls. The frozen samples of the selected cases and controls were sent and thawed for analysis by SomaScan in 2016. No thawing or refreezing was performed during the interim period between 2011-2016. VTE biomarkers were measured by validated immunoassay-based, viscosity-based detection instruments, henceforth referred to as laboratory instruments. D-dimer total concertation (ng/mL) was measured by the Vidas D-dimer immunoassay (BioMérieux, Basingstoke, UK). The activity (international units per millilitre; IU/ml) of protein C, protein S, antithrombin (using chromogenic assay), prothrombin, coagulation factor IX, and coagulation factor XI were measured by STA-R coagulation analyser (Diagnostica Stago, S.A.S, Asnières sur Seine, France). Fibrinogen total concertation (g/L) was measured by STA-R coagulation analyser (Diagnostica Stago, S.A.S, Asnières sur Seine, France)[11].

Samples were sent to SomaLogic (Boulder, CO, USA) and measured by the SomaScan platform. The instrument measured 1310 total proteins of which 24 proteins failed the quality check and were flagged. We selected eight VTE biomarkers that were measured by laboratory instruments and successfully measured by SomaScan: D-dimer, prothrombin, protein C, protein S, antithrombin, fibrinogen, coagulation Factors IX, and XI. One control sample failed quality control was excluded.

To compare the agreement and interchangeability of the different measures we used the 95% agreement statistical method[12]. Since SomaScan uses relative fluoresces units (RFU) as measures for protein concentration, and the laboratory instruments measure absolute protein

concentrations or activity (IU/ml), we applied an alternative method to assess agreement if measurements are on different units, as described by Bland & Altman[12]. First, we performed a linear regression per biomarker with laboratory instruments measures as the outcome and SomaScan measures as the independent variable. Second, we used the regression models to predict estimates of the laboratory instruments values using SomaScan values. After checking the normality of the residuals, we log transformed the D-Dimer measurements as their distributions were very skewed. Finally, we calculated the 95% prediction interval (PI) to represent the equivalent of 95% limits of agreement[12]. This method is equivalent to comparing the mean differences of the two measurement methods. If the bias is consistent and the mean difference is close to 0 the result would show narrow prediction intervals. Consequently, the two methods would be interchangeable and in good agreement[13]. It is difficult to define hard cut-off points for the intervals. Therefore, judging the agreement is considered a clinical question rather than a statistical one[14].

4 DISCUSSION AND RESULTS

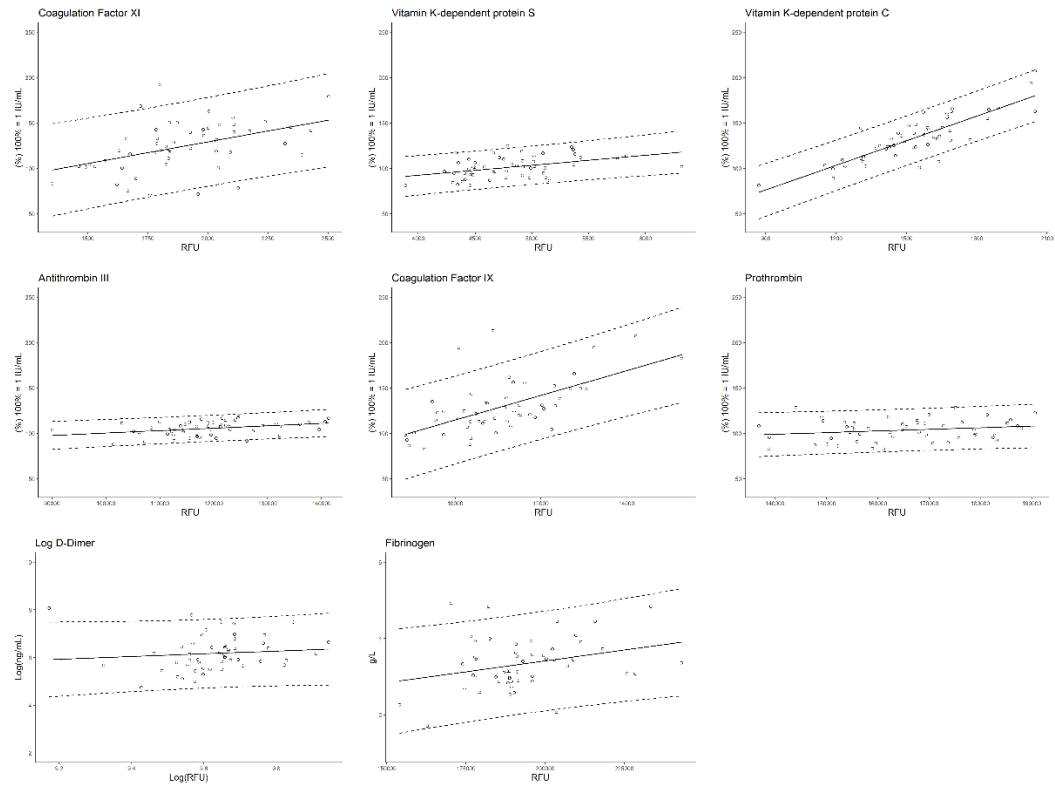
We examined the agreement of SomaScan measurements of VTE biomarkers with the laboratory instruments. The table and figures for the results are shown in Table 1 and Figure 1. Although a particular biomarker may seem to be in good agreement due to oblique slopes, the appropriate indicator for agreement is the width of the prediction interval around the average fit of the regression line[12].

Table 1: Average fit and average prediction intervals for SomaScan and laboratory instruments for the coagulation factors.

Protein Name	Units	Average Fit [Average 95% PIs]	Average Width
Coagulation Factor XI	(%) 100% = 1 IU/mL	124.3 [75.0-173.5]	98.5
Protein S	(%) 100% = 1 IU/mL	101.4 [80.2-122.7]	42.6
Protein C	(%) 100% = 1 IU/mL	133.9 [106.4-161.4]	55.1
Antithrombin	(%) 100% = 1 IU/mL	105.5 [91.0-120.1]	29.1
Coagulation Factor IX	(%) 100% = 1 IU/mL	129.6 [81.0-178.2]	97.2
Prothrombin	(%) 100% = 1 IU/mL	103.7 [80.3-127.1]	46.9
D-Dimer			
Log	Log(ng/mL*)	6.2 [4.7-7.6]	2.9
Back-transformed	ng/mL*	492.7 [110.0-1998.2]	1888.2
Fibrinogen	g/L	3.3 [2.0-4.7]	2.6

Activity of coagulation factor XI, protein S, protein C, antithrombin, coagulation factor IX, and prothrombin were measured by the same instrument and use IU/mL units. D-dimer and fibrinogen total concentrations were measured by immunoassay instruments. *D-dimer was assayed using the Vidas D-dimer assay. Unit type used was FEU = Fibrinogen equivalent unit (500 ng FEU/mL = 250 ng D-dimer/mL). Abbreviations: PIs: Prediction Interval; IU/mL: international units per millilitre; Average Width: the average difference between the lower and upper limits of the prediction interval.

Figure 1: 95% Limits of Agreements plots for each VTE biomarker.



Narrow prediction intervals indicate higher agreement between the SomaScan (x-axis) and laboratory instruments (y-axis). Antithrombin had the narrowest interval and, therefore, the best agreement. D-dimer had the poorest agreement as indicated by the wide interval in the log transformed plot. Abbreviations: RFU: relative fluoresces units; IU/mL: international units per millilitre.

Overall, the results indicate poor agreement of SomaScan with the validated laboratory instruments. The narrowest prediction interval, and thus the best agreement, were observed for antithrombin (average fit [95% PI]: 105.52 IU/mL [90.98-120.06]) followed by protein S (average fit [95% PI]: 101.44 IU/mL [80.15-122.73]), Prothrombin (average fit [95% PI]: 103.7 IU/mL [80.27-127.13]) and protein C (average fit: [95% PI] 133.87 IU/mL [106.35-161.4]). Factor IX (average fit [95% PI]: 129.62 IU/mL [81.03-178.21]) and factor XI (average fit [95% PI]: 124.26 IU/mL [75-173.53]) had a wide mean prediction interval and low agreement. The prediction interval for fibrinogen also had a wide interval (average fit [95% PI]: 3.3 g/L [2.0-4.7]). This result indicates that the predicted value of fibrinogen was within a prediction range (average width) of 2.6 g/L (~80%) of the laboratory measurement. Considering the wide prediction interval width and the fact that the normal range of fibrinogen is between 2 and 5 g/L[15], we concluded that the SomaScan measurements of fibrinogen are in poor agreement with the laboratory instrument. Finally, agreement between SomaScan and laboratory instruments for D-dimer had the widest average interval (average fit [95% PI]: 6.2 [4.7-7.6]) among the measured markers as shown in Figure 1. The values and width of the interval for the log D-dimer plot may seem normal compared

to the plots of the other measurements. However, unlike the other biomarkers, D-dimer was measured in ng/mL units and was log transformed in order to fulfil the requirement of normality for the analysis. After back-transforming the values, the agreement was very poor as indicated by the extremely wide prediction interval (average fit [95% PI]: 492.7 ng/mL [110.0-1998.2]).

The current study demonstrated poor agreement of SomaScan VTE measures with the laboratory instruments, particularly for D-dimer, which is a particularly important VTE biomarker[1, 4, 6], despite passing quality control. This poor agreement could explain the reported lack of association between SomaScan D-dimer measurement and the risk of DVT[16].

Despite the advantages of SomaScan for high throughput measurements, the observed disagreement could be due to some of the platform's shortcomings[17]. Some factors that can affect binding affinity are aptamer cross reactivity, genetic variations, post-translational modifications, and the complexity and stability of the target protein structure. Moreover, SomaScan measurements are quantitative and not qualitative and would not be able to detect qualitative defects in the analysis. It is important to note that our assessment was only of the eight VTE biomarkers available in our study and cannot be generalized to the agreement of the remaining proteins measured by SomaScan for our dataset.

Two previous studies assessed the association between SomaScan measurements and VTE SNPs in the SomaScan protein genome wide association study (pGWAS)[8, 10]. Since several genetic loci associated with VTE biomarkers have previously been identified, multiple hits were expected. However only factor XI and protein C measured by SomaScan associated with three loci and one locus, respectively. The lack of genetic correlations with the SomaScan measures for the biomarkers further supports our findings of poor agreement.

Possible limitations of the study are the usage of activity measurements for biomarkers versus the relative concentration reported by SomaScan. However, both D-dimer and fibrinogen showed poor agreement despite being measured as concentration measures. Moreover, viscosity-based activity measurements, such as the STA-R analyser used here, are considered the standard for VTE studies[18-20]. Furthermore, the recommended the sample size for Bland-Altman methods is usually N>100[21]. Our small size may affect the accuracy of the width of the 95% agreement intervals. However, we found that the agreement is very poor for some of the biomarkers, such as D-dimer, which cannot be fully explained by the sample size. Finally, it is unlikely the storage time of the plasma samples before the SomaScan analysis caused major degradation. Since the blood was collected, the samples were stored in -80 °C and the sampled aliquots were used for the primary analysis. Afterwards the samples were not thawed until the analysis by SomaScan five years later. Several studies have shown that these conditions were optimal for the storage of plasma samples and maintain minimal degradation[22-24]. Therefore, storage time and conditions are an unlikely cause to the disagreement in our results. Nevertheless, comparing the agreement of SomaScan with total concentrations for the other biomarkers and in larger studies may provide further insight.

5 CONCLUSION

The 95% limits agreement is a simple and effective statistical method for comparing measurements by different methods. We believe it is important to apply this type of analysis to compare the measurements of exciting novel high throughput platforms with current established measurements; thereby limiting measurement errors from affecting the results and conclusions based on such platforms.

In conclusion, despite the promising applications of aptamers for proteomics studies, we found that the applied SomaScan platform is not interchangeable with validated laboratory instruments for the VTE markers in our study. Therefore, caution is needed when applying SomaScan measurements for hypothesis driven VTE studies using these markers. Whether this is also true for other biomarkers for VTE remains to be determined. It is clear that more studies of agreement with larger sample size and additional markers are needed.

6 FUNDING

This project has been sponsored by the Leiden University Fund / Nypels van der Zee fonds. THE-VTE study was supported by the Netherlands Organization for Scientific Research (NWO) – grant number 916.56.157 (VENI).]. Tariq Faquih was supported by the King Abdullah Scholarship Program and King Faisal Specialist Hospital & Research Centre [No. 1012879283].

7 REFERENCES

1. Heit JA. Epidemiology of venous thromboembolism. Nat Rev Cardiol. 2015;12(8):464-74.
2. Naess IA, Christiansen SC, Romundstad P, Cannegieter SC, Rosendaal FR, Hammerstrom J. Incidence and mortality of venous thrombosis: a population-based study. J Thromb Haemost. 2007;5(4):692-9.
3. Halaby R, Popma CJ, Cohen A, Chi G, Zacarkim MR, Romero G, et al. D-Dimer elevation and adverse outcomes. Journal of thrombosis and thrombolysis. 2015;39(1):55-9.
4. Konstantinides SV, Meyer G, Becattini C, Bueno H, Geersing G-J, Harjola V-P, et al. 2019 ESC Guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS): The Task Force for the diagnosis and management of acute pulmonary embolism of the European Society of Cardiology (ESC). European Heart Journal. 2019;41(4):543-603.
5. Cristea IM, Gaskell SJ, Whetton AD. Proteomics techniques and their application to hematology. Blood. 2004;103(10):3624-34.
6. Pabinger I, Ay C. Biomarkers and Venous Thromboembolism. 2009;29(3):332-6.
7. Gold L, Ayers D, Bertino J, Bock C, Bock A, Brody EN, et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. PLoS One. 2010;5(12):e15004.
8. Suhre K, Arnold M, Bhagwat AM, Cotton RJ, Engelke R, Raffler J, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat Commun. 2017;8:14357.
9. Ganz P, Heidecker B, Hveem K, Jonasson C, Kato S, Segal MR, et al. Development and Validation of a Protein-Based Risk Score for Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease. JAMA. 2016;315(23):2532-41.
10. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. Nature. 2018;558(7708):73-9.
11. van Hylckama Vlieg A, Baglin CA, Luddington R, MacDonald S, Rosendaal FR, Baglin TP. The risk of a first and a recurrent venous thrombosis associated with an elevated D-dimer level and an elevated thrombin potential: results of the THE-VTE study. J Thromb Haemost. 2015;13(9):1642-52.
12. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. 2003;22(1):85-93.
13. Giavarina D. Understanding Bland Altman analysis. Biochemia medica. 2015;25(2):141-51.
14. Bland JM. Interpreting the limits of agreement: do I have good or bad agreement? 2009 [updated 2009 March 202021 January 11]. Available from: https://www-users.york.ac.uk/~mb55/meas/interlim.htm.
15. Kattula S, Byrnes JR, Wolberg AS. Fibrinogen and Fibrin in Hemostasis and Thrombosis. Arteriosclerosis, thrombosis, and vascular biology. 2017;37(3):e13-e21.
16. Tala JA, Polikoff LA, Pinto MG, Li S, Trakas E, Miksa M, et al. Protein biomarkers for incident deep venous thrombosis in critically ill adolescents: An exploratory study. 2020;67(4):e28159.
17. Joshi A, Mayr M. In Aptamers They Trust: The Caveats of the SOMAscan Biomarker Discovery Platform from SomaLogic. Circulation. 2018;138(22):2482-5.
18. Justine B, Thomas S, Bruno P, Marc GB, Anne-Françoise S-S, Aurélien L. Évaluation des performances de l'automate STA R Max®(Stago) pour les paramètres d'hémostase de routine. Annales de Biologie Clinique. 2018;76(2):143-9.
19. Flanders MM, Crist R, Safapour S, Rodgers GM. Evaluation and Performance Characteristics of the STA-R Coagulation Analyzer. Clinical Chemistry. 2002;48(9):1622-4.
20. Cupaiolo R, Govaerts D, Blauwaert M, Cauchie P. Performance evaluation of a new Stago(®) automated haemostasis analyser: The STA R Max(®) 2. Int J Lab Hematol. 2019;41(6):731-7.
21. Bland JM. How can I decide the sample size for a study of agreement between two methods of measurement? 2004 [updated 2004 January 122021 January 1]. Available from: https://www-users.york.ac.uk/~mb55/meas/sizemeth.htm.
22. Hubel A, Spindler R, Skubitz AP. Storage of human biospecimens: selection of the optimal storage temperature. Biopreservation and biobanking. 2014;12(3):165-75.

23. Tworoger SS, Hankinson SE. Collection, processing, and storage of biological samples in epidemiologic studies: sex hormones, carotenoids, inflammatory markers, and proteomics as examples. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2006;15(9):1578-81.
24. Wagner-Golbs A, Neuber S, Kamlage B, Christiansen N, Bethan B, Rennefahrt U, et al. Effects of Long-Term Storage at -80 °C on the Human Plasma Metabolome. Metabolites. 2019;9(5).

Chapter 3

A Workflow for Missing Values Imputation of Untargeted Metabolomics Data

Tariq Faquih [1], Maarten van Smeden [2], Jiao Luo [1], Saskia le Cessie [1,3], Gabi Kastenmüller [4,5], Jan Krumsiek [6], Raymond Noordam [7], Diana van Heemst [7], Frits R. Rosendaal [1],Astrid van Hylckama Vlieg [1], Ko Willems van Dijk [8,9,10] and Dennis O. Mook-Kanamori [1,11,12,*]

[1] Department of Clinical Epidemiology, Leiden University Medical Center, Postal Zone C7-P, PO Box 9600, 2300 RC Leiden, The Netherlands; T.O.Faquih@lumc.nl (T.F.); J.Luo@lumc.nl (J.L.); S.le_Cessie@lumc.nl (S.l.C.); F.R.Rosendaal@lumc.nl (F.R.R.); A.van_Hylckama_Vlieg@lumc.nl (A.v.H.V.)

[2] Julius Center for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht University, 8, 3584 Utrecht, The Netherlands; M.vanSmeden@umcutrecht.nl

[3] Department of Biomedical Data Sciences, Section Medical Statistics and Bioinformatics, Leiden University Medical Center, 2, 2333 Leiden, The Netherlands

[4] Institute of Bioinformatics and Systems Biology, Helmholtz-Zentrum München, 85764 Neuherberg, Germany; g.kastenmueller@helmholtz-muenchen.de

[5] Institute of Experimental Genetics, Genome Analysis Center, Helmholtz Zentrum München, 85764 Neuherberg, Germany

[6] Department of Physiology, Institute for Computational Biomedicine, Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY 10065, USA; jak2043@med.cornell.edu

[7] Department of Internal Medicine, Section of Gerontology and Geriatrics, Leiden University Medical Center, 2333ZA Leiden, The Netherlands; R.Noordam@lumc.nl (R.N.); D.van_Heemst@lumc.nl (D.v.H.)

[8] Einthoven Laboratory for Experimental Vascular Medicine, Leiden University Medical Center, 2, 2333 Leiden, The Netherlands; K.Willems_van_Dijk@lumc.nl

[9] Department of Internal Medicine, Division of Endocrinology, Leiden University Medical Center, 2, 2333 Leiden, The Netherlands

[10] Department of Human Genetics, Leiden University Medical Center, 2, 2333 Leiden, The Netherlands

[11] Department of Public Health and Primary Care, Leiden University Medical Center, 2, 233 Leiden, The Netherlands

[12] Metabolon Inc., Morrisville, NC 27560, USA

* Correspondence: D.O.Mook@lumc.nl

Metabolites. **2020**. 10:486. doi:10.3390/metabo10120486 PMID: 33256233

1 ABSTRACT:

Metabolomics studies have seen a steady growth due to the development and implementation of affordable and high-quality metabolomics platforms. In large metabolite panels, measurement values are frequently missing and, if neglected or sub-optimally imputed, can cause biased study results. We provided a publicly available, user-friendly *R* script to streamline the imputation of missing endogenous, unannotated, and xenobiotic metabolites. We evaluated the multivariate imputation by chained equations (MICE) and k-nearest neighbors (kNN) analyses implemented in our script by simulations using measured metabolites data from the Netherlands Epidemiology of Obesity (NEO) study (n = 599). We simulated missing values in four unique metabolites from different pathways with different correlation structures in three sample sizes (599, 150, 50) with three missing percentages (15%, 30%, 60%), and using two missing mechanisms (completely at random and not at random). Based on the simulations, we found that for MICE, larger sample size was the primary factor decreasing bias and error. For kNN, the primary factor reducing bias and error was the metabolite correlation with its predictor metabolites. MICE provided consistently higher performance measures particularly for larger datasets (n > 50). In conclusion, we presented an imputation workflow in a publicly available *R* script to impute untargeted metabolomics data. Our simulations provided insight into the effects of sample size, percentage missing, and correlation structure on the accuracy of the two imputation methods.

Keywords: imputation; multiple imputation using chained equations; k-nearest neighbors; untargeted metabolomics; metabolon; simulation; workflow

2 INTRODUCTION

Metabolomics studies have seen a steady growth due to the development and implementation of affordable and high-quality metabolomics platforms. These platforms can be split into two categories: targeted and untargeted metabolomics platforms based on their approach to metabolite identification [1,2]. Targeted platforms are focused on a known prespecified set of metabolites, while untargeted platforms aim to detect as many metabolites as possible in the sample without the need for explicit prior knowledge of their identity. The metabolite signatures detected (i.e., mass to charge ratio, m/z, or retention times) are subsequently matched in a metabolite library to determine their identity. Currently, both targeted and untargeted platforms can detect over 1000 metabolites in a single biological sample (e.g., blood, saliva, and urine). A typical issue with both these platform types is missing values from the measurement.

Missing values in metabolomics data are problematic for subsequent analyses, may be neglected, and are often mishandled or ignored. A common misconception is that missing values in metabolomics data are exclusively due to metabolites with a very low concentration, i.e., below the limit of detection of the instrument. Although in many circumstances the majority of missing values can be due to low concentrations, it has been shown that missing values can also be caused by biological and/or technical variation [3–5]. Based on the assumption that not reaching the limit of detection exclusively causes missingness, missing values are often handled with one or more of the following procedures:

For each metabolite the missing values are replaced (“imputed”) with a single value, such as the minimum detection level or half the minimum detection level. This approach results in overrepresentation of a single value in the population distribution. This may affect subsequent analyses and may cause biased results, regardless of the cause of missing values [5,6]. Furthermore, metabolites could be missing in some individuals because they are not biologically present in their

system. Therefore, imputing these missing values will cause bias in the analysis. For example, if the metabolites for metformin are imputed, both diabetic patients who use the drug as well as and non-diabetic individuals who do not use the drug will have values for the it. This is a prominent issue in platforms such as Metabolon™ (Metabolon Inc., Durham, NC, USA) that include xenobiotic metabolites (e.g., metabolites from external sources such as medications).

Metabolites with a missing percentage above an arbitrary cut-off value (for example 20%) are removed from the dataset due to "too much missingness" regardless of the metabolite identity. By applying a cut-off above which metabolites are removed from the dataset, or, in the most extreme case only using the complete cases, data are unnecessarily discarded, that could have been of importance to the research question. Furthermore, this exclusion can affect further pathway analysis, such as metabolite set enrichment analysis, that explore possible pathway connections for the measured metabolites [7].

Several studies have evaluated imputation methods for metabolomics data. The consensus from these studies has so far been that imputation using half the minimum value leads to more bias than other methods and, consequently, this method is discouraged [3,8]. One alternative imputation method that has been recommended for metabolomics is the k-nearest neighbors (kNN) imputation [6,9]. An extensive simulation was performed that evaluated and compared 31 methods of imputation in a simulated untargeted metabolomics data provided by the Metabolon™ platform [6]

These methods included univariate methods such as half-minimum imputation and multivariate methods such as variations of kNN and multivariate imputation by chained equations (MICE). Two methods were concluded to have the best performance:

kNN on observations with variable pre-selection ("kNN-obs-sel"), a two-step method that incorporates the standard kNN algorithm with a preselection of a group of metabolites that are most correlated with the metabolite with missing values (i.e., auxiliary metabolites). Therefore, the neighbors selected by kNN will have similar metabolomic profiles [6].

MICE using the predictive mean matching method ("MICE-pmm"). Like kNN-obs-sel, the most correlated metabolites were used for the imputation. The imputed values are then selected from distribution of possible values to produce multiple imputed datasets [10–12].

In this paper, we expand upon the meticulous evaluation of the imputation methods by Do et al. [6], which was performed on an older version of the metabolomics platform that detects a smaller set of metabolites (*n* = 517). Furthermore, we set out to take unannotated (i.e., unidentified metabolites in the library) and xenobiotic metabolites into account. The recent Metabolon™ panel in use (Discovery HD4) has increased the number of metabolites to >1000, which includes more unannotated and xenobiotic metabolites. As more scientists are using metabolomics data in their research, it is helpful to have a user-friendly workflow for imputation using the best available methods. We provided this imputation workflow and a user-friendly *R* script to streamline the imputation of the Metabolon™ HD4 panel using kNN-obs-sel and MICE-pmm. Furthermore, we evaluated the imputations by the script in several scenarios with different missingness conditions by a resampling simulation analysis using measured metabolomics data from the Netherlands Epidemiology of Obesity (NEO) study.

3 RESULTS

3.1 Metabolomic Data Characteristics

Metabolomics measurements in 599 individuals between the ages of 45 and 65 with normal BMI distribution from the NEO study identified 1365 metabolites. Detailed information regarding the population are provided in the methods section and Appendix A. Known metabolites were annotated with their chemical name, super pathway, sub pathway, compound identifiers from various metabolite databases, and information regarding their biochemical properties. A total of 840 metabolites were from various endogenous pathways, 229 metabolites were characterized as xenobiotics, and 296 metabolites were unannotated (lacking information regarding chemical name and pathway). Of the 1365 identified metabolites, 800 (58.6%) contained missing values and the median number of missing metabolites per observation was 228 (38%) (Table 1).

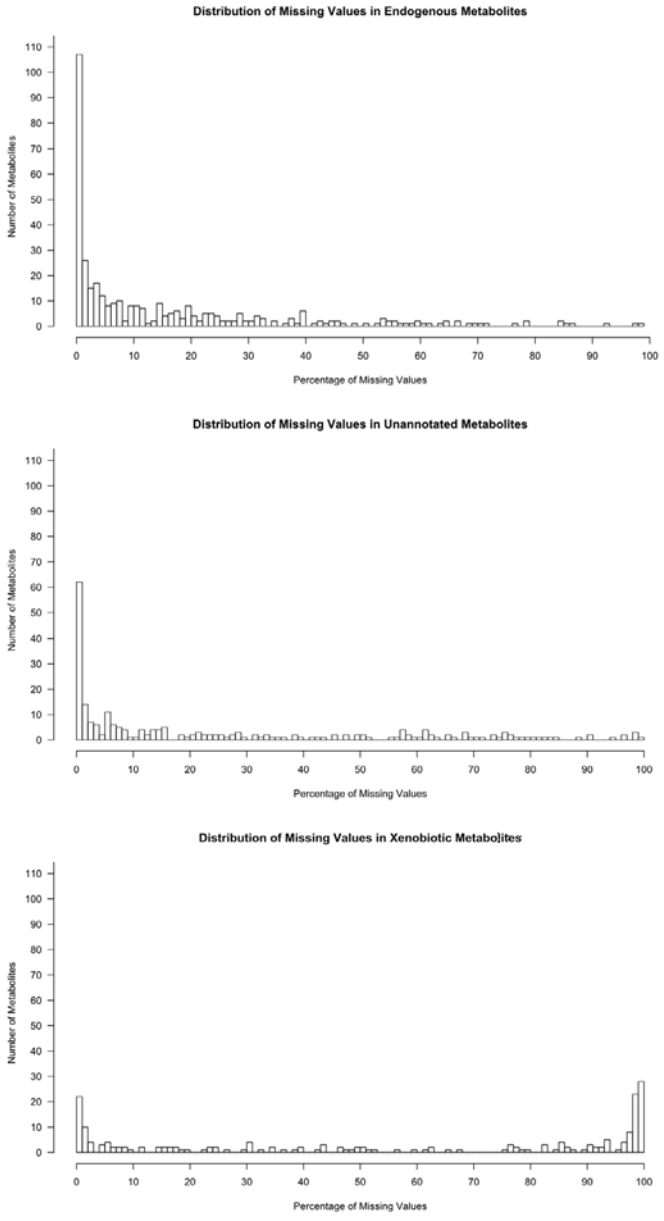
Table 1. Summary of missing data in the Netherlands Epidemiology of Obesity (NEO) study.

Missing Data	Metabolite Groups			
	Endogenous (*n* = 840)	Unannotated (*n* = 296)	Xenobiotics (*n* = 229)	Total (*n* = 1365)
Metabolites with missing values, *n* (%)	367 (43.7)	236 (79.7)	197 (86.0)	800 (58.6)
Missing metabolites per observation, median (range)	57 (23–94)	59 (31–112)	110 (79–149)	228 (152–343)

In the NEO study, 1365 metabolites were measured in 599 individuals (observations).

We plotted the distribution of missing values in each metabolite group (Figure 1). The distribution of the number of missing values of the unannotated metabolites was similar to that of the endogenous metabolites rather than the xenobiotic metabolites. This suggests that most unannotated metabolites are most likely from an endogenous source, similar to the annotated endogenous metabolites, and are most likely expected to be present in all our participants.

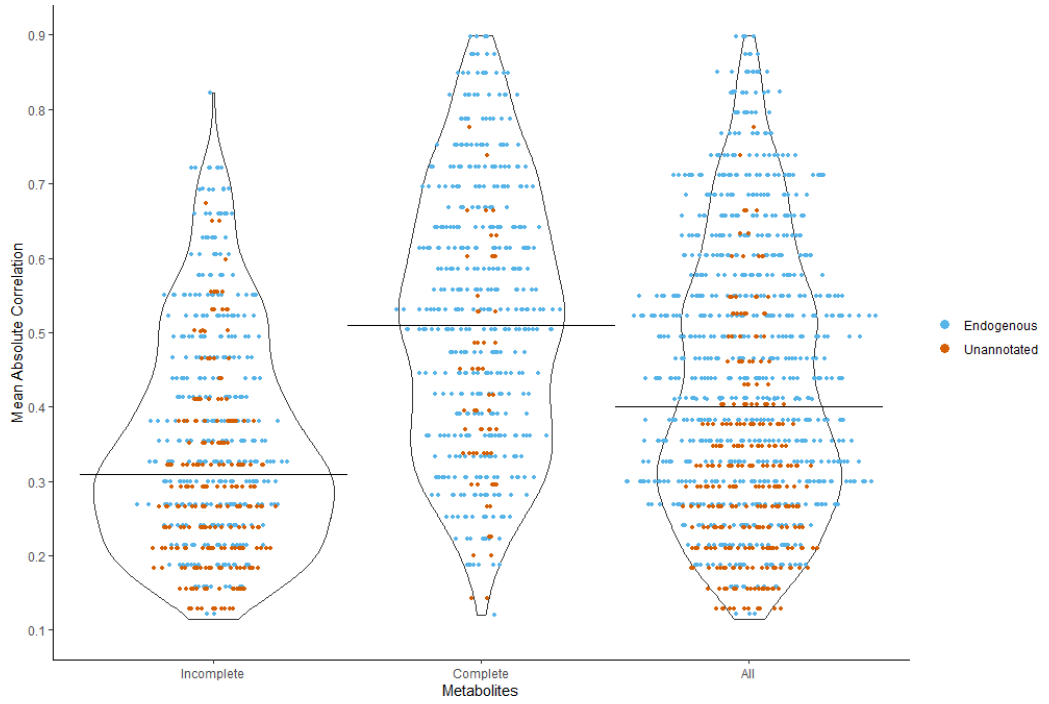
Figure 1. Distribution of the missing values in each metabolite group.



The Pearson's pairwise-complete correlation matrix for the endogenous and unannotated metabolites was calculated using all the metabolites (complete with no missing values and incomplete). For each incomplete metabolite we selected up to 10 complete metabolites with the highest absolute Pearson's correlation (auxiliary metabolites). If the metabolite was not correlated with 10 metabolites (due to high missingness), then we selected the available correlated metabolites. We then calculated the mean value of the Pearson's correlations for these metabolites. Figure 2 shows the distribution of the mean of the auxiliary absolute correlations with further details in Table B1.

The 82% of the incomplete metabolites had a mean absolute Pearson's correlation coefficient lower than 0.5 with their auxiliary metabolites. Overall, the median of the median absolute Pearson's correlation coefficient was 0.4 (0.09–0.89), indicating a generally low intercorrelation between the metabolites.

Figure 2. Distribution of the mean absolute correlations for the complete (without missing values) and incomplete (with missing values) endogenous and unannotated metabolites in the NEO dataset.



3.2 Availability

The imputation script [13] streamlines the workflow by calculating the correlation matrix, selecting the auxiliary metabolites, and imputing the missing values of the metabolites using the provided data from the user. The script requires a dataset, a list of xenobiotic and non-xenobiotic metabolites (endogenous/unannotated), and a choice for the method of imputation (MICE-pmm or kNN-obs-sel). The script and example files can be found at: https://github.com/tofaquih/imputation_of_untargeted_metabolites.

3.3 Performance Evaluation

To evaluate our imputation framework, we applied it to impute metabolites with missing values in the measured NEO dataset (*n* = 599) using kNN-obs-sel and MICE-pmm. All metabolites were imputed apart from 12 metabolites (3 endogenous, 9 unannotated) in the dataset that had >90% missingness and were subsequently treated as xenobiotic and imputed to 0. As mentioned in the methods section, extremely high missingness limits the amount of data needed to impute the metabolites and to find auxiliary metabolites. High missingness in the 3 endogenous metabolites could have been caused by technical or biological issues, or they

could represent misannotated xenobiotic metabolites. The 9 unannotated metabolites were likely xenobiotic metabolites.

Simulations were performed to compare the performance of the imputation method (MICE- pmm or kNN-obs-sel). As detailed in the Methods section, we generated 144 resampling simulation scenarios, using four metabolites from independent pathways and varying mean correlations with auxiliary metabolites (PC(32:2) (mean absolute correlation = 0.64), urate (mean absolute correlation = 0.49), glutamate (mean absolute correlation = 0.49), succinylcarnitine (mean absolute correlation =0.36)), three sample sizes (50, 150, 599), three percentage of missing (15%, 30%, 60%), and two missing mechanisms missing mechanisms (missing completely at random (MCAR) and probabilistic limit of detection (PLoD)). The percentage biases from the simulation are presented in Figure 3 and Table 2. Root mean squared errors (RMSE) are shown in Figure 4, Table B5, and Table B6. The mean and standard deviation of the estimates from the simulation are provided in Table B3 and Table B4 using MCAR and PLoD mechanisms, respectively. We used nested loop plots [14] to produce all the figures.

Figure 3. Nested loop plot of the percentage bias of the four metabolites from the simulation. The horizontal axis in each box represents the missing percentage and is split per sample size. Abbreviations: MCAR: missing completely at random; PLoD: probabilistic limit of detection.

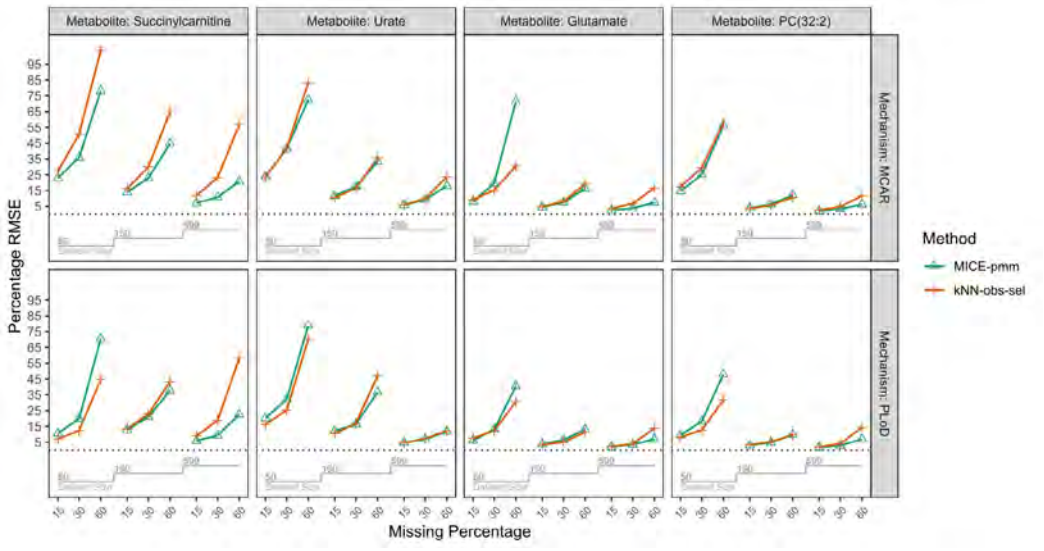


Figure 4. Nested loop plot of the root mean squared error (RMSE) of the four metabolites from the simulation. To simplify comparability in the plot we converted the RMSE values to a percentage by subtracting then dividing the RMSE values by the corresponding true estimates (in sample size n = 599). The horizonal axis in each box represents the missing percentage and is split per sample size. Abbreviations: MCAR: missing completely at random; PLoD: probabilistic limit of detection.

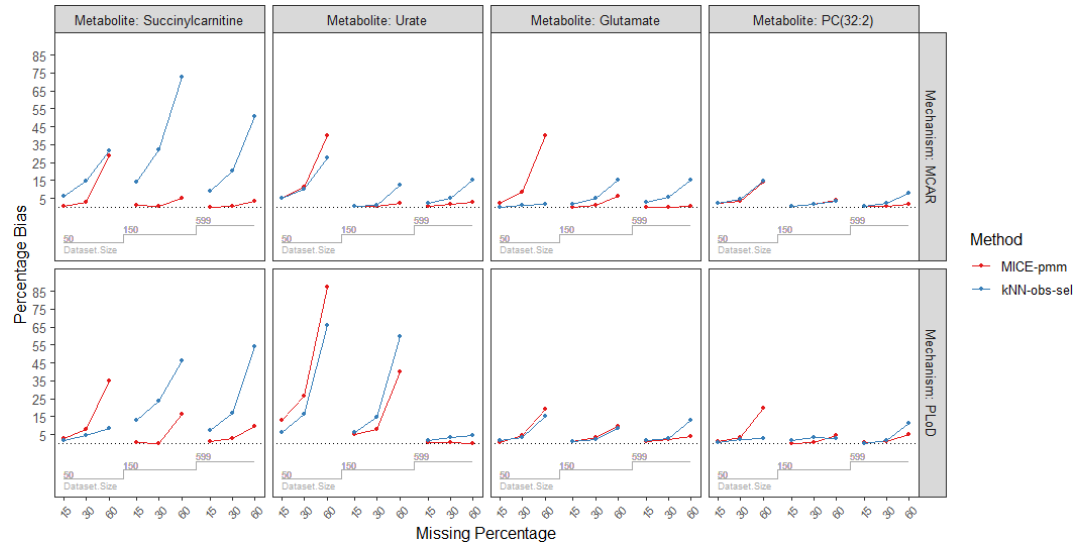


Table 2. Percentage bias of the imputation methods across different parameters on different metabolites including multivariate imputation by chained equations (MICE)-pmm with a single imputation.

Missing Mechanism	Sample Size	Missing Percentage	PC(32:2)		Metabolites/Imputation Method				Urate	
			MICE-pmm	kNN-obs-sel	Succinylcarnitine		Glutamate		MICE-pmm	kNN-obs-sel
					MICE-pmm	kNN-obs-sel	MICE-pmm	kNN-obs-sel		
MCAR	*n* = 50	15%	2.0	2.1	0.3	6.0	2.2	0.2	4.9	4.9
		30%	3.2	4.5	2.8	14.4	8.6	1.1	11.3	10.2
		60%	13.9	14.5	28.9	31.3	40.2	2.0	39.9	27.6
	n = 150	15%	0.7	0.5	1.4	13.8	0.2	1.9	0.7	0.5
		30%	1.9	1.8	0.4	31.9	1.1	5.1	0.7	1.3
		60%	3.9	3.2	5.2	72.7	6.2	15.0	2.09	12.5
	n = 599	15%	0.3	0.8	0.1	9.3	0.2	2.6	0.6	2.2
		30%	0.7	2.4	0.6	20.1	0.1	5.6	1.5	4.9
		60%	1.9	7.9	3.5	50.9	0.7	15.1	2.6	15.0
PLoD	*n* = 50	15%	1.3	0.8	2.8	1.8	0.7	1.8	12.7	6.2
		30%	3.5	2.2	7.7	4.3	4.2	3.6	26.3	16.5
		60%	19.5	3.0	34.8	8.3	19.3	15.0	87.4	66.1
	n = 150	15%	0.2	1.6	0.7	12.8	1.3	0.8	5.3	6.3
		30%	0.3	3.3	0.2	23.4	3.0	2.3	8.0	14.9
		60%	4.2	2.6	16.2	46.1	9.4	8.6	39.8	59.7
	n = 599	15%	0.5	0.1	0.9	7.3	0.9	1.4	0.5	1.9
		30%	1.1	1.6	2.5	16.7	2.1	2.9	0.6	3.3
		60%	4.9	11.5	9.7	54.3	4.1	13.1	0.1	4.3

Abbreviations: MCAR: missing completely at random; PLoD: probabilistic limit of detection.

3.3.1 Sample Size

We observed a decrease of bias and RMSE as the sample size increased (Figures 3 and 4). This trend was consistent for MICE-pmm for each metabolite, with the percentage bias median (range):

8.2 (0.3–87.4) in *n* = 50 decreasing to median (range): 0.8 (0.1–9.7) in *n* = 599. However, increasing sample size did not improve imputation with kNN-obs-sel. Overall, percentage bias was median (range): 4.7 (0.2–66.1) in *n* = 50, median (range): 5.7 (0.5–72.7) in *n* = 150, and median (range): 5.2 (0.1– 54.3). Furthermore, in some scenarios, bias and RMSE increased in larger sample sizes even with the same missing percentage and missing mechanism; this was particularly noticeable for succinylcarnitine (mean absolute correlation = 0.36) where the percentage bias increased in 60% missing from 8.3% in *n* = 50, to 46.1% in *n* = 150, and 54.3% in *n* = 599. Finally, for sample sizes of *n* = 50 and 150, MICE-pmm had lower bias than kNN-obs-sel but a RMSE higher or similar to kNN-obs- sel.

3.3.2 Percentage of Missing

In scenarios with 15% and 30% missing, MICE-pmm and kNN-obs-sel showed low bias and RMSE across all sample sizes. At 15% missing, MICE-pmm had a percentage bias of median (range):

0.7 (0.1–12.7), while kNN-obs-sel had a percentage bias of median (range): 1.9 (0.1–13.8). At 30% missing, MICE-pmm had a percentage bias of median (range): 2.0 (0.1–26.2) and kNN-obs-sel had a percentage bias of median (range): 4.4 (1.1–31.9). Finally, in 60% missing MICE-pmm had a percentage bias of median (range): 7.8 (0.1–87.4) and kNN-obs-sel had a percentage bias of median (range): 14.7 (1.9–72.7). Overall, MICE-pmm had lower bias in all missing percentages than kNN-obs- sel. However, the percentage bias for kNN-obs-sel was often lower than that of MICE-pmm at 30% and 60% missing in *n* = 50.

3.3.3 Correlation Strength with the Auxiliary Metabolites

We compared the percentage bias and RMSE of both imputation methods for the four metabolites to assess the influence of correlation strength of the auxiliary metabolites as shown in Tables 2 and B2 and Figures 3 and 4. We observed that availability of auxiliary metabolites with higher correlation for the imputation greatly reduced the bias and RMSE in both methods. In PC(32:2), the metabolite with the highest mean correlation (mean absolute correlation = 0.64), had the lowest bias overall. Percentage bias was median (range): 1.9 (0.2–19.5) with the MICE-pmm imputation and median (range): 2.3 (0.1–14.5) with kNN-obs-sel imputation. Glutamate (mean absolute correlation = 0.49) had median (range): 2.1 (0.1–40.2) percentage bias with MICE-pmm imputation and median (range): 2.7 (0.2–15.1) with kNN-obs-sel. Similarly, imputation of urate (mean absolute correlation = 0.49) using MICE-pmm had median (range): 3.8 (0.1–87.4) percentage bias and median (range): 6.2 (0.5–66.1) using kNN-obs-sel. In contrast, the percentage bias was much higher for the metabolite with the lowest mean correlation, Succinylcarnitine (mean absolute correlation = 0.36), with median (range): 2.6 (0.1–34.8) percentage bias using MICE-pmm imputation and median (range): 15.5 (1.8–72.7) with kNN-obs-sel. Moreover, the bias reached very high percentages in urate and succinylcarnitine compared to PC(32:2) and glutamate in the *n* = 50 subset.

3.3.4 Missing Mechanisms

We used two mechanisms for missingness, MCAR and PLoD, in our simulations. Since PLoD is fundamentally missing not at random (MNAR), causing lower concentrations to have a higher likelihood of missingness, we examined how PLoD affects the performance of MICE-pmm and kNN-obs-sel compared to MCAR scenarios. MCAR scenarios had a percentage bias median (range): 1.9 (0.1–40.2) with MICE-pmm imputation and median (range): 5.3 (0.2–72.7) with kNN-obs-sel. PLoD scenarios had a percentage bias median (range): 3.2 (~0–87.4) with MICE-pmm imputation and median (range): 4.3 (0.1–66.1) with kNN-obs-sel. However, the RMSE (Figure 4, Tables B5 and B6) was lower in PLoD for MICE-pmm (median (range): 11.1 (1.8–79)) than in MCAR (median (range):

14.3 (2.0–78.0)) and similarly lower for kNN-obs-sel in PLoD scenarios (median (range): 12.1 (2.1–70.3)) than MCAR (median (range): 16.5 (2.6–103.9)). Overall, imputing in PLoD scenarios lead to higher bias but lower RMSE compared to MCAR.

4 DISCUSSION

Several simulation studies have evaluated different imputation methods for missing data in metabolomic datasets [3,6,9,15,16]. Nevertheless, the "half the minimum" method of imputation remains in use despite studies showing its sub-optimal performance [3,6,9,15,16]. In this study, we followed up on previous work and provided a framework and complementary *R* script on GitHub

[13] that streamlines the imputation of untargeted metabolomics data. The script provides univariate imputation of zero for missing values considered to be truly absent in xenobiotics and two options of multivariate imputation methods for the remaining metabolites.

Overall, for the four metabolites we used in the simulation, we observed several factors that influenced the performance of each imputation method with different degrees. In the four metabolites we used, MICE-pmm performed better overall across different simulated scenarios. This performance is especially better in PLoD, which represents a missing mechanism similar to that of real metabolomics data [6]. MICE-pmm performance decreased the most in smaller sample sizes, somewhat less by the metabolite auxiliary correlation and the least by the missing percentage. Interestingly, the negative effect of missing percentage diminished as the sample size increased (n = 150 and n = 599). On the other hand, unlike MICE-pmm, kNN-obs-sel performance was decreased most by a higher percentage of missingness and low metabolite auxiliary correlation, which was not improved by increased sample size. A possible explanation is the nature of the kNN-obs-sel method. kNN-obs-sel focused on finding the nearest neighbors based on the correlated metabolites. If it failed to find strongly correlated metabolites, due to the metabolite naturally a having low correlation or due to a large amount of missing values, it selected weak neighbors. Therefore, even at larger sample sizes (150 and 599) the performance of the kNN-obs-sel method remained poor if the missing percentage was large and the metabolite had poor correlation.

4.1 Advantages and Disadvantages of MICE-pmm for Metabolomics

Unlike kNN imputations, we found few papers in the literature regarding the use of MICE imputation for metabolomics. The MICE-pmm imputation is a more intricate method for generating the imputation values. First, the imputation is repeated multiple times in order to assess the uncertainty of the imputation and provide standard errors of the estimates. Second, MICE-pmm

imputation is more compatible with both normally distributed and skewed metabolites than kNN [10]. Third, MICE imputation utilizes discrete and continues variables for imputation. Therefore, MICE-pmm can include additional biologically relevant predictors and the outcome of the analysis of interest, improving the quality of the imputation [17]. These features explain the robustness of MICE-pmm in situations with low correlated auxiliary variables and high missingness.

However, MICE-pmm has some disadvantages. First, small sample sizes negatively affected the performance of MICE-pmm because this forces duplication and reuse of the same individuals [10]. Second, MICE imputation may require more computational run time and is somewhat more complicated to use than kNN because multiple imputed datasets are generated that require a pooling step for the analysis. We shortened computational time by using the latest *MICE R* package and by setting the number of multiple imputations to 5, which has been shown to be a suitable number of imputations [10]. This caused the running time for the complete imputation using MICE-pmm to be equal to that of kNN-obs-sel for the NEO dataset (n = 599). Furthermore, to test the speed of the script, we duplicated and stacked the NEO dataset to create larger datasets (n = 5400 and n = 20,000); MICE- pmm completed the imputations faster than kNN-obs-sel (Table B7). Third, with MICE-pmm it is not possible to apply further analysis such as lasso regression or random forest, which are common analysis methods used in metabolomics [7,18,19]. This is because MICE-pmm uses multiple datasets with Rubin's Rules to pool the estimates of the analysis per dataset. One solution is to use the kNN- obs-sel method, as it always creates a single dataset for analysis. A second alternative would be to use MICE-pmm with a single imputation [m = 1], which can be specified in our script, and use that single dataset in the multivariate analysis. It should be noted that MICE-pmm with m = 1 still performed better than kNN-obs-sel for the larger sample sizes (see Tables B2 and B6 and Figures B1 and B2).

4.2 Limitations

Several methodological issues should be considered. Firstly, our evaluation was done using 599 samples, limited by available metabolomics data in the NEO study. Although this number is not particularly small, future research should be performed in larger datasets. Secondly, we assumed that all missing xenobiotics values are truly missing and replaced them by zero. This could be explored further by incorporating MICE-pmm or kNN-obs-sel to specifically impute xenobiotic metabolites from the same medication sources in persons taking the medication. Furthermore, it could be possible to use questionnaire and clinical data as imputation predictors in MICE-pmm to impute related xenobiotic metabolites. Thirdly, we did not explore alternative methods for MICE to handle small data sizes, such as regulation and penalization. Fourth, our simulation did not evaluate the variance estimators such as type-I and type-II errors or confidence interval coverage. Fifth, metabolites with very large missingness will have high bias and error in the imputation and should be interpreted with caution. Finally, the data do not provide the explicit cause of the missing values and, therefore, we could only assume if the values were truly missing, missing completely at random, or missing due to other reasons. Future studies which explore the causes of missingness will also allow us to impute the missingness more effectively.

5 MATERIALS AND METHODS

5.1 Population Characteristics

The resampling simulation analyses were performed in the NEO study. This study has been extensively described elsewhere [20] and in Appendix A. The NEO study was accepted by the Medical Ethics committee of the Leiden University Medical Center under protocol P08.109. The study is also registered at clinicaltrials.gov under number NL21981.058.08 / P08.109. All participants gave written informed consent [20]. Fasting state serum samples from a sub-population (*n* = 599) of the NEO study were sent for untargeted metabolomics measurements at Metabolon Inc. (Durham, NC, USA) using their Metabolon™ Discovery HD4 platform. In brief, this process involves four independent ultra-high-performance liquid chromatography mass spectrometry (UHPLC-MS/MS) platforms [21,22]. Two platforms used positive ionization reverse phase chromatography, one used negative ionization reverse phase chromatography, and one used hydrophilic interaction liquid chromatography (HILIC) negative ionization [22]. In total, 1365 serum metabolites were measured which included 840 endogenous, 296 unannotated, and 229 xenobiotic metabolites.

5.2 Imputation Methods

Following our examination of the missing data distribution in the NEO study (Figure 2), we decided the xenobiotic metabolites and non-xenobiotic metabolites (endogenous/unannotated) with different imputations. For xenobiotic metabolites, we assumed missing values are truly missing values. For example, when a medication metabolite concentration is missing, it is most likely that the participant is not taking the medication. Therefore, we decided to impute xenobiotic metabolites to zero, as imputing the values (with MICE, kNN, or half-min) would cause bias due to skewed distribution and false positives. For the non-xenobiotic metabolites (endogenous/unannotated), the missing pattern suggests that the unannotated metabolites are most likely endogenous. Therefore, we decided to impute the endogenous and the unannotated metabolites as a single group using the multivariate imputation methods of MICE-pmm and kNN-obs-sel. For these two multivariate methods, we first estimated a correlation matrix for all applicable/non-xenobiotic metabolites from which to select 10 auxiliary metabolites to be used for imputation.

For non-xenobiotic metabolites, we assumed that they are metabolites with truly missing values only if less than 90% of values were missing. This cut-off was necessary for multiple reasons: 1) it became nearly impossible to find auxiliary metabolites for imputation, 2) unannotated metabolites with high missing values are likely xenobiotic and therefore most likely truly missing, and 3) it became statistically problematic to perform multivariate imputation with such high missingness— particularly in small sample sizes [23].

In this study, we used MICE-pmm with 10 auxiliary metabolites to impute the missing values and generated 5 imputed datasets (m = 5). In addition to the auxiliary metabolites, we included further predictors by adding the clinical variables for the outcome (BMI) and the covariates (age and sex) used in the analysis model for the MICE-pmm imputation. The addition of these variables is required in MICE imputations to avoid bias in the results [23,24]. We used kNN-obs-sel only with 10 auxiliary metabolites to impute the missing values. Details regarding the imputation methods are provided in Appendix A. In our script, we incorporated the *R* package *mice* version 3.6.0 [10] for the MICE-pmm imputations and the package *VIM* version 4.8.0 [25] in the kNN-obs-sel imputations.

5.3 Evaluation Analysis and Missing Value Simulation

For the simulation, the analysis of interest was an ordinary least squares regression model with body mass index (BMI) as the outcome and age, sex, and a selected metabolite as the exposures. For the purpose of our study, BMI was used as the outcome for two reasons: (1) BMI is a variable that was measured in all our participants, and (2) BMI is strongly associated with many metabolites and commonly studied in metabolomics [26].

Four metabolites were used, selected based on the following criteria: (1) the metabolite had no missing values in the original NEO dataset, (2) the metabolite must have a strong association with BMI in our Metabolon™ data as well as in the literature using Metabolon™ [26], (3) the four metabolites must be from different biological pathways, and (4) the metabolites must have different mean correlations with their auxiliary metabolites. We found 6 out of 473 complete endogenous metabolites in NEO that fulfilled these criteria. We then narrowed the selection to one metabolite per pathway. Accordingly, we selected four metabolites: PC(32:2) (mean absolute correlation 0.64) from the lipid super pathway; succinylcarnitine (mean absolute correlation = 0.36) from the energy super pathway, the nucleotide urate (mean absolute correlation 0.49), and the amino acid glutamate (mean absolute correlation 0.49). Information regarding the metabolites is provided in Table 3.

Table 3

Metabolite Full Name	Mean Absolute Correlation	Super Pathway	Sub Pathway	Estimate *n* = 599	Estimate *n* = 150	Estimate *n* = 50
PC(32:2)	0.64	Lipid	Plasmalogen	-4.18×10^{-7}	-3.64×10^{-7}	-4.38×10^{-7}
Urate	0.49	Nucleotide	Purine Metabolism	1.39×10^{-8}	9.58×10^{-9}	9.69×10^{-9}
Glutamate	0.49	Amino Acid	Glutamate Metabolism	1.83×10^{-7}	2.89×10^{-8}	1.66×10^{-8}
Succinylcarnitine	0.36	Energy	TCA Cycle	2.84×10^{-6}	1.53×10^{-6}	4.53×10^{-6}

Abbreviations. Mean absolute correlation: mean of the 10 absolute Pearson's correlations from the metabolite correlation matrix. Estimate is the regression coefficient from the model
BMI~age + sex + metabolite. Therefore, the estimates are the mean increase in BMI per 1 unit increase of the metabolite.

We compared the performance of the two imputation methods by simulating missing values using the NEO dataset (*n* = 599). All simulations were performed on three datasets: the original dataset of 599 participants, and on two randomly sampled sub datasets of size *n* = 150 and *n* = 50. The distribution of age, sex, and BMI was maintained in the sub datasets of 50 and 150 individuals. We used the same sub datasets for the all corresponding simulation scenarios. Generating the subsets with different random sampling did not change the estimates drastically (not shown). It should bepointed out that the selected auxiliary metabolites differed slightly between the sub datasets. Metabolite levels were log transformed and standardized (mean of 0 and variance of 1). We calculated the estimates for each metabolite in the complete datasets separately to be used later for the bias and RMSE calculations. In the different simulation scenarios, we induced different percentages of missingness (15, 30, and 60%), and under two different mechanisms, MCAR and PLoD. In the PLoD missing scenarios, the odds of a value being missing increased as the concentration decreases. The total number of missing values was divided per quantile of the

metabolite as follows: 40% into the lower quantile, 50% into the middle quantile, and 10% in the upper quantile.

The evaluation was done by (1) performing the linear regression analysis and obtaining the estimate of the regression coefficient using the complete metabolites data in each subset (Table 3), (2) simulating missing values, (3) imputing missing values using the two imputation methods, (4) estimating the regression coefficient using the imputed data, and (5) evaluating the difference between the estimate of the complete data for that subset and the estimate using the imputed methods, (6) repeating step 2 to 5 1000 times per simulation scenario. The performance of the imputation methods was evaluated using the following measures: raw bias, which is the difference between the real estimate and the mean of the simulations estimates, which can be a positive or a negative value; percentage bias, which is the raw bias divided by real estimate for easier interpretation and comparison [27]; the RMSE, which is the square root of the mean squared difference between estimated; and true value, this measure combines the bias and variance of the simulated estimates into a single measure and represents the precision of the method [28] (Appendix A).

Thus, in total, we used three datasets (n = 50, 150, 599), four metabolites (PC(32:2), succinylcarnitine, urate, glutamate), three missingness percentages (15%, 30%, 60%), two missing mechanisms (MCAR and PLoD), and evaluation by two imputation methods (kNN-obs-sel and MICE-pmm) for a total of 144 possible scenarios. Each of these scenarios was repeated 1000 times.

5.4 Imputation Workflow

To simplify the procedure of imputing missing data, we wrote an *R* script that calculates the correlation matrix between the different metabolites, selects the auxiliary metabolites with the largest correlation, imputes the xenobiotic metabolites with univariate imputation, and imputes the endogenous metabolites with a multivariate imputation (either kNN-obs-sel or MICE-pmm), which can be found on our GitHub repository [13].

6 CONCLUSIONS

In conclusion, we provided a workflow for handling missing values in untargeted metabolomics data using univariate imputation for xenobiotics and multivariate imputation using MICE-pmm or kNN-obs-sel for endogenous and unannotated metabolites. We further evaluated MICE-pmm and kNN-obs-sel in different simulated scenarios. Our evaluation showed that the performance of both methods is affected by three different factors, namely the metabolite mean correlation with auxiliary metabolites, the sample size, and the missing percentage. For MICE-pmm, sample size was the primary factor affecting bias and error inversely. For kNN-obs-sel, the primary factor affecting bias and RMSE was the metabolite correlation with the predictors, which, when high, can provide low bias and RMSE even in small sample sizes (n = 50). Since most of our metabolites had low mean correlation, MICE-pmm provided consistently higher performance measures than kNN-obs-sel and, as a result, we suggest using MICE-pmm imputation for untargeted metabolomics, particularly for larger datasets (n > 50).

Author Contributions: Conceptualization, T.F., A.v.H.V., K.W.v.D., D.O.M.-K., G.K. and J.K.; formal analysis, T.F.; software, T.F.; funding acquisition, D.v.H., R.N., K.W.v.D. and F.R.R.; methodology, S.l.C., M.v.S. and T.F.;

resources, R.N. and F.R.R.; supervision A.v.H.V., K.W.v.D., D.O.M.-K. and M.v.S.; validation, J.L.; writing— original draft, T.F.; writing—review & editing, all authors. All authors have read and agreed to the published version of the manuscript.

Funding: The NEO study is supported by the participating Departments, the Division and the Board of Directors of the Leiden University Medical Centre, and by the Leiden University, Research Profile Area 'Vascular and Regenerative Medicine'. The analyses of metabolites are funded by the VENI grant (ZonMW-VENI Grant 916.14.023) of D.O.M.-K. D.v.H. and R.N. were supported by a grant of the VELUX Stiftung [grant number 1156].

J.L. was supported by the China Scholarship Counsel [No. 201808500155]. T.F. was supported by the King Abdullah Scholarship Program and King Faisal Specialist Hospital & Research Center [No. 1012879283].

Acknowledgments: The authors of the NEO study thank all individuals who participated in the Netherlands Epidemiology in Obesity study, all participating general practitioners for inviting eligible participants, and all research nurses for collection of the data. We thank the NEO study group, Pat van Beelen, Petra Noordijk and Ingeborg de Jonge for the coordination, lab, and data management of the NEO study.

Conflicts of Interest: Dennis Mook-Kanamori is a part-time clinical research consultant for Metabolon, Inc. All other co-authors have no conflicts of interest to declare.

7 APPENDIX A

A1. NEO Study Design

The Netherlands Epidemiology of Obesity (NEO) study is a population-based, prospective cohort study of individuals aged 45–65 years, with an oversampling of overweight individuals or individuals with obesity. Men and women aged between 45 and 65 years with a self-reported BMI of 27 kg/m^2 or higher, living in the greater area of Leiden (in the West of the Netherlands) were eligible to participate in the NEO study. In addition, all inhabitants aged between 45 and 65 years from one municipality (Leiderdorp) were invited, irrespective of their BMI. Recruitment of participants started in September 2008 and was completed at the end of September 2012. In total, 6671participants were included, of whom 5217 had a BMI of 27 kg/m^2 or higher. The NEO study was accepted by the Medical Ethics committee of the Leiden University Medical Center under protocol P08.109. The study is also registered at clinicaltrials.gov under number NL21981.058.08 / P08.109. All participants gave written informed consent. Participants were invited to come to the NEO study center of the LUMC for one baseline study visit after an overnight fast. A blood sample of 108 mL was taken from the participants after an overnight fast of at least 10 h [20]. From the Leiderdorp subpopulation (n = 1671) we selected 599 Caucasian individuals with normal BMI distribution and sent their serum samples for metabolomics analysis using the Metabolon platform and for examination in this paper.

A1.1. Evaluation Measures

In addition to using bias and RMSE we also converted these measures to percentages. This was necessary because the estimates of the analysis model of the metabolites in complete NEO dataset varied in magnitude and scale. For example, in sample size n = 599, the estimate for urate was 1.39×10^{-8} and for succinylcarnitine was 2.84×10^{-6} (full details in Table 3 of the main manuscript). Percentage bias was calculated by dividing the bias in each sample size set by the estimate calculated for the respective sample size. RMSE percentage was calculated by subtracting then dividing all scenarios for each metabolite by the corresponding true coefficient in sample size n = 599 and multiplying by 100.

A1.2. Imputation Methods

The first step in our workflow was creating a correlation matrix for the metabolites in the dataset. For each metabolite with missing values (X), we selected the ten metabolites without missing values with the strongest absolute correlation $|r|$ to X from the correlation matrix. Our metabolomics dataset was generated on the latest measuring platform which greatly expanded the number of metabolites but reduced the overall intercorrelation of the data. This reduction of the intercorrelation is partly explained by the inclusion of remote metabolites in smaller pathways.

In standard kNN, distances are used to select closest neighbors to the observation with missing values. In kNN-obs-sel, for each metabolite we used up to 10 auxiliary metabolites as predictors and imputed the missing values by taking the average of the 10 nearest neighbors (K = 10) observations.Multiple imputation using chained equations (MICE) is used for incomplete data in multiple variables and may use discrete, categorical, and continuous variables of different units for the imputation [29]. When using the option predicted mean matching, it yields several different datasets with imputed values obtained from observed cases. The analysis of interest is then performed on each of the imputed datasets separately and the results are pooled afterwards as described by White et al., (2010) [23] and other articles [29–31]. Given that kNN-obs-sel calculates the mean from the auxiliary variables, it was only possible to use metabolites (with the same units and scale) for the imputation. In contrast, we used clinical variables sex and age in addition to the auxiliary metabolites as predictors. Furthermore, the outcome, BMI, was added as well. Adding the outcome and the covariates is essential in MICE imputations to avoid bias and underestimation in the imputation results as shown in simulation studies [24] and discussed in several sources [17,23]. Adding the clinical variables and the outcome in our study was an additional step that was not used in the simulation study by Do et al. [6].

8 APPENDIX B

Table B1. Distribution of the mean correlation for the incomplete endogenous and unannotated metabolites in NEO.

Mean Correlation	0.1–0.19, *n* (%)	0.2–0.29, *n* (%)	0.3–0.39, *n* (%)	0.4–0.49, *n* (%)	0.5–0.59, *n* (%)	0.6–0.69, *n* (%)	0.7–0.79, *n* (%)	0.8–0.89, *n* (%)	Total
Endogenous	32 (8.79)	86 (23.63)	98 (26.92)	58 (15.93)	53 (14.56)	29 (7.97)	7 (1.92)	1 (0.27)	364
Unannotated	53 (23.35)	83 (36.56)	50 (22.03)	22 (9.69)	15 (6.61)	1 (0.44)	3 (1.32)	0 (0)	227
Combined	85 (14.38)	169 (28.6)	148 (25.04)	80 (13.54)	68 (11.51)	30 (5.08)	10 (1.69)	1 (0.17)	591

Approximately 80% of the metabolites have a mean correlation below 0.5 with their respective top 10 correlated metabolites in the correlation matrix.

Table B2. Percentage bias of the imputation methods across different parameters on different metabolites including MICE-pmm with a single imputation.

Missing Mechanism	Sample Size	Missing Percentage	Metabolites/Imputation Method											
			PC(32:2)			Succinylcarnitine			Glutamate			Urate		
			MICE-pmm	kNN-obs-sel	MICE-pmm [m = 1]	MICE-pmm	kNN-obs-sel	MICE-pmm [m = 1]	MICE-pmm	kNN-obs-sel	MICE-pmm [m = 1]	MICE-pmm	kNN-obs-sel	MICE-pmm [m = 1]
MCAR	*n* = 50	15%	2.0	2.1	1.4	0.3	6.0	0.4	2.2	0.2	2.3	4.9	4.9	5.5
		30%	3.2	4.5	3.4	2.8	14.4	2.6	8.6	1.1	8.8	11.3	10.2	10.1
		60%	13.9	14.5	13.2	28.9	31.3	29.7	40.2	2.0	40.7	39.9	27.6	40.9
	n = 150	15%	0.7	0.5	0.9	1.4	13.8	1.7	0.2	1.9	0.3	0.7	0.5	0.5
		30%	1.9	1.8	2.2	0.4	31.9	0.9	1.1	5.1	1.1	0.7	1.3	0.1
		60%	3.9	3.2	3.4	5.2	72.7	0.6	6.2	15.0	6.0	2.09	12.5	2.1
	n = 599	15%	0.3	0.8	0.3	0.1	9.3	0.2	0.2	2.6	0.3	0.6	2.2	0.4
		30%	0.7	2.4	0.8	0.6	20.1	0.3	0.1	5.6	0.1	1.5	4.9	1.8
		60%	1.9	7.9	1.7	3.5	50.9	4.6	0.7	15.1	0.8	2.6	15.0	2.4
PLoD	*n* = 50	15%	1.3	0.8	1.0	2.8	1.8	3.4	0.7	1.8	0.8	12.7	6.2	12.5
		30%	3.5	2.2	2.8	7.7	4.3	7.5	4.2	3.6	4.1	26.3	16.5	25.5
		60%	19.5	3.0	18.6	34.8	8.3	34.8	19.3	15.0	19.6	87.4	66.1	89.4
	n = 150	15%	0.2	1.6	0.2	0.7	12.8	0.7	1.3	0.8	1.4	5.3	6.3	5.4
		30%	0.3	3.3	0.6	0.2	23.4	0.4	3.0	2.3	3.1	8.0	14.9	8.7
		60%	4.2	2.6	4.2	16.2	46.1	17.2	9.4	8.6	9.1	39.8	59.7	38.2
	n = 599	15%	0.5	0.1	0.5	0.9	7.3	1.0	0.9	1.4	0.9	0.5	1.9	0.5
		30%	1.1	1.6	1.2	2.5	16.7	2.9	2.1	2.9	2.2	0.6	3.3	0.6
		60%	4.9	11.5	4.9	9.7	54.3	9.2	4.1	13.1	4.4	0.1	4.3	0.6

Abbreviations: MCAR: missing completely at random; PLoD: probabilistic limit of detection; MICE-pmm [m = 1]: MICE-pmm with a single imputation.

Table B3. Mean and standard deviation of the estimates of the imputation methods using the MCAR missing mechanism, the three sample sizes, the three missing percentages, and four metabolites.

Sample Size	Missing Percentage	Metabolites/Imputation Method/Mean Estimate (SD)									
		PC(32:2)		Succinylcarnitine		Glutamate		Urate			
		MICE-pmm	kNN-obs-sel	MICE-pmm	kNN-obs-sel	MICE-pmm		kNN-obs-sel MICE-pmm		kNN-obs-sel	
n = 50	0%	-4.18 × 10^{-7}		2.84 × 10^{-6}		1.83 × 10^{-7}		1.39 × 10^{-8}			
	15%	-4.46 × 10^{-7} (6.04 × 10^{-8})	-4.47 × 10^{-7} (7.24 × 10^{-8})	4.52 × 10^{-6} (6.38 × 10^{-7})	4.8 × 10^{-6} (7.38 × 10^{-7})	2.82 × 10^{-8} (1.29 × 10^{-9})	2.88 × 10^{-8} (1.69 × 10^{-9})	9.11 × 10^{-9} (3.29 × 10^{-9})	1.01 × 10^{-8} (3.27 × 10^{-9})		
	30%	-4.52 × 10^{-7} (1.04 × 10^{-7})	-4.57 × 10^{-7} (1.21 × 10^{-7})	4.39 × 10^{-6} (1.02 × 10^{-6})	5.15 × 10^{-6} (1.24 × 10^{-6})	2.64 × 10^{-8} (2.61 × 10^{-9})	2.86 × 10^{-8} (2.79 × 10^{-9})	8.5 × 10^{-9} (5.66 × 10^{-9})	1.06 × 10^{-8} (5.72 × 10^{-9})		
	60%	-3.77 × 10^{-7} (2.28 × 10^{-7})	-5.01 × 10-7 (2.34 × 10-7)	3.22 × 10^{-6} (1.73 × 10^{-6})	5.96 × 10^{-6} (2.58 × 10^{-6})	1.73 × 10^{-8} (6.06 × 10^{-9})	2.83 × 10^{-8} (5.57 × 10^{-9})	5.75 × 10^{-9} (9.36 × 10^{-9})	1.22 × 10^{-8} (1.13 × 10^{-8})		
n = 150	0%	-3.64 × 10^{-7}		1.53 × 10^{-6}		2.89 × 10^{-8}		9.58 × 10^{-9}			
	15%	-3.66 × 10^{-7} (1.57 × 10^{-8})	-3.62 × 10^{-7} (1.53 × 10^{-8})	1.54 × 10^{-6} (4.09 × 10^{-7})	1.77 × 10^{-6} (4.17 × 10^{-7})	1.66 × 10^{-8} (8.24 × 10^{-10})	1.69 × 10^{-8} (8.15 × 10^{-10})	9.76 × 10^{-9} (1.62 × 10^{-9})	9.74 × 10^{-9} (1.46 × 10^{-9})		
	30%	-3.71 × 10^{-7} (2.57 × 10^{-8})	-3.58 × 10^{-7} (2.29 × 10^{-8})	1.52 × 10^{-6} (6.43 × 10^{-7})	1.99 × 10^{-6} (6.94 × 10^{-7})	1.64 × 10^{-8} (1.37 × 10^{-9})	1.74 × 10^{-8} (1.21 × 10^{-9})	9.75 × 10^{-9} (2.45 × 10^{-9})	9.82 × 10^{-9} (2.29 × 10^{-9})		
	60%	-3.78 × 10^{-7} (4.87 × 10^{-8})	-3.52 × 10^{-7} (4.3 × 10^{-8})	1.48 × 10^{-6} (1.2 × 10^{-6})	2.7 × 10^{-6} (1.48 × 10^{-6})	1.56 × 10^{-8} (2.79 × 10^{-9})	1.91 × 10^{-8} (2.47 × 10^{-9})	9.89 × 10^{-9} (4.61 × 10^{-9})	1.09 × 10^{-8} (4.81 × 10^{-9})		
n = 599	0%	-4.38 × 10^{-7}		4.53 × 10^{-6}		1.66 × 10^{-8}		9.69 × 10^{-9}			
	15%	-4.2 × 10^{-7} (8.44 × 10^{-9})	-4.22 × 10^{-7} (1.05 × 10^{-8})	2.84 × 10^{-6} (1.96 × 10^{-7})	3.1 × 10^{-6} (1.94 × 10^{-7})	1.84 × 10^{-8} (4.43 × 10^{-10})	1.88 × 10^{-8} (4.77 × 10^{-10})	1.4 × 10^{-8} (8.08 × 10^{-10})	1.42 × 10^{-8} (7.59 × 10^{-10})		
	30%	-4.21 × 10^{-7} (1.33 × 10^{-8})	-4.28 × 10^{-7} (1.76 × 10^{-8})	2.85 × 10^{-6} (3.08 × 10^{-7})	3.43 × 10^{-6} (3.35 × 10^{-7})	1.84 × 10^{-8} (7.05 × 10^{-10})	1.94 × 10^{-8} (7.16 × 10^{-10})	1.41 × 10^{-8} (1.27 × 10^{-9})	1.46 × 10^{-8} (1.19 × 10^{-9})		
	60%	-4.26 × 10^{-7} (2.52 × 10^{-8})	-4.51 × 10^{-7} (3.54 × 10^{-8})	2.79 × 10^{-6} (5.88 × 10^{-7})	4.34 × 10^{-6} (7.2 × 10^{-7})	1.82 × 10^{-8} (1.34 × 10^{-9})	2.11 × 10^{-8} (1.27 × 10^{-9})	1.43 × 10^{-8} (2.47 × 10^{-9})	1.6 × 10^{-8} (2.55 × 10^{-9})		

Estimates are the regression coefficient from the model BMI ~ age + sex + metabolite. Therefore, the estimates are the mean increase in BMI per 1 unit increase of the metabolite. The 0% rows are the estimates from the real data before amputing and imputing the missing values.

Table B4. Mean and standard deviation of the estimates of the imputation methods using the PLoD missing mechanism, the three sample sizes, the three missing percentages, and four metabolites.

Sample Size	Missing Percentage	Metabolites/Imputation Method/Mean Estimate (SD)							
		PC(32:2)	kNN-obs-sel	Succinylcarnitine	kNN-obs-sel	Glutamate	kNN-obs-sel	Urate	kNN-obs-sel
		MICE-pmm		MICE-pmm		MICE-pmm		MICE-pmm	
n = 50	0%	-4.18×10^{-7}		2.84×10^{-6}		1.83×10^{-7}		1.39×10^{-8}	
	15%	-4.32×10^{-7} (3.91×10^{-8})	-4.34×10^{-7} (3.35×10^{-8})	4.4×10^{-6} (2.56×10^{-7})	4.62×10^{-6} (1.79×10^{-7})	2.87×10^{-8} (1.09×10^{-9})	2.94×10^{-8} (1.29×10^{-9})	8.36×10^{-9} (2.53×10^{-9})	8.99×10^{-9} (2.19×10^{-9})
	30%	-4.22×10^{-7} (7.47×10^{-8})	-4.28×10^{-7} (5.19×10^{-8})	4.15×10^{-6} (4.4×10^{-7})	4.72×10^{-6} (2.83×10^{-7})	2.77×10^{-8} (2.07×10^{-9})	2.99×10^{-8} (1.99×10^{-9})	7.07×10^{-9} (3.7×10^{-9})	8×10^{-9} (3.14×10^{-9})
	60%	-3.52×10^{-7} (1.82×10^{-7})	-4.51×10^{-7} (1.32×10^{-7})	3.05×10^{-6} (1.18×10^{-6})	4.96×10^{-6} (1.19×10^{-6})	2.33×10^{-8} (4.89×10^{-9})	3.32×10^{-8} (3.59×10^{-9})	1.21×10^{-9} (7.1×10^{-9})	3.25×10^{-9} (7.47×10^{-9})
n = 150	0%	-3.64×10^{-7}		1.53×10^{-6}		2.89×10^{-8}		9.58×10^{-9}	
	15%	-3.65×10^{-7} (1.26×10^{-8})	-3.58×10^{-7} (1.22×10^{-8})	1.51×10^{-6} (3.88×10^{-7})	1.72×10^{-6} (3.61×10^{-7})	1.64×10^{-8} (6.63×10^{-10})	1.67×10^{-8} (5.85×10^{-10})	9.18×10^{-9} (1.62×10^{-9})	9.08×10^{-9} (1.31×10^{-9})
	30%	-3.65×10^{-7} (1.96×10^{-8})	-3.52×10^{-7} (1.86×10^{-8})	1.49×10^{-6} (6.11×10^{-7})	1.86×10^{-6} (5.51×10^{-7})	1.61×10^{-8} (1.07×10^{-9})	1.7×10^{-8} (8.47×10^{-10})	8.91×10^{-9} (2.14×10^{-9})	8.25×10^{-9} (1.87×10^{-9})
	60%	-3.79×10^{-7} (3.83×10^{-8})	-3.74×10^{-7} (3.83×10^{-8})	1.66×10^{-6} (1.05×10^{-6})	2.11×10^{-6} (9.8×10^{-7})	1.5×10^{-8} (1.8×10^{-9})	1.8×10^{-8} (1.51×10^{-9})	5.83×10^{-9}(3.36×10^{-9})	3.91×10^{-9} (3.08×10^{-9})
n = 599	0%	-4.38×10^{-7}		4.53×10^{-6}		1.66×10^{-8}		9.69×10^{-9}	
	15%	-4.2×10^{-7} (7.17×10^{-9})	-4.18×10^{-7} (8.61×10^{-9})	2.82×10^{-6} (1.59×10^{-7})	3.05×10^{-6} (1.5×10^{-7})	1.82×10^{-8} (3.37×10^{-10})	1.86×10^{-8} (3.1×10^{-10})	1.4×10^{-8} (6.57×10^{-10})	1.42×10^{-8} (5.63×10^{-10})
	30%	-4.23×10^{-7} (1.1×10^{-8})	-4.25×10^{-7} (1.52×10^{-8})	2.79×10^{-6} (2.52×10^{-7})	3.34×10^{-6} (2.34×10^{-7})	1.8×10^{-8} (5.37×10^{-10})	1.89×10^{-8} (4.81×10^{-10})	1.4×10^{-8} (9.58×10^{-10})	1.44×10^{-8} (8.97×10^{-10})
	60%	-4.39×10^{-7} (2.02×10^{-8})	-4.66×10^{-7} (3.56×10^{-8})	2.58×10^{-6} (5.62×10^{-7})	4.37×10^{-6} (6.13×10^{-7})	1.76×10^{-8} (9.99×10^{-10})	2.07×10^{-8} (8.04×10^{-10})	1.39×10^{-8} (1.64×10^{-9})	1.45×10^{-8} (1.57×10^{-9})

Estimates are the regression coefficient from the model BMI ~ age + sex + metabolite. Therefore, the estimates are the mean increase in BMI per 1 unit increase of the metabolite. The 0% rows are the estimates from the real data before amputing and imputing the missing values.

Table B5. RMSE of the imputation methods across different parameters on different metabolites.

Missing Mechanism	Sample Size	Missing Percentage	Metabolites/Imputation Method							
			PC(32:2)		Succinylcarnitine		Glutamate		Urate	
			MICE-pmm	kNN-obs-sel	MICE-pmm	kNN-obs-sel	MICE-pmm	kNN-obs-sel	MICE-pmm	kNN-obs-sel
MCAR	*n* = 50	15%	6.56×10^{-8}	7.48×10^{-8}	6.5×10^{-7}	7.73×10^{-7}	3.32×10^{-9}	3.3×10^{-9}	1.44×10^{-9}	1.69×10^{-9}
		30%	1.04×10^{-7}	1.22×10^{-7}	1.02×10^{-6}	1.44×10^{-6}	5.76×10^{-9}	5.8×10^{-9}	3.59×10^{-9}	2.8×10^{-9}
		60%	2.33×10^{-7}	2.49×10^{-7}	2.22×10^{-6}	2.95×10^{-6}	1.01×10^{-8}	1.16×10^{-8}	1.31×10^{-8}	5.59×10^{-9}
	n = 150	15%	1.71×10^{-8}	1.69×10^{-8}	3.97×10^{-7}	4.6×10^{-7}	1.62×10^{-9}	1.46×10^{-9}	8.24×10^{-10}	8.75×10^{-10}
		30%	2.63×10^{-8}	2.49×10^{-8}	6.48×10^{-7}	8.5×10^{-7}	2.45×10^{-9}	2.29×10^{-9}	1.38×10^{-9}	1.47×10^{-9}
		60%	5.23×10^{-8}	4.78×10^{-8}	1.27×10^{-6}	1.84×10^{-6}	4.62×10^{-9}	4.96×10^{-9}	2.97×10^{-9}	3.51×10^{-9}
	n = 599	15%	8.64×10^{-9}	1.08×10^{-8}	$2{,}00 \times 10^{-7}$	3.35×10^{-7}	8.12×10^{-10}	8.17×10^{-10}	4.45×10^{-10}	6.71×10^{-10}
		30%	1.28×10^{-8}	1.95×10^{-8}	3.08×10^{-7}	6.6×10^{-7}	1.28×10^{-9}	1.36×10^{-9}	7.05×10^{-10}	1.24×10^{-9}
		60%	2.63×10^{-8}	5.04×10^{-8}	5.91×10^{-7}	1.61×10^{-6}	2.5×10^{-9}	3.29×10^{-9}	1.35×10^{-9}	3.04×10^{-9}
PLoD	*n* = 50	15%	3.99×10^{-8}	3.23×10^{-8}	2.98×10^{-7}	$2{,}00 \times 10^{-7}$	2.81×10^{-9}	2.26×10^{-9}	1.11×10^{-9}	1.38×10^{-9}
		30%	7.36×10^{-8}	5.36×10^{-8}	5.66×10^{-7}	3.48×10^{-7}	4.47×10^{-9}	3.51×10^{-9}	2.4×10^{-9}	2.24×10^{-9}
		60%	1.87×10^{-7}	1.25×10^{-7}	$2{,}00 \times 10^{-6}$	1.27×10^{-6}	1.1×10^{-8}	9.79×10^{-9}	7.42×10^{-9}	5.62×10^{-9}
	n = 150	15%	1.33×10^{-8}	1.36×10^{-8}	3.62×10^{-7}	3.8×10^{-7}	1.7×10^{-9}	1.45×10^{-9}	6.97×10^{-10}	5.99×10^{-10}
		30%	1.96×10^{-8}	2.15×10^{-8}	5.97×10^{-7}	6.46×10^{-7}	2.27×10^{-9}	2.36×10^{-9}	1.18×10^{-9}	9.27×10^{-10}
		60%	4.42×10^{-8}	3.98×10^{-8}	1.07×10^{-6}	1.23×10^{-6}	5.11×10^{-9}	6.55×10^{-9}	2.39×10^{-9}	2.07×10^{-9}
	n = 599	15%	7.54×10^{-9}	8.52×10^{-9}	1.67×10^{-7}	2.56×10^{-7}	6.6×10^{-10}	6.22×10^{-10}	3.77×10^{-10}	4.03×10^{-10}
		30%	1.17×10^{-8}	1.64×10^{-8}	2.64×10^{-7}	5.32×10^{-7}	9.61×10^{-10}	1.01×10^{-9}	6.55×10^{-10}	7.2×10^{-10}
		60%	2.96×10^{-8}	5.82×10^{-8}	6.4×10^{-7}	1.65×10^{-6}	1.64×10^{-9}	1.68×10^{-9}	1.25×10^{-9}	2.53×10^{-9}

Abbreviations: MCAR: missing completely at random; PLoD: probabilistic limit of detection; MICE-pmm [m = 1]: MICE-pmm with a single imputation.

Table B6. Percentage RMSE of the imputation methods across different parameters on different metabolites including MICE-pmm with a single imputation.

Missing Mechanism	Sample Size	Missing Percentage	Metabolites/Imputation Method											
			PC(32:2)			Succinylcarnitine			Glutamate			Urate		
			MICE-pmm	kNN-obs-sel	MICE-pmm [m = 1]	MICE-pmm	kNN-obs-sel	MICE-pmm [m = 1]	MICE-pmm	kNN-obs-sel	MICE-pmm [m = 1]	MICE-pmm	kNN-obs-sel	MICE-pmm [m = 1]
MCAR	*n* = 50	15%	14.6	17.4	18.3	22.9	27.2	29.0	7.9	9.2	10.9	23.9	23.7	29.7
		30%	25.1	29.4	31.3	35.9	50.7	45.1	19.6	15.3	24.6	41.4	41.7	50.6
		60%	56.2	57.9	71.2	78.2	103.9	96.1	71.4	30.5	78.0	72.5	83.3	89.1
	n = 150	15%	3.8	3.7	5.0	14.0	16.2	17.1	4.5	4.8	5.9	11.6	10.5	14.1
		30%	6.4	5.7	8.2	22.8	29.9	27.7	7.5	8.0	9.8	17.6	16.5	21.1
		60%	12.1	10.7	15.6	44.7	64.8	51.8	16.2	19.1	20.0	33.2	35.6	39.3
	n = 599	15%	2.0	2.6	2.7	7.0	11.8	8.9	2.4	3.7	3.2	5.8	5.9	7.0
		30%	3.3	4.8	4.0	10.8	23.2	13.3	3.8	6.8	4.7	9.2	9.8	11.6
		60%	6.3	11.6	8.0	20.8	56.7	27.1	7.4	16.6	8.8	18.0	23.6	21.4
PLoD	*n* = 50	15%	9.4	8.1	12.5	10.5	7.0	15.5	6.1	7.5	8.3	20.2	16.2	23.9
		30%	18.2	12.6	23.3	19.9	12.3	25.5	13.1	12.2	17.0	32.1	25.2	37.2
		60%	48.1	31.8	61.0	70.4	44.7	81.0	40.4	30.6	48.9	79.0	70.3	90.5
	n = 150	15%	3.0	3.2	4.1	12.7	13.4	15.7	3.8	3.3	4.9	12.2	10.4	13.7
		30%	4.7	5.3	6.2	21.0	22.7	25.2	6.4	5.1	7.9	16.3	17.0	19.2
		60%	9.9	9.4	13.4	37.7	43.3	49.3	13.0	11.3	16.7	36.7	47.1	42.5
	n = 599	15%	1.8	2.1	2.4	5.9	9.0	7.7	2.1	2.2	2.6	4.7	4.5	6.0
		30%	2.8	3.9	3.7	9.3	18.7	12.6	3.6	3.9	4.4	6.9	7.3	8.6
		60%	6.9	14.3	8.3	22.5	58.1	27.4	6.8	13.8	8.1	11.8	12.1	15.2

We converted the RMSE values to a percentage by subtracting then dividing the RMSE values by the corresponding true estimates (in sample size *n* = 599). Abbreviations: MCAR: missing completely at random; PLoD: probabilistic limit of detection; MICE-pmm [m = 1]: MICE-pmm with a single imputation.

Table B7. Runtime for MICE-pmm and kNN-obs-sel imputations using different datasets.

Imputation Method	Dataset Sizes		
	***n* = 599**	***n* = 5400**	***n* = 20,000**
MICE-pmm (minutes)	1.9	13.4	138.9
kNN-obs-sel (minutes)	0.7	16.2	210.7

Imputation was applied to the actual NEO dataset (*n* = 599; 58% metabolites contain missing values) and two oversampled datasets generated from the NEO data. With *n* = 599, kNN-obs-sel was slightly faster. However, MICE-pmm imputation took a shorter time to complete the imputations in larger datasets.

Figure B1. Nested loop plot of the percentage bias of the four metabolites from the simulation including MICE-pmm with a single imputation. The X axis in each box represents the missing percentage and is split per sample size. Abbreviations: MCAR: missing completely at random; PLoD: probabilistic limit of detection; MICE-pmm [m = 1]: MICE-pmm with a single imputation.

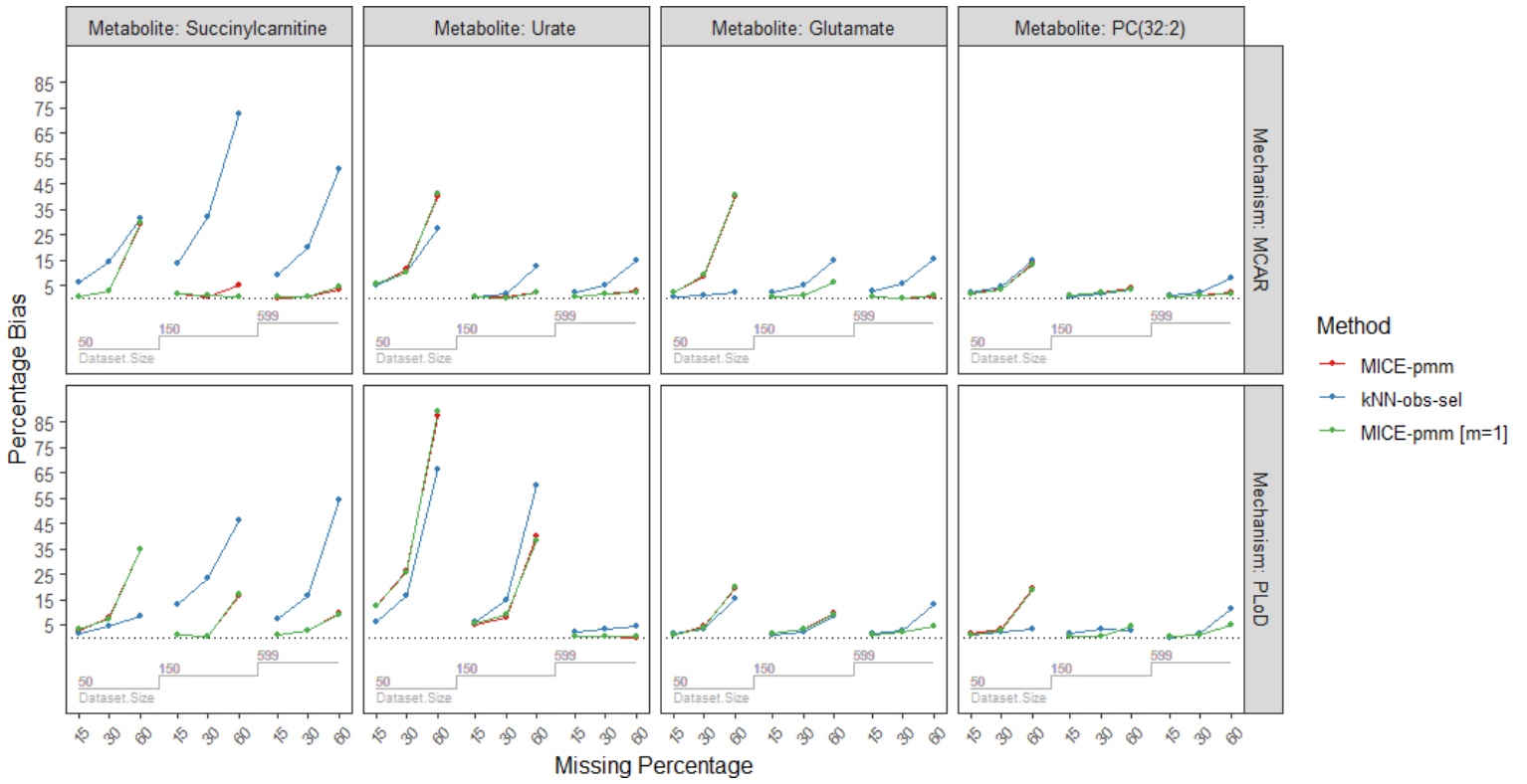
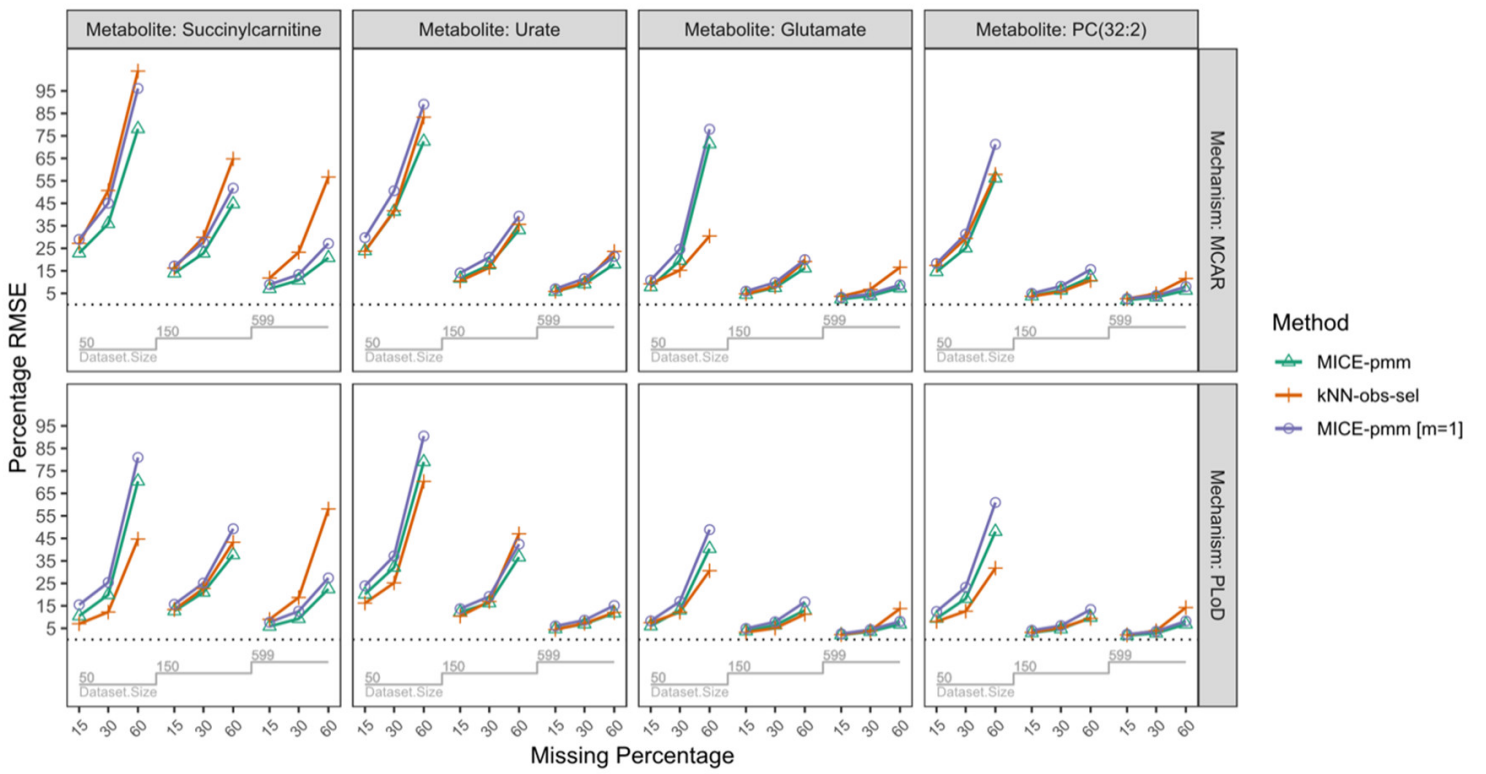


Figure B2. Nested loop plot of the percentage RMSE of the four metabolites from the simulation including MICE-pmm with a single imputation. To simplify comparability in the plot we converted the RMSE values to a percentage by subtracting then dividing the RMSE values by the corresponding true estimates (in sample size *n* = 599). The horizontal axis in each box represents the missing percentage and is split per sample size. Abbreviations: MCAR: missing completely at random; PLoD: probabilistic limit of detection; MICE-pmm [m = 1]: MICE-pmm with a single imputation.



9 REFERENCES

1. Suhre, K.; Meisinger, C.; Döring, A.; Altmaier, E.; Belcredi, P.; Gieger, C.; Chang, D.; Milburn, M.V.; Gall, W.E.; Weinberger, K.M.; et al. Metabolic Footprint of Diabetes: A Multiplatform Metabolomics Study in an Epidemiological Setting. *PLoS ONE* **2010**, *5*, e13953, doi:10.1371/journal.pone.0013953.
2. Schrimpe-Rutledge, A.C.; Codreanu, S.G.; Sherrod, S.D.; McLean, J.A. Untargeted Metabolomics Strategies—Challenges and Emerging Directions. *J. Am. Soc. Mass Spectrom.* **2016**, *27*, 1897–1905, doi:10.1007/s13361-016-1469-y.
3. Wei, R.; Wang, J.; Su, M.; Jia, E.; Chen, S.; Chen, T.; Ni, Y. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci. Rep.* **2018**, *8*, 1–10, doi:10.1038/s41598-017-19120-0.
4. Karpievitch, Y.V.; Dabney, A.R.; Smith, R.D. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinform.* **2012**, *13*, S5, doi:10.1186/1471-2105-13-s16-s5.
5. Hrydziuszko, O.; Viant, M.R. Missing values in mass spectrometry based metabolomics: An undervalued step in the data processing pipeline. *Metabolomics* **2012**, *8*, 161–174, doi:10.1007/s11306-011-0366-4.
6. Do, K.T.; Wahl, S.; Raffler, J.; Molnos, S.; Laimighofer, M.; Adamski, J.; Suhre, K.; Strauch, K.; Peters, A.; Gieger, C.; et al. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics* **2018**, *14*, 128.
7. Alonso, A.; Marsal, S.; Juliã, A. Analytical Methods in Untargeted Metabolomics: State of the Art in 2015. *Front. Bioeng. Biotechnol.* **2015**, *3*, 23, doi:10.3389/fbioe.2015.00023.
8. Deng, Y.; Chang, C.; Ido, M.S.; Long, Q. Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data. *Sci. Rep.* **2016**, *6*, 21689, doi:10.1038/srep21689.
9. Gromski, P.S.; Xu, Y.; Kotze, H.L.; Correa, E.; Ellis, D.I.; Armitage, E.G.; Turner, M.L.; Goodacre, R. Influence of Missing Values Substitutes on Multivariate Analysis of Metabolomics Data. *Metabolites* **2014**, *4*, 433–452, doi:10.3390/metabo4020433.
10. Van Buuren, S. *Flexible Imputation of Missing Data*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018.
11. Little, R.J.A. Missing-Data Adjustments in Large Surveys. *J. Bus. Econ. Stat.* **1988,** *6*, 287–296.
12. Rubin, D.B. Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. *J. Bus. Econ. Stat.* **1986**, *4*, 87–94.
13. Faquih, T. *Imputation of Untargeted Metabolites Official Release*, Version v1.3; Zenodo: Meyrin, Switzerland, 2020. Available online: https://zenodo.org/record/4167193 (accessed on 31 October 2020).
14. Rücker, G.; Schwarzer, G. Presenting simulation results in a nested loop plot. *BMC Med. Res. Methodol.* **2014**, *14*, 129, doi:10.1186/1471-2288-14-129.
15. Shah, J.; Rai, S.N.; DeFilippis, A.P.; Hill, B.G.; Bhatnagar, A.; Brock, G. Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies. *BMC Bioinform.* **2017**, *18*, 114, doi:10.1186/s12859-017-1547-6.
16. Di Guida, R.; Engel, J.; Allwood, J.W.; Weber, R.J.M.; Jones, M.R.; Sommer, U.; Viant, M.R.; Dunn, W.B. Non-targeted UHPLC-MS metabolomic data processing methods: A comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics* **2016**, *12*, 1–14, doi:10.1007/s11306-016-1030-9.
17. Molenberghs, G.; Kenward, M. *Missing Data in Clinical Studies*; Wiley: Hoboken, NJ, USA, 2007.
18. Wang, J.; Li, Z.F.; Chen, J.; Zhao, H.; Luo, L.; Chen, C.; Xu, X.; Zhang, W.; Gao, K.; Li, B.; et al. Metabolomic identification of diagnostic plasma biomarkers in humans with chronic heart failure. *Mol. BioSyst.* **2013**, *9*, 2618, doi:10.1039/c3mb70227h.
19. Yousri, N.A.; Bayoumy, K.; Elhaq, W.G.; Mohney, R.P.; Al Emadi, S.; Hammoudeh, M.; Halabi, H.; Masri, B.; Badsha, H.; Uthman, I.; et al. Large Scale Metabolic Profiling identifies Novel Steroids linked to Rheumatoid Arthritis. *Sci. Rep.* **2017**, *7*, 1–9, doi:10.1038/s41598-017-05439-1.

20. De Mutsert, R.; Heijer, M.D.; Rabelink, T.J.; Smit, J.W.A.; Romijn, J.A.; Jukema, J.W.; De Roos, A.; Cobbaert, C.M.; Kloppenburg, M.; Le Cessie, S.; et al. The Netherlands Epidemiology of Obesity (NEO) study: Study design and data collection. *Eur. J. Epidemiol.* **2013**, *28*, 513–523, doi:10.1007/s10654-013-9801-3.
21. Evans, A.; Bridgewater, B.; Liu, Q.; Mitchell, M.; Robinson, R.; Dai, H.; Stewart, S.; DeHaven, C.; Miller, L.J.M. High Resolution Mass Spectrometry Improves Data Quantity and Quality as Compared to Unit Mass Resolution Mass Spectrometry in High-Throughput Profiling Metabolomics. *J. Postgenomics Drug Biomark. Dev.* **2014**, *4*, 1–7, doi:10.4172/2153-0769.1000132.
22. Rhee, E.P.; Waikar, S.S.; Rebholz, C.M.; Zheng, Z.; Perichon, R.; Clish, C.B.; Evans, A.M.; Avila, J.; Denburg, M.R.; Anderson, A.H.; et al. Variability of Two Metabolomic Platforms in CKD. *Clin. J. Am. Soc. Nephrol.* **2019**, *14*, 40.
23. White, I.R.; Royston, P.; Wood, A.M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* **2011**, *30*, 377–399, doi:10.1002/sim.4067.
24. Moons, K.G.; Donders, R.A.; Stijnen, T.; Harrell, F.E. Using the outcome for imputation of missing predictor values was preferred. *J. Clin. Epidemiol.* **2006**, *59*, 1092–1101, doi:10.1016/j.jclinepi.2006.01.009.
25. Kowarik, A.; Templ, M. Imputation with the R Package VIM. *J. Stat. Softw.* **2016,** *74*, 16.
26. Cirulli, E.T.; Guo, L.; Swisher, C.L.; Shah, N.; Huang, L.; Napier, L.A.; Kirkness, E.F.; Spector, T.D.; Caskey, C.T.; Thorens, B.; et al. Profound Perturbation of the Metabolome in Obesity Is Associated with Health Risk. *Cell Metab.* **2019**, *29*, 488–500.e2, doi:10.1016/j.cmet.2018.09.022.
27. Demirtas, H.; Freels, S.A.; Yucel, R.M. Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. *J. Stat. Comput. Simul.* **2008**, *78*, 69–84, doi:10.1080/10629360600903866.
28. Morris, T.P.; White, I.R.; Crowther, M.J. Using simulation studies to evaluate statistical methods. *Stat. Med.* **2019**, *38*, 2074–2102, doi:10.1002/sim.8086.
29. Van Buuren, S.; Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **2011**, *45*, 1–67, doi:10.18637/jss.v045.i03.
30. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; John Wiley & Sons, Inc.: New York, NY, USA, 1987.
31. Rubin, D.B. Multiple Imputation After 18+ Years. *J. Am. Stat. Assoc.* **1996**, *91*, 473–489.

Chapter 4

Robust metabolomic age prediction based on a wide selection of metabolites

Key words: metabolomics, ridge regression, prediction, aging, sleep, cardiometabolic disease, depression

Tariq Faquih[1], Astrid van Hylckama Vlieg[1], Praveen Surendran[2,3,4,5], Ruifang Li-Gao[1,6], Renée de Mutsert[1], Frits R. Rosendaal[1], Raymond Noordam[7], Diana van Heemst[7], Ko Willems van Dijk[8,9,10], Dennis Mook-Kanamori[1,11]

[1] Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands; T.O.Faquih@lumc.nl (T.O.F.); R.Li@lumc.nl (R.L.-G.); R.de_Mutsert@lumc.nl (R.d.M); F.R.Rosendaal@lumc.nl (F.R.R.); A.van_Hylckama_Vlieg@lumc.nl (A.v.H.V.); D.O.Mook@lumc.nl (D.O.M.-K.)

[2] British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; ps629@medschl.cam.ac.uk (P.S.)

[3] British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK; ps629@medschl.cam.ac.uk (P.S.)

[4] Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK; ps629@medschl.cam.ac.uk (P.S.)

[5] Rutherford Fund Fellow, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; ps629@medschl.cam.ac.uk (P.S.)

[6] Metabolon, Inc. Morrisville, North Carolina, United State of America; R.Li@lumc.nl (R.L.-G.)

[7] Department of Internal Medicine, Section of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands; R.Noordam@lumc.nl (R.N.); D.van_Heemst@lumc.nl (D.v.H.)

[8] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; K.Willems_van_Dijk@lumc.nl (K.W.v.D)

[9] Department of Internal Medicine, Division of Endocrinology, Leiden University Medical Center, Leiden, The Netherlands; K.Willems_van_Dijk@lumc.nl (K.W.v.D)

[10] Einthoven Laboratory for Experimental Vascular Medicine, Leiden University Medical Center, Leiden, The Netherlands; K.Willems_van_Dijk@lumc.nl (K.W.v.D)

[11] Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands; D.O.Mook@lumc.nl (D.O.M.-K.)

Manuscript in preparation

1 ABSTRACT

Chronological age is a major risk factor for numerous diseases. However, the biological aging process is complex and often does not align with chronological age. Metabolite levels are thought to reflect the integrated effects of both genetic and environmental factors, including age, and thus may provide a signature for biological age. Here, we set out to develop a rigorous metabolomic age prediction model by applying ridge regression and bootstrapping with 826 metabolites (of which 678 endogenous and 148 xenobiotic) measured in 11977 individuals (age range: 18-75 years old) from the INTERVAL study (Cambridge, UK). Subsequently, the metabolomic age prediction model was applied in the Netherlands Epidemiology of Obesity Study (NEO) (n=599) to quantify the difference between metabolomic and chronological age (Δ age). We assessed the influence of cardiovascular disease (CVD), hypertension, type 2 diabetes, obesity/body mass index (BMI), depression, and sleep duration on the Δ age. The metabolomic age models using both endogenous and xenobiotic metabolites demonstrated high correlation with chronological age (R^2=0.82). In NEO, CVD associated with increased Δ age by approximately 12 years and obese BMI (45kg/m^2) associated with increased Δ age by approximately 8.5 years, respectively. In summary, we developed robust models for predicting metabolomic age in a large relatively healthy population with a wide age range. We further demonstrated that Δ age can potentially reflect the effects of CVD and obesity/BMI in the NEO study.

2 INTRODUCTION

Chronological age is a major risk factor for a multitude of diseases (1, 2). The biology of aging is a complex and multifactorial process that is influenced by lifestyle and environmental factors (3-5). However, it is evident that the rate of aging varies between individuals, wherein some individuals are able to live to an older chronological age without age-related diseases and disability compared with individuals in the same age group (3).This suggests that chronological age does not fully align with the biological aging process. Thus, several studies aimed to capture the signature of biological changes due to aging by predicting age using biological factors. Such studies used DNA methylation (6) and proteins (7).

Metabolomic profiling aims to identify small molecules that are mostly substrates and products of cell metabolism (metabolites). The number of metabolomics studies has increased in recent years due to major technological advances and availability of commercial and noncommercial analyses platforms. In addition, the current platforms have improved capability to detect and quantify large numbers of endogenous and xenobiotic metabolites. Since individual metabolomic profiles reflect the influences of both genetic and acquired factors, they are thought to provide a holistic representation of biological processes, such as aging (8, 9). Furthermore, metabolomic profiles are strongly affected by chronological age (4, 10) and sex (11, 12), they have been used to develop prediction models of chronological age (i.e., the metabolomic age) (13-16). However, predicting metabolomic age using metabolomics had limited success or faced methodological limitations due to several reasons. First, some studies used an insufficiently small sample size to predict age using hundreds or even thousands of metabolite predictors (10, 17). The inclusion of larger number of predictors than the sample often causes large overfitting and bias in such models. Second, studies can be limited by the age distribution of the cohort study. This restricts the model to a specific age range which affects the generalizability of the model in other studies and other age groups (10). Third, lack of generalizability can also be caused if the model was developed in cohorts with an oversampling of individuals with specific disease outcomes or a specific population (7, 13). Fourth, even when the previous considerations are addressed, oversimplified statistical methodology and application can lead to a flawed, biased, and an overfitted model—such as the case with stepwise selection models(18). Finally, the model's validity and generalizability are seldom examined in external studies or different populations via external validation (19-21).

In this study, we aimed to develop a model to predict metabolomic age in a large healthy population with a widespread age range, using a single untargeted metabolomics platform. To attain this goal, we developed a prediction model using data from the INTERVAL study (22) (University of Cambridge, Cambridge, UK). Metabolomic profiles were available in 11977 participants as measured using Metabolon's (Durham, North Carolina, USA) untargeted metabolomics platform. These measurements included a broad range of endogenous and xenobiotic metabolites (n=1,363) from various biochemical pathways, thereby enabling the capture of metabolites related to a vast range of ageing effects. The model was subsequently applied to the Netherlands Epidemiology of Obesity Study (NEO) (n=599) to determine the effects of six health-related phenotypes associated with aging (23, 24): cardiovascular disease, hypertension, type 2 diabetes, obesity and body weight, depression, and sleep duration on the difference between metabolomic age and chronological age (14, 15).

3 METHODS

3.1 INTERVAL Study

The INTERVAL study is a prospective cohort study of approximately 50,000 participants nested within a pragmatic randomized sample of blood donors (22). Between 2012 and 2014, blood donors, aged 18 years and older, were consented and recruited from 25 National Health Service Blood and Transplant (NHSBT) static donor centers across the UK. Individuals with major disease (myocardial infarction, stroke, cancer etc.) as well as those who reported being unwell or having had recent illness or infection or did not fulfill the other criteria required for blood donation (22, 25) were ineligible for the study. Therefore, participants included in the study were predominantly healthy. Participants completed online questionnaires addressing basic lifestyle and health-related information, including self-reported height and weight, ethnicity, current smoking status, alcohol consumption, doctor-diagnosed anemia, use of medications (hormone replacement therapy, iron supplements) and menopausal status (22). Untargeted metabolomic data were available in 11979 individuals (age range 18 - 75). Two individuals had incorrect or missing height/weight values and were therefore excluded from the study. Thus, the final sample size for the current study was 11977.

3.2 Untargeted metabolomic measurements

Untargeted metabolomic measurements were quantified at Metabolon Inc. (Durham, North Carolina, USA) using Metabolon™ Discovery HD4 platform. In brief, this process involves four independent ultra-high-performance liquid chromatography mass spectrometry (UHPLC-MS/MS) platforms (26, 27). Two using positive ionization reverse phase chromatography, one using negative ionization reverse phase chromatography, and one using hydrophilic interaction liquid chromatography negative ionization (27). Known metabolites were annotated at Metabolon Inc. with chemical names, super pathways, sub pathways, biochemical properties, and compound identifiers from various metabolite databases. Metabolomic measurements in the INTERVAL study were conducted in three batches (n=4087, 4566, and 3326). Subsequent harmonization and quality checks were performed between the batches by Metabolon. All metabolite measurements were scaled to a median of 1.

3.3 Selection of Predictors

We aimed to select metabolites consistently and reliably measured by the Metabolon platform. In total 1411 metabolites were measured in the INTERVAL study. First, we removed metabolites completely missing in at least one of the batches (n=175). Second, we excluded metabolites without annotation/unnamed (n=258), keeping only endogenous and xenobiotic metabolites. These metabolites were excluded as they are inconsistently measured by the platform and are highly variable between batches and studies. Moreover, the lack of full annotation increases the uncertainty that we are using the same metabolite across batches and studies and leaves no secondary information for further verification. Third, we excluded metabolites measured in less than 100 individuals (n=30). Fourth, we excluded metabolites that were not measured in the NEO study (n=122). The final set of metabolites included 826 metabolites, with 678 endogenous and 148 xenobiotic metabolites (Figure 1).

3.4 Missing value imputation

Missing values were imputed using the pipeline as described in our previous work (28). In brief, endogenous metabolites were imputed by multiple imputation using chained equations method to generate five imputed datasets (m=5). For each metabolite with missing values, we used the

outcome variable (i.e., age), 5-10 highly correlated related metabolites, body mass index (BMI), center number where the blood samples were collected, and the batch number to impute the missing values. Xenobiotic metabolites were imputed to zero to account for true missingness.

3.5 Model Development

For the development of the model, we used ridge regression (29) to reduce potential overfitting. Cross validation was performed (n=10) in each imputed dataset to calculate the optimal shrinkage term (lambda). Subsequently the mean of the lambda values was used to develop the model on the stacked imputed datasets (i.e., all 5 datasets combined as one). Accordingly, the weight of the observations in the stacked dataset was set to 1/m = 1/5 = 0.2 (20). As the outcome (age) is a continuous variable, we assessed the fit of the model by deriving the R^2. As an additional sensitivity test, we used generalized additive model (GAM) to examine and calculate the R^2 for the nonlinear correlation. Internal validation was performed using bootstrapping (b=100). Bootstrapping results were used for the optimization of R^2 and the calculation of the mean squared error (MSE) and the mean absolute error (MAE) of model. Two models were developed using two sets of the selected metabolite predictors. First, we used the full set of endogenous and xenobiotic metabolites (n=826) to develop "model A". Second, we used only the endogenous metabolites (n=678) to develop "model B". As previous studies found that metabolomic profiles (12, 30) and aging (31-33) are influenced by the sex of individuals, we included sex as an additional predictor in both models.

3.6 Sample size considerations

The primary database used to create the prediction model was the INTERVAL study (n= 11977) with 826 predictors. we used the formulas described by Riley et al. (34) to confirm that our sample size (n) and number of predictor (p) are sufficient to minimize overfitting and provide high precision. First, we used n and p to check that the calculated Copas global shrinkage factor(35, 36) is above the recommended 0.9 threshold (34). Based on this calculation the estimated shrinkage factor was 0.95 if the adjusted R^2 of the model was assumed to be 0.7. Second, we calculated the sample size required to ensure a small difference between the R^2 and the adjusted R^2 for the development model. Assuming the adjusted R^2 was 0.7 again and a small desired R^2 difference (R^2_{diff}=0.025), then the sample size required to achieve this should be at least n=9913. Third, we checked the sample size required for precise residual standard deviation of the model. Accordingly, we found the multiplicative margin of error (MMOE) to be less than 10% (MMOE =1.3%) using our n and p in the INTERVAL study. Finally, we checked the precision of the mean predicted outcome value (predicted age) of the model. We used n and p for the INTERVAL study and assumed that predicted age would have a mean of 45 and a variance of 35. Accordingly, the upper and lower bounds were approximately 45.34 and 44.65 respectively. Thus, the MMOE for the mean predicted outcome was less 1% (MMOE= 45.34 /45 = 1.007 = 0.7%). Therefore, the sample size of the INTERVAL study was optimal to minimize overfitting, optimism, and provide a precise estimation of the residual standard deviation and mean predicted values.

3.7 Metabolomic Age and Health-Related Phenotypes

3.7.1 NEO Study

The Netherlands Epidemiology of Obesity (NEO) study is a population-based, prospective cohort study of individuals aged 45–65 years, with an oversampling of individuals with overweight or obesity. Men and women aged between 45 and 65 years with a self-reported BMI of 27 kg/m^2

or higher, living in the greater area of Leiden (in the West of the Netherlands) were eligible to participate in the NEO study. In addition, all inhabitants aged between 45 and 65 years from one municipality (Leiderdorp) were invited, irrespective of their BMI. Recruitment of participants started in September 2008 and completed at the end of September 2012. In total, 6,671 participants have been included. Participants were invited to come to the NEO study center of the LUMC for one baseline study visit after an overnight fast. A blood sample of 108 mL was taken from the participants after an overnight fast of at least 10 hours. Untargeted metabolomics measurements are available in the 599 individuals from Leiderdorp sub-population (single batch). Therefore, only 599 individuals were included in this study. All metabolite measurements were scaled to a median of 1. Missing values in the metabolites measurements was conducted using the same method in the INTERVAL study using a single imputation.

3.7.2 Health-Related phenotypes in the NEO study

Several health-related phenotypes, including cardiometabolic risk factors, are associated with both metabolomic profiles (12, 37) and aging (23, 24). In the NEO study, individual patient data was available for the following health-related phenotypes: cardiovascular disease (CVD), type 2 diabetes (T2D), hypertension, BMI, depressive symptoms, and total sleep duration. At baseline of the NEO study, 21 individuals had self-reported cardiovascular disease, which was defined as a composite trait composed of myocardial infarction, angina, congestive heart failure, stroke, and peripheral vascular disease. In addition, 89 individuals had type 2 diabetes, based on impaired fasting glycaemia (6.1-7.0 mmol/L), fasting plasma glucose higher or equal to 7.0 mmol/L, or self-reported diabetes mellitus 1 or 2 medication. Hypertension was characterized in 213 individuals, who had systolic blood pressure >=140mmHg or diastolic blood pressure >=90mmHg. BMI was calculated based on physical examination measurements and 79 participants had a BMI in the obese range (>35 kg/m^2). Depression score (1-84) was derived from the Inventory of Depressive Symptomatology (IDS) questionnaire. We defined current depression status as a score of 14 and above (23). Using the IDS questionnaire, 111 participants were characterized as depressed at baseline. Total Sleep duration was derived from self-reported questionnaire; specifically, from the question: "on an average day, how much sleep do you get?". We further categorized sleep duration using the 5th percentile to define the shortest sleep duration, 5th to 20th as short, 20th to 80th as medium, 80th to 95th as long, and 95th and above as the longest sleep duration. Mean sleep duration was 7 hours of which the majority (71%) reported a total sleep duration between 6 and 8 while short sleep duration (5.5-6 hours) accounted for 15.9%.

3.7.3 Estimating the Effects of Health-Related Phenotypes on Metabolomic Age

The developed prediction model was applied to derive metabolomic age for 599 individuals in the NEO study with metabolomics data. To assess the effects of health-related phenotypes on the difference between metabolomic age and chronological age—henceforth referred to as Δ age—with the available 6 health-related phenotypes. For a better interpretation of the BMI effect on the Δ age, we centered it on the median healthy BMI range (BMI = 22 kg/m^2) as defined by the World Health Organization(38). We used a simple linear regression analysis for each of the six clinical phenotypes as exposures separately. The Δ age was the outcome variable in each analysis.

4 RESULTS

4.1 Population characteristics

Characteristics of the total INTERVAL study population and age subgroups are summarized in Table 1. In total 11977 individuals were included and had a normal age distribution with a mean age of 45 years and a range of 18 – 75 years. The number of men and women was approximately equal (50.2% were men). BMI was largely in the recommended range of 18.5-24.9 kg/m^2 (38) with a mean BMI of 22.8 kg/m^2 and was similar in all age groups.

Characteristics of the NEO study population are summarized in Table 2. The mean age was 56 years and ranged from 45 to 66. Mean BMI was 25.9, slightly over the recommended BMI (38) and was similar in men and women.

Table 1. Characteristics of the INTERVAL study population.

	Total	18 to 25	25 to 35	35 to 45	45 to 55	55 to 65	65 to 75
N	11977	1178	2245	2152	2867	2602	811
Age (years), mean (range)	45 (18-75)						
Men, n (%)	6019 (50.2%)	485 (41.2%)	965 (42.3%)	1042 (48.4%)	1560 (54.4%)	1480 (56.9%)	546 (67.3%)
BMI (kg/m^2), mean (SD)	22.8 (4.2)	21.3 (4.0)	22.0 (4.2)	23.3 (4.4)	23.5 (4.3)	23.0 (3.9)	22.6 (3.5)

Table 2. Characteristics of the NEO study population.

	Total
n	599
Age (years), mean (range)	56 (45-65)
Men, n (%)	284 (47.7)
BMI (kg/m^2), mean (SD)	25.9 (4.0)
Obesity, n (%)	79 (13.2)
CVD, n (%)	21 (3.5)
T2D, n (%)	89 (14.9)
Hypertension, n (%)	213 (35.6)
Depression, n (%)	111 (19)
Total Sleep Duration(hours), mean (SD)	7 (0.98)
Shortest (<5.5), n (%)	40 (6.7)
Short (5.5 - 6), n (%)	95 (15.9)
Medium (6 - 8), n (%)	425 (71.0)
Long (8 - 8.5), n (%)	16 (27)
Longest (>8.5), n (%)	21 (3.5)

4.2 Metabolomic Age Prediction

Prediction models A (endogenous plus xenobiotic metabolites) and B (endogenous metabolites only) were developed in the INTERVAL study using ridge regression. The workflow including metabolite selection, missing value imputation, and analyses are summarized in Figure 1. Internal validation using bootstrapping (b=100) and optimization provided an R^2 of 0.83 (MSE=31, MAE=4.4) for model A, and 0.82 (MSE=33.7, MAE=4.6) for model B. GAM R^2 was slightly higher for both models, 0.85 for model A and 0.84 for model B (Figure 2, Supplementary Table 1). Full tables with the intercept, sex, and metabolite coefficients for model A and B are provided in supplementary Table 2. This table also contains the mean values for the metabolites from the INTERVAL study.

Figure 1: Flowchart of the selection of predictor metabolites and the development steps for the metabolomic age prediction model in the INTERVAL study

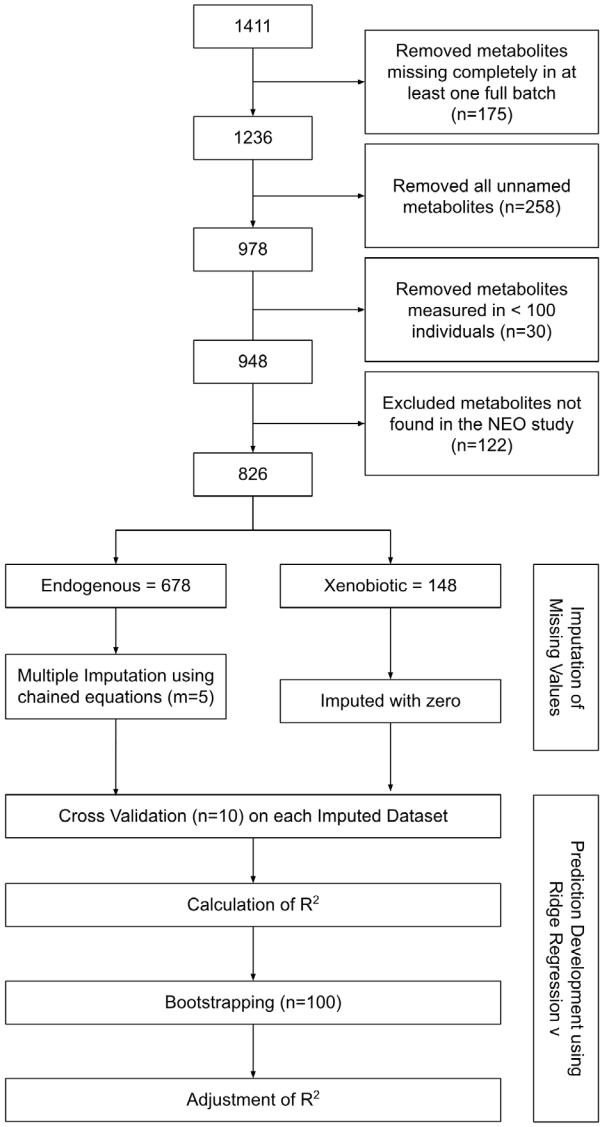
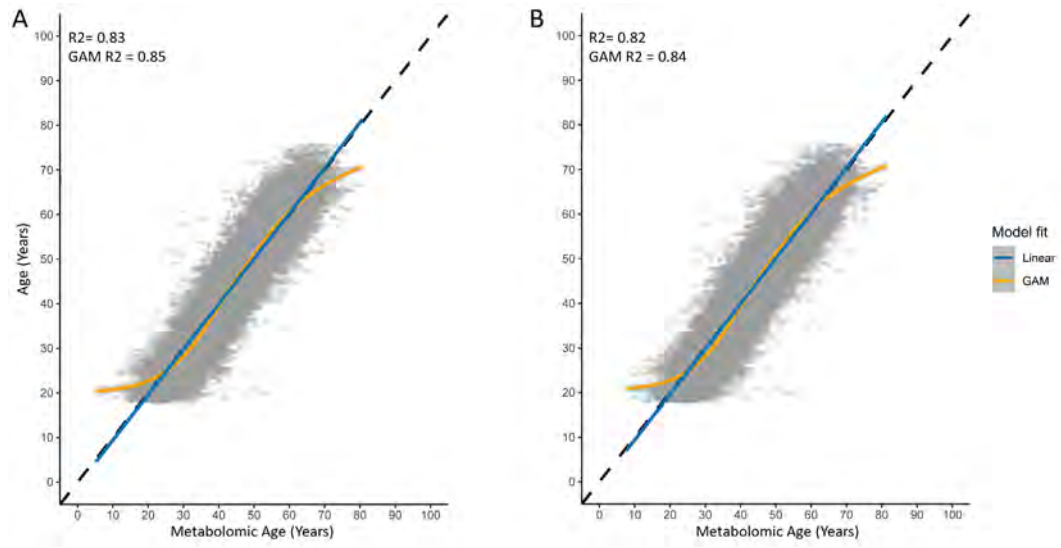


Figure 2: Correlation plots of the Metabolomic age (predicted age) on the horizontal axis and the chronological age on the vertical axis for model A (A) and model B (B). The data used is the stacked imputed datasets in the INTERVAL study. Abbreviations: GAM, generalized additive model.



4.3 Chronological age versus Metabolomic Age

The coefficient estimates from model A and B from the INTERVAL study were subsequently applied to the NEO data to predict metabolomic age. We first performed calibration in the large in the NEO data to readjust the intercept (model A from 51.97 to 50.83; model B from 51.85 to 49.59). Subsequently, the effect of the phenotypes CVD, BMI/obesity, T2D, hypertension, depression, and total sleep duration on Δ age were assessed, for both model A and model B. First, we plotted Δ age against chronological age and highlighted individuals with each respective phenotype for model A and model B (Figure 3 A-F; and Figure 4 A-F). Overall, model A showed higher variance in Δ age (σ^2=170) compared with model B (σ^2=84.1), notably for individuals with a negative health trait (i.e., with CVD or depression etc.).

Figure 3: Scatter plot representing the age difference (Δ age)—as predicted using model A—on the y-axis and chronological age on the x-axis in the NEO study. Each subplot highlights individuals with specific phenotypes.

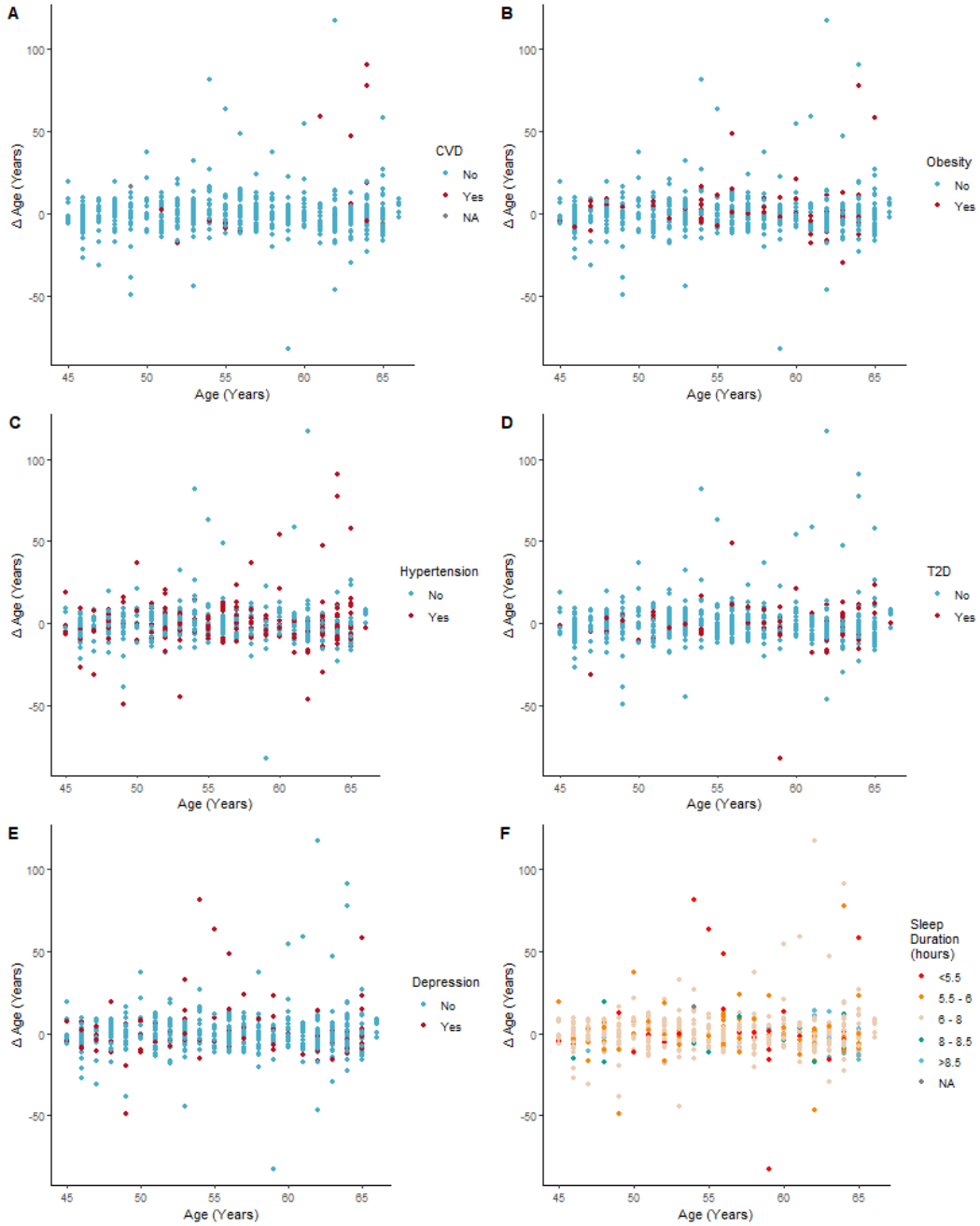
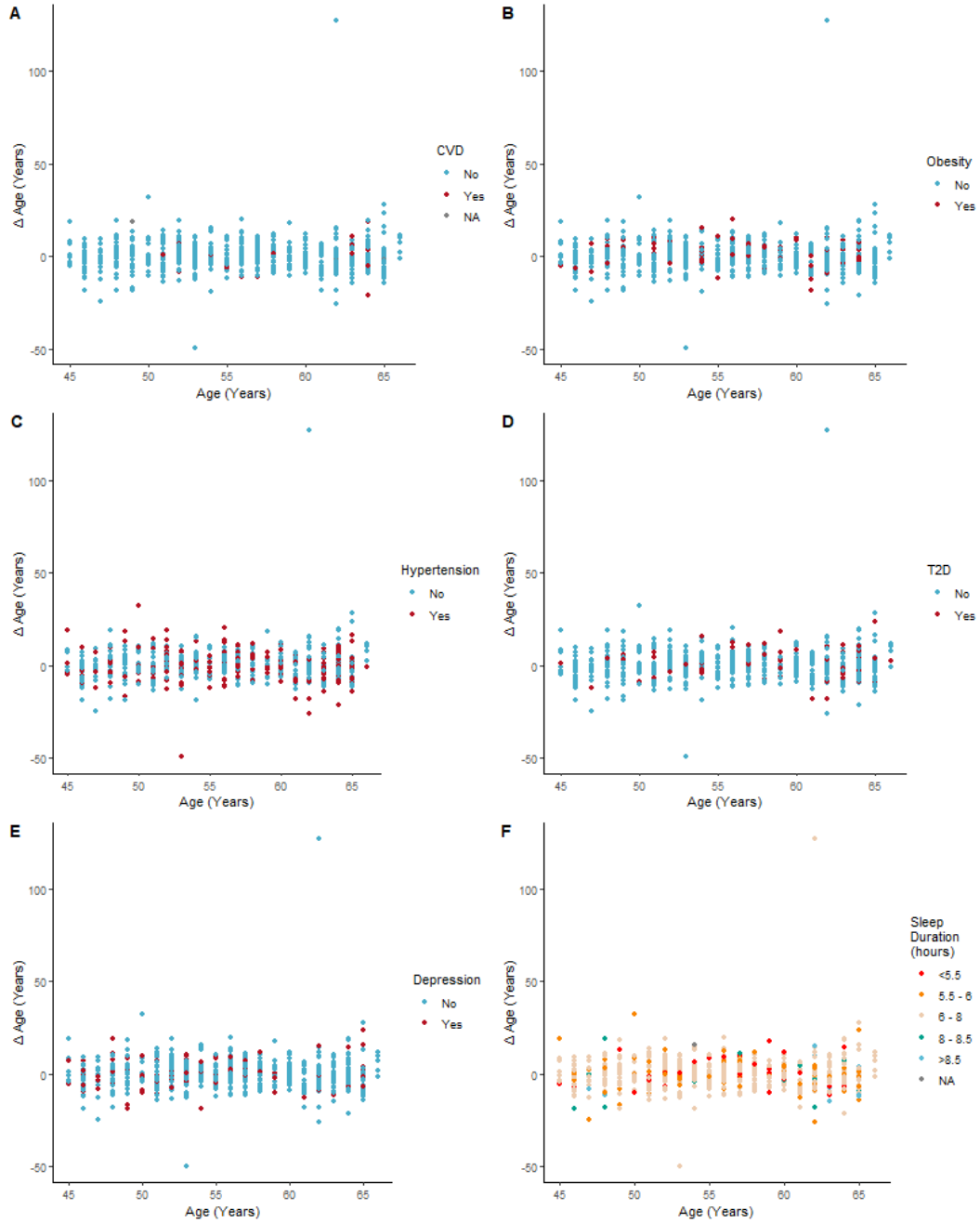


Figure 4: Scatter plot representing the age difference (Δ age)—as predicted using model B—on the y-axis and chronological age on the x-axis in the NEO study. Each subplot highlights individuals with specific phenotypes.



Second, we assessed the effects of the health-related phenotypes on the Δ age using linear regression. These results are shown in Figure 5 and Supplementary Table 3. In both models, BMI was associated with an increase of Δ age (beta [95% confidence intervals (CI)]: model A = 0.37 [0.1 – 0.64]; model B = 0.25 [0.06 – 0.43] years/ kg/m2). Obesity was also associated with an increase of Δ age but with wide CI in both models (model A = 3.31 [0.14 – 6.49]; model B = 1.89 [-0.27 – 4.06]). Thereafter, using the estimates from models A and B, we calculated the age at BMI's 22, 25, 35, and 45 kg/m2 (Figure 6). Overall, model A showed higher Δ age estimations at all BMI levels. At 35 kg/m2, the projected Δ age increased by 3.7 and 2.4 years in models A and B, respectively, as compared with 25 kg/m2. Whereas at BMI of 45 kg/m2, the projected Δ age rose by 8.5 and 5.6 years in models A and B, respectively, compared with BMI of 25 kg/m2.

CVD was strongly associated with an increased Δ age in model A (12.13 years [6.34 – 17.92]) but not model B (-0.89 years [-4.89 – 3.09]). No associations were found in either models for T2D, hypertension, depression, or total sleep duration.

Figure 5: A forest plot of estimates of phenotypes with the Δ age using model A (A) and model B (B).

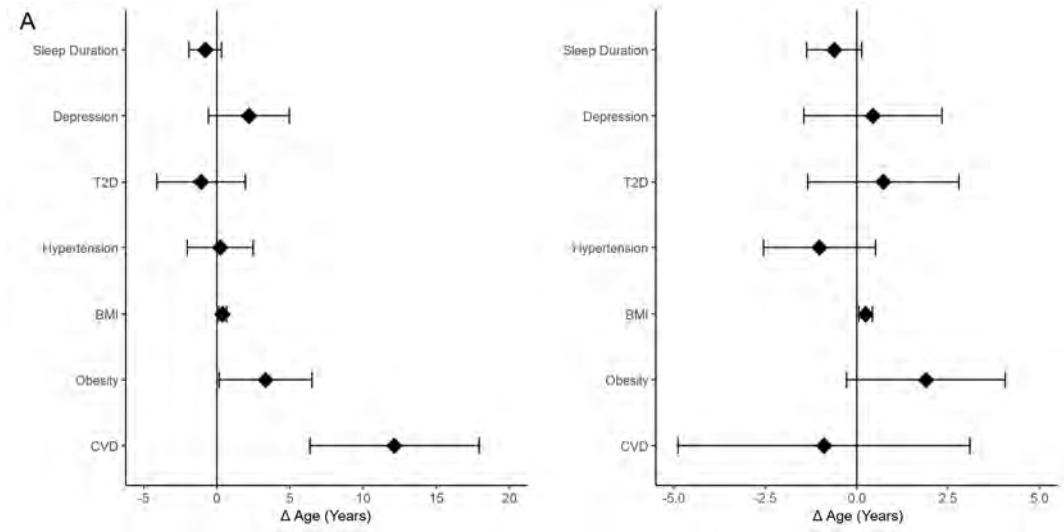
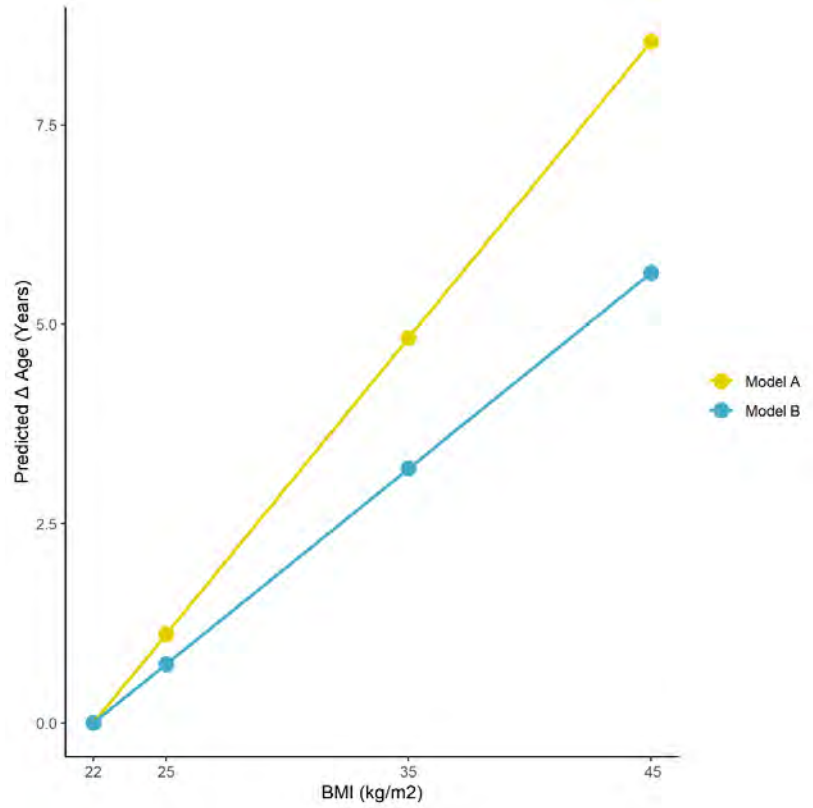


Figure 6: Predicted Δ age at 22,25,35, and 45 kg/m2 based on the estimates from model A and model B of BMI in the NEO study



5 DISCUSSION

In this study, we developed a prediction model for metabolomic age based on metabolite measurements, including a wide range of endogenous and xenobiotic metabolites belonging to a variety of biochemical pathways. The metabolomic measurements were performed in a single study using the same metabolomic platform and harmonized for within study and between batch variation. Importantly, we used multiple imputation, ridge regression, and bootstrapping(20) to develop and internally validate the metabolomic age prediction models. We developed two models, with the first (model A) using both endogenous and xenobiotic metabolites (n=826) and the second (model B) using the endogenous metabolites only (n=678). Both models had high adjusted R^2 (model A = 0.83; Model B = 0.82), however, model B had slightly higher MSE and MAE, indicating higher error for the predicted values.

5.1 Ridge Regression for Metabolomic Age

Our metabolomic age model is based on ridge regression which generates shrinkage factors that shrink metabolite coefficients with weak influence on the model to small values. Unlike methods such as LASSO and elastic net regression, this shrinkage never causes the coefficients to reach zero. Therefore, unlike those methods, ridge regression does not perform selection of predictors/metabolites due to the shrinkage term. Thus, the model consistently includes the

same metabolites during the cross validation and internal validation by bootstrapping as well as during redevelopment and external validation in studies such as NEO (20). Ridge regression thus ensures rigorousness and consideration of all the selected metabolite predictors. In our model we have found the nonlinear GAM R^2 to be slightly higher than linear R^2 in both models. Therefore, predicting metabolomic age from this model in a nonlinear form, by adding a quadratic term or using a spline function, can further expand the model flexibility.

5.2 The effect of health traits on metabolomic age

After developing the metabolomic age prediction models, we used the NEO study for two purposes. First, we used the NEO study to assess the viability of the prediction models. This was done by estimating the metabolomic age in the NEO study and readjusted the intercepts of the models accordingly (calibration in the large). The readjustment of the intercepts was minimal in the NEO metabolomic age models. Second, we used the models in NEO to examine the influence of six health phenotypes on the Δ age. We expected that the cases with negative health phenotypes had a higher Δ age (BMI/obesity, T2D, hypertension, CVD, depression, and short sleep). On the other hand, we expected no effect or a lower Δ age for individuals who did not suffer from these phenotypes or reported longer sleep durations.

Overall, Model A showed wider Δ age distributions compared with model B. As expected, our results showed that BMI, obesity, and particularly CVD had a strong effect on the Δ age using model A. Despite the equivalent R^2 for model A and B in the INTERVAL study, only BMI was associated with the Δ age estimated from model B. Since model A includes xenobiotic metabolites, medication related metabolites are likely important for capturing the effects of CVD and obesity on metabolomic age. The effect of BMI on age is pronounced in our estimation of Δ age using BMI's of 25, 35 and 45 kg/m^2. Indeed, at an obese BMI of 45 kg/m^2 the predicted increase in Δ age was 8.5 and 5.6 years for models A and B, respectively.

None of the other health traits; sleep duration, depression, T2D, or hypertension affected Δ age in either model. This was unexpected as for example T2D and hypertension are associated with CVD risk and have increased incidence rates in older individuals (24, 39). In addition, a previous study on metabolic age has shown that metabolites were a strong predictor of T2D and CVD (14). One possible reason for this is the exclusion of insulin dependent T2D individuals and those with a history of major disease outcomes such as myocardial infarction and stroke in the INTERVAL study due to the blood donor eligibility criteria in UK (40). This exclusion criteria may have affected the ability of metabolomic age, particularly in model B, to reflect these phenotypes in the Δ age in the NEO study, which did not have such exclusions. However, Δ age from model A showed a clear association with CVD. Therefore, other factors, such as the small sample size or unknown factors, could be involved possibly in addition to the exclusion criteria in the INTERVAL study. Further investigation is required in a larger sample size to verify the metabolomic age and the factors affecting the models. Depression was similarly not associated with the metabolomic age developed in a previous study. In contrast, they found an association between depression and their epigenetic, proteomic, and transcriptomic age predictors in the same population (7). Regarding the sleep duration results, some studies reported a U-shaped association between sleep and health outcomes (41). In this study we only examined the linear association in the NEO study. Thus, further examination of different association patterns for sleep are needed in future studies. Finally, we avoided examining etiological associations between the metabolites with Δ age and the six phenotypes as this was beyond the scope of paper. However, further exploration of etiological associations for specific metabolites with metabolomic age and health related phenotypes could be of interest for future studies.

5.3 Strengths and Limitations

A major strength of our current study is the development of robust metabolomic age prediction models based on a large number of metabolites measured in a large cohort with a wide age distribution. This sample size was confirmed to fit the required criteria for developing a model with low overfitting (34). Furthermore, the INTERVAL study included relatively healthy blood donors as participants, making it a suitable study to develop a metabolomic age without being affected with selection bias based on specific disease-related inclusion criteria. Unlike previous metabolomic age predictors (alternatively referred to as metabolomic "clocks"), we developed our model using the latest Metabolon metabolomics platform that measures a large selection of metabolites. A benefit of this platform is the inclusion of xenobiotics, such as those derived from medication or pollution as well as endogenous metabolites. Thus, we were able to include metabolites originating from internal and external sources. The xenobiotic metabolites represent some of the acquired environmental and lifestyle exposures of individuals that could play a role in biological aging.

Regarding the limitations, because the INTERVAL study does not have patient data for health-related outcomes, we could not assess the effects of different disease outcomes and health phenotypes on the metabolomic age. A second limitation was the small sample size included from of the NEO study relative to the number of predictors (n<p) and the narrower age range. Therefore, applying the models in the NEO study lead to loss in power and affected the reliability and performance of the models (20). Another limitation was the low number of events in some of the selected phenotypes in the NEO study. This was the case for CVD (n=21), T2D (n=89) and short/long sleep duration. It is possible that the larger number of metabolites from endogenous and xenobiotics may capture the metabolomic phenotype of some of the health traits more accurately in larger sample size. In addition, the results for sleep duration could have been affected by the sleep variable used. In NEO sleep duration was derived from a self-reported questionnaire, which is prone to recall bias. For future studies, sleep duration and quality quantification can be improved by using objectively measured actigraphy data.

5.4 External Validation for Metabolomic Age and Future Applications

Prediction models may suffer from overfitting due to selection bias, small sample size, methodology limitations, and lack of internal validation and calibration during development. However, another common issue regarding prediction modelling, such as the case with metabolomic age, is the challenge to externally validate them. This a common issue with prediction models in general. Indeed, few studies perform external validation of prediction models (19, 42). The reasons for the lack of external validation include the of difficulty applying and reproducing the prediction model method, lack of the full prediction variables to develop the model in the new dataset, or a lack of an appropriate effective sample size for external validation. Without external validation the quality of the models cannot be properly assessed, and a model could still be overfitted despite presenting good results during internal validation (20, 21, 43). We took advantage of the sample size and wide age range to use a stringent ridge regression and internal validation method to develop the metabolomic age model. The resulting model demonstrated a high R^2 for the metabolomic age and, as shown in NEO, was influenced by CVD and BMI. Accordingly, future robust external validation, using weak and moderate calibrations (43) would be valuable for the metabolomic age models presented in this paper. Furthermore, examination of the association between different phenotypes, such as those used in the NEO study, in the external validation would be valuable to provide more statistical power to assess their influence on metabolomic age and Δ age. Indeed, this could improve and verify the assessment of the phenotypes known

to influence metabolomic profiles (14, 24, 39, 44) and predicted metabolomic age as reported in our study and previous studies.

Several age prediction models have been developed that utilize different biological measurements such as targeted metabolomics (7, 14), proteomics (7), and DNA methylation (6). Here, we used the Metabolon untargeted metabolomics platform for the prediction of metabolomic age. In addition to our aim of addressing the primary issues with the development of metabolomic age models, the Metabolon platform expanded on the range of metabolites that can potentially be a better predictor of age. For example, previous metabolomic age studies that could not measure or include xenobiotic metabolites. In our study, we were able to include this additional group of metabolites in model A. We found that model A's Δ age, but not model B's Δ age, was greatly increased by CVD and obesity. This apparent additional predictive value of xenobiotics can be further investigated in future studies. Furthermore, Model A may also be used in tandem with metabolomic age from targeted platforms, and biological clocks of different biological molecules measurements similar to the work by Jansen et al. (7), to possibly improve or compare their predictive performance and their ability of capturing the effects of health-related phenotypes.

6 CONCLUSIONS

We developed metabolomic age prediction models in a large relatively healthy population using a wide array of endogenous and xenobiotic metabolites. In model A with the endogenous and xenobiotic metabolites and in model B with endogenous metabolites only, the R^2 of the linear fit was 0.82 and 0.83, respectively. The hypotheses that the predicted metabolomic age reflects metabolomic age and that health-related phenotypes increase metabolomic age was subsequently tested in the NEO study. These analyses revealed that obesity and CVD increased metabolomic age only in the model A, indicating possibly higher predictive value from external influences as reflected by the xenobiotic metabolites. We provided the full list of metabolites and their coefficients for both models. This data can enable other researchers to replicate our metabolomic age prediction model, externally validate it in their own studies with different disease outcomes and combine them with other age prediction models.

7 REFERENCES

1. North BJ, Sinclair DA. The intersection between aging and cardiovascular disease. Circ Res. 2012;110(8):1097-108.
2. Broglio SP, Eckner JT, Paulson HL, Kutcher JS. Cognitive decline and aging: the role of concussive and subconcussive impacts. 2012;40(3):138.
3. Lopez-Otin C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Cell. 2013;153(6):1194-217.
4. Hoffman JM, Lyu Y, Pletcher SD, Promislow DEL. Proteomics and metabolomics in ageing research: from biomarkers to systems biology. Essays Biochem. 2017;61(3):379-88.
5. Brooks-Wilson AR. Genetics of healthy aging and longevity. 2013;132(12):1323-38.
6. Horvath S. DNA methylation age of human tissues and cell types. Genome Biology. 2013;14(10):3156.
7. Jansen R, Han LK, Verhoeven JE, Aberg KA, van den Oord EC, Milaneschi Y, et al. An integrative study of five biological clocks in somatic and mental health. eLife. 2021;10.
8. Rattray NJW, Deziel NC, Wallach JD, Khan SA, Vasiliou V, Ioannidis JPA, et al. Beyond genomics: understanding exposotypes through metabolomics. Hum Genomics. 2018;12(1):4.
9. Alonso A, Marsal S, Julia A. Analytical methods in untargeted metabolomics: state of the art in 2015. Front Bioeng Biotechnol. 2015;3:23.
10. Rutledge J, Oh H, Wyss-Coray T. Measuring biological age using omics data. Nature Reviews Genetics. 2022.
11. Martin FJ, Montoliu I, Kussmann M. Metabonomics of ageing - Towards understanding metabolism of a long and healthy life. Mech Ageing Dev. 2017;165(Pt B):171-9.
12. Yu Z, Zhai G, Singmann P, He Y, Xu T, Prehn C, et al. Human serum metabolic profiles are age dependent. Aging Cell. 2012;11(6):960-7.
13. Macdonald-Dunlop E, Taba N, Klarić L, Frkatović A, Walker R, Hayward C, et al. A catalogue of omics biological ageing clocks reveals substantial commonality and associations with disease risk. Aging. 2022;14(2):623-59.
14. van den Akker EB, Trompet S, Barkey Wolf JJH, Beekman M, Suchiman HED, Deelen J, et al. Metabolic Age Based on the BBMRI-NL (1)H-NMR Metabolomics Repository as Biomarker of Age-related Disease. Circulation Genomic and precision medicine. 2020;13(5):541-7.
15. Hertel J, Friedrich N, Wittfeld K, Pietzner M, Budde K, Van der Auwera S, et al. Measuring Biological Age via Metabonomics: The Metabolic Age Score. J Proteome Res. 2016;15(2):400-10.
16. Rist MJ, Roth A, Frommherz L, Weinert CH, Kruger R, Merz B, et al. Metabolite patterns predicting sex and age in participants of the Karlsruhe Metabolomics and Nutrition (KarMeN) study. PLoS One. 2017;12(8):e0183228.
17. Hwangbo N, Zhang X, Raftery D, Gu H, Hu S-C, Montine TJ, et al. A Metabolomic Aging Clock Using Human Cerebrospinal Fluid. The Journals of Gerontology: Series A. 2021;77(4):744-54.
18. Smith G. Step away from stepwise. Journal of Big Data. 2018;5(1):32.
19. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clinical Kidney Journal. 2020;14(1):49-58.
20. Steyerberg EW. Clinical Prediction Models. 2nd ed. Cham, Switzerland: Springer International Publishing; 2019 2019.
21. Bleeker SE, Moll HA, Steyerberg EW, Donders ART, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research:: A clinical example. Journal of Clinical Epidemiology. 2003;56(9):826-32.
22. Moore C, Sambrook J, Walker M, Tolkien Z, Kaptoge S, Allen D, et al. The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. Trials. 2014;15:363-.
23. Emami M, Agbaedeng TA, Thomas G, Middeldorp ME, Thiyagarajah A, Wong CX, et al. Accelerated

Biological Aging Secondary to Cardiometabolic Risk Factors Is a Predictor of Cardiovascular Mortality: A Systematic Review and Meta-analysis. Canadian Journal of Cardiology. 2022;38(3):365-75.
24. Buford TW. Hypertension and aging. Ageing research reviews. 2016;26:96-111.
25. NHS Blood Donation Who can give blood 2022 [updated 2022/10/07/. Available from: https://www.blood.co.uk/who-can-give-blood.
26. Evans A, Bridgewater B, Liu Q, Mitchell M, Robinson R, Dai H, et al. High resolution mass spectrometry improves data quantity and quality as compared to unit mass resolution mass spectrometry in high-throughput profiling metabolomics. 2014;4(2):1.
27. Rhee EP, Waikar SS, Rebholz CM, Zheng Z, Perichon R, Clish CB, et al. Variability of Two Metabolomic Platforms in CKD. Clinical Journal of the American Society of Nephrology. 2019;14(1):40.
28. Faquih T, van Smeden M, Luo J, le Cessie S, Kastenmüller G, Krumsiek J, et al. A Workflow for Missing Values Imputation of Untargeted Metabolomics Data. Metabolites. 2020;10(12).
29. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. 1970;12(1):55-67.
30. Saner C, Harcourt BE, Pandey A, Ellul S, McCallum Z, Kao K-T, et al. Sex and puberty-related differences in metabolomic profiles associated with adiposity measures in youth with obesity. Metabolomics. 2019;15(5):75.
31. Hägg S, Jylhävä J. Sex differences in biological aging with a focus on human studies. eLife. 2021;10:e63425.
32. Nakamura E, Miyao KJTJoGSABS, Sciences M. Sex differences in human biological aging. 2008;63(9):936-44.
33. McCrory C, Fiorito G, McLoughlin S, Polidoro S, Cheallaigh CN, Bourke N, et al. Epigenetic clocks and allostatic load reveal potential sex-specific drivers of biological aging. 2020;75(3):495-503.
34. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE, Jr., Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. Stat Med. 2019;38(7):1262-75.
35. Copas JB. Regression, prediction and shrinkage. Journal of the Royal Statistical Society: Series B. 1983;45(3):311-35.
36. Copas JB. Using regression models for prediction: shrinkage and regression to the mean. Statistical methods in medical research. 1997;6(2):167-83.
37. Regan JA, Shah SH. Obesity Genomics and Metabolomics: a Nexus of Cardiometabolic Risk. Current cardiology reports. 2020;22(12):174.
38. WHO WHO. A healthy lifestyle - WHO recommendations. World Health Organization: WHO. 2010.
39. Einarson TR, Acs A, Ludwig C, Panton UH. Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of scientific evidence from across the world in 2007–2017. Cardiovascular Diabetology. 2018;17(1):83.
40. Watts M. Can People with Diabetes Give Blood? 2022 [updated 2022/09/08/. Available from: https://www.diabetes.co.uk/can-people-with-diabetes-give-blood.html.
41. Knutson KL, Turek FW. The U-shaped association between sleep and health: the 2 peaks do not mean the same thing. Sleep. 2006;29(7):878-9.
42. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. Journal of Clinical Epidemiology. 2015;68(1):25-34.
43. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: the Achilles heel of predictive analytics. BMC Medicine. 2019;17(1):230.
44. Ahola-Olli AV, Mustelin L, Kalimeri M, Kettunen J, Jokelainen J, Auvinen J, et al. Circulating metabolites and the risk of type 2 diabetes: a prospective study of 11,896 young adults from four Finnish cohorts. Diabetologia. 2019;62(12):2298-309.

8 FUNDING

The NEO study is supported by the participating Departments, the Division, and the Board of Directors of the Leiden University Medical Centre, and by the Leiden University, Research Profile Area 'Vascular and Regenerative Medicine'. The analyses of metabolites are funded by the VENI grant (ZonMW-VENI Grant 916.14.023) of D.O.M.-K., D.v.H. and R.N. were supported by a grant of the VELUX Stiftung [grant number 1156]. T.O.F. was supported by the King Abdullah Scholarship Program and King Faisal Specialist Hospital & Research Center [No. 1012879283].

9 CONFLICTS OF INTEREST

R.L.-G. is a part-time clinical research consultant for Metabolon, Inc. All other co- authors have no conflicts of interest to declare.

10 ACKNOWLEDGMENTS AND DISCLOSURES

The authors of the NEO study thank all participants, all participating general practitioners for inviting eligible participants, all research nurses for data collection, and the NEO study group: Pat van Beelen, Petra Noordijk, and Ingeborg de Jonge for coordination, laboratory, and data management.

11 AUTHOR CONTRIBUTIONS

T.O.F.- conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing-original draft. R.L.-G.- validation, writing – review & editing. P.S. – supervision, conceptualization, project administration, resources, funding acquisition, writing – review & editing - R.d.M.- project administration, resources, funding acquisition, writing – review & editing. R.N. and D.V.H.- funding acquisition, writing – review & editing. F.R.R.- funding acquisition. A.v.H.V. and K.W.v.D- conceptualization, supervision, writing – review & editing. D.O.M.-K.- conceptualization, supervision, funding acquisition, writing – review & editing.

Part II

Epidemiological Research and Advanced Data Analysis

Chapter 5

Analyses of metabolites and biochemical pathways associated with hepatic triglyceride content indicate extensive metabolite dysregulation

Tariq O. Faquih [1], Jan Bert van Klinken [2,3], Ruifang Li-Gao [1,4], Raymond Noordam [5], Diana van Heemst [5], Sebastiaan Boone [1], Patricia A Sheridan [4], Gregory Michelotti [4], Hildo Lamb [6], Renée de Mutsert [1], Frits R. Rosendaal [1], Astrid van Hylckama Vlieg [1], Ko Willems van Dijk [2, 7, 8], Dennis O. Mook-Kanamori [1, 9 *]

[1] Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands; T.O.Faquih@lumc.nl (T.O.F.); R.Li@lumc.nl (R.L.-G.); s.c.boone@lumc.nl (S.C.B.); R.de_Mutsert@lumc.nl (R.d.M); F.R.Rosendaal@lumc.nl (F.R.R.); A.van_Hylckama_Vlieg@lumc.nl (A.v.H.V.); D.O.Mook@lumc.nl (D.O.M.-K.)
[2] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; J.B.van_Klinken@lumc.nl (J.B.v.K); K.Willems_van_Dijk@lumc.nl (K.W.v.D)
[3] Laboratory Genetic Metabolic Diseases, Amsterdam UMC, University of Amsterdam, Departments of Clinical Chemistry and Pediatrics, Amsterdam Gastroenterology Endocrinology Metabolism, Amsterdam, The Netherlands; J.B.van_Klinken@lumc.nl (J.B.v.K).
[4] Metabolon, Inc. Morrisville, North Carolina, United State of America; R.Li@lumc.nl (R.L.-G.); psheridan@metabolon.com (P.A.S.); GMichelotti@metabolon.com (G.M.)
[5] Department of Internal Medicine, Section of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands; R.Noordam@lumc.nl (R.N.); D.van_Heemst@lumc.nl (D.v.H.)
[6] Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands; h.j.lamb@lumc.nl (H.L.)
[7] Department of Internal Medicine, Division of Endocrinology, Leiden University Medical Center, Leiden, The Netherlands; K.Willems_van_Dijk@lumc.nl (K.W.v.D)
[8] Einthoven Laboratory for Experimental Vascular Medicine, Leiden University Medical Center, Leiden, The Netherlands; K.Willems_van_Dijk@lumc.nl (K.W.v.D)
[9] Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands; D.O.Mook@lumc.nl (D.O.M.-K.)

Correspondence
Dennis O. Mook-Kanamori, Department of Clinical Epidemiology, Leiden University Medical Center & Department of Public Health and Primary Care, Leiden University and Medical Center, PO box 9600, 2300 RC Leiden, The Netherlands. Email: D.O.Mook@lumc.nl

Main body word count: 4321; Tables: 1; Figures: 4

Abbreviations
NAFLD, Non-alcoholic fatty liver disease; HTGC, hepatic triglyceride content; GMM, Gaussian graphical model; GSMM, genome scale metabolic model; BCAA, branched chain amino acids; IR, insulin resistance; T2D, type 2 diabetes; NEO, The Netherlands Epidemiology of Obesity Study; ^{1}H-MRS, Proton magnetic resonance spectroscopy, BMI, body mass index; MRI, magnetic resonance imaging; HOMA1-IR, homeostatic model assessment index for insulin resistance; CVD, cardiovascular disease; UHPLC-MS/MS, ultra-high-performance liquid chromatography mass spectrometry; HILIC, hydrophilic interaction liquid chromatography; HMR2, Human metabolic reactions database; EBIC, Extended Bayesian information criterium; PC, phosphatidylcholines; PE, phosphatidylethanolamines; PI, phosphatidylinositol; GPC, Glycerophosphorylcholine; HCER, hexosyl/glucosylceramide ; LCER, lactosylceramide ; AKG, alpha ketoglutarate; BCKA, branched chain keto acids.

Funding information
The NEO study is supported by the participating Departments, the Division, and the Board of Directors of the Leiden University Medical Centre, and by the Leiden University, Research Profile Area 'Vascular and Regenerative Medicine'. The analyses of metabolites are funded by the VENI grant (ZonMW-VENI Grant 916.14.023) of D.O.M.-K., D.v.H. and R.N. were supported by a grant of the VELUX Stiftung [grant number 1156]. T.O.F. was supported by the King Abdullah Scholarship Program and King Faisal Specialist Hospital & Research Center [No. 1012879283].

Conflicts of Interest
R.L.-G. is a part-time clinical research consultant for Metabolon, Inc. P.A.S. is an associate director and G.M. is the science director at Metabolon, Inc. All other co- authors have no conflicts of interest to declare.

Acknowledgments and Disclosures
The authors of the NEO study thank all participants, all participating general practitioners for inviting eligible participants, all research nurses for data collection, and the NEO study group: Pat van Beelen, Petra Noordijk, and Ingeborg de Jonge for coordination, laboratory, and data management.

Author Contributions
T.O.F.- conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing-original draft. J.B.v.K- formal analysis, software, visualization, writing – review & editing. R.L.-G.- validation, writing – review & editing. R.d.M.- project administration, resources, funding acquisition, writing – review & editing. S.C.B. - writing – review & editing. R.N. and D.V.H.- funding acquisition, writing – review & editing. P.A.S. - resources, writing – review & editing. G.M.- resources. H.L.- resources. F.R.R.- funding acquisition. A.v.H.V. and K.W.v.D- conceptualization, supervision, writing – review & editing. D.O.M.-K.- conceptualization, supervision, funding acquisition, writing – review & editing.

Manuscript submitted, under revision

1 ABSTRACT

Background & Aims

Non-alcoholic fatty liver disease (NAFLD) is characterized by the pathological accumulation of triglycerides in hepatocytes and is associated with insulin resistance, atherogenic dyslipidemia, and cardiometabolic diseases. Thus far, the extent of metabolic dysregulation associated with hepatic triglyceride accumulation has not been fully addressed. In this study we aimed to identify metabolites associated hepatic triglyceride content (HTGC) and map these associations using network analysis.

Methods

To gain insight in the spectrum of metabolites associated with hepatic triglyceride accumulation, we performed a comprehensive plasma metabolomics screening of 1,363 metabolites in apparently healthy middle aged (age 45-65) individuals (N=496) in whom HTGC was measured by proton magnetic resonance spectroscopy. An atlas of metabolite-HTGC associations, based on univariate results, was created using correlation-based Gaussian graphical model (GGM) and genome scale metabolic model (GSMM) network analyses.

Results

Our analyses revealed that 118 metabolites were univariately associated with HTGC (P-value $< 6.59\times10^{-5}$), including 106 endogenous, 1 xenobiotic, and 11 partially characterized/uncharacterized metabolites. These associations were mapped to several biological pathways including branched amino acids (BCAA), diglycerols, sphingomyelin, glucosyl- and lactosyl- ceramides. We also identified novel a possible HTGC related pathway connecting glutamate, metabolonic lactone sulfate, and X-15245 using the GGM network. The full interactive metabolite-HTGC atlas is provided online: https://tofaquih.github.io/AtlasLiver/.

Conclusions

The combined network and pathway analyses indicated extensive associations between BCAA and the lipids pathways with HTGC. Moreover, we report a novel pathway glutamate-metabolonic lactone sulfate-X-15245 with a potential strong association with HTGC. These findings can aid elucidating HTGC metabolomic profiles and provide insight in novel drug targets for treatment or monitoring of NAFLD.

2 INTRODUCTION

Nonalcoholic fatty liver disease (NAFLD) is a highly prevalent liver condition and a common cause of liver disease. It is estimated that NAFLD has a global prevalence of approximately 25% (95% CI: 22 – 28) (1, 2). NAFLD is considered a metabolic disease and is strongly associated with cardiovascular disease, insulin resistance (IR), type 2 diabetes (T2D), obesity, dyslipidemia, and hypertension. NAFLD is diagnosed when the accumulation of triglycerides in the liver exceeds 5%, in people without excessive alcohol intake and alternative causes for liver disease, such as hepatitis infection (1). Assessment of triglyceride content in the liver is commonly measured by ultrasonography, due to its low coast and wide availability, but quantitative assessment is mostly performed using proton magnetic resonance spectroscopy (^{1}H-MRS) (3). The term NAFLD covers a wide range of liver damage levels, including minor steatosis to major cirrhosis. Triglyceride deposition occurs depending on, amongst other factors, diet and fasting status (4). Pathological hepatic triglyceride accumulation is the consequence of an imbalance between hepatic uptake of endogenous triglycerides and fatty acids, hepatic triglyceride secretion, de-novo lipogenesis and fatty acid oxidation. Disturbances in these processes are strongly associated with insulin resistance and may also cause further progression of metabolic diseases such as T2D and NAFLD (4). The development of NAFLD as well as the progression to steatosis and cirrhosis varies greatly between individuals. This is due to the complex and multifactorial pathogenesis of fatty liver diseases (5). In addition to environmental acquired factors (6, 7), genetic factors also play an important role (8).

Several studies have been performed to gain insight into the complex etiology of NAFLD by applying metabolomics (9). Changes in circulating metabolites are thought to reflect the composite of both environmental, acquired, and genetic factors of an individual. Metabolomics can thus provide holistic insight to capture the complexity of multi-factorial diseases such as NAFLD (10, 11). Current high-throughput untargeted metabolomic platforms are capable of measuring and mapping over 1000 metabolites from an array of biological pathways from a single biological sample (e.g., blood, urine, or saliva). In addition to well-annotated endogenous metabolites, both xenobiotic metabolites derived from the diet and medications as well as uncharacterized metabolites are reported (12). Although metabolomic analysis has been previously performed in patients with NAFLD (13) most studies had limited sample sizes or focused on a specific subset of metabolites using targeted metabolomics methods (9).

Here we aimed to elucidate the HTGC metabolomic profile as assessed by proton magnetic resonance spectroscopy (^{1}H-MRS), in a middle-aged population (N=496) using an untargeted metabolomics platform (1,363 metabolites). We further created a comprehensive atlas of HTGC-associated metabolites and pathways from our results using two pathway analysis approaches that allows flexible and interactive examination of our results.

3 METHODS

3.1 Study Population

For our present study we included 599 participants from the Netherlands Epidemiology of Obesity (NEO) with available metabolomics data from the general Leiderdorp subpopulation. The Leiderdorp subpopulation was included based on postcode and age (45-65 years) only. We excluded 103 participants who did not undergo direct assessment of the hepatic triglyceride content (HTGC) by ^{1}H-MRS. Therefore, final number of participants included in the present study was 496. The characteristics of the included participants are presented in Table 1. Metabolites were measured using the Metabolon™ Discovery HD4 platform (Metabolon Inc., Durham, North Carolina, USA). In total, 1,363 serum metabolites were measured of which 840 metabolites were from various endogenous pathways (56% lipids, 31% amino acids, 13% other), 227 xenobiotic metabolites, and 296 metabolites that were uncharacterized (unknown chemical structure and biological properties). Further details regarding the study design, HTGC assessment procedure, and metabolite measurements are described in detail in previous works (14, 15) and in the Supplementary Materials.

The NEO study was approved by the Medical Ethics committee of the Leiden University Medical Centre under protocol P08.109. The study is also registered at clinicaltrials.gov under number NL21981.058.08 / P08.109. All participants gave written informed consent (14).

3.2 Statistical Analysis

For the analysis we used multiple linear regressions to test the associations between the metabolites from the untargeted platform with the outcome HTGC. We further adjusted for potential confounding by including sex, age, total body fat, alcohol intake, and lipid lowering medication in the models.

Natural log-transformation was applied to the outcome variable, HTGC, as it was heavily skewed. One individual had a measurement of 0 units of HTGC and was imputed to half the minimum (0.1) before the log-transformation. Missing values in the measured metabolites were imputed using multiple imputation by chained equations as described in our previous work (12). Details regarding the analysis, imputation and scaling of the metabolites are described in the Supplementary Methods.

Additionally, to examine the known sex differences in metabolites, we performed the analysis separately for men and women. We further stratified the women subgroup by menopausal status to examine the metabolomic profile after menopause. Menopausal status was defined as a binary variable based on a questionnaire (Supplementary Material) wherein postmenopausal women were coded as 1 and premenopausal and perimenopausal women were coded as 2. Finally, as sensitivity analysis, we adjusted for HOMA-IR to investigate if the associations of metabolites with HTGC were dependent on IR, particularly for those known to be associated with IR i.e., amino acids and carbohydrates.

3.3 Pathway Analysis

Significant metabolites from the main analysis and the sex stratified analysis were subsequently analyzed with two pathway/network analysis methods: Gaussian Graphical Model (GGM) and Genome Scale Metabolic Model (GSMM). GGM has been used and described in previous studies as a viable approach for the visualization and reconstruction of biological pathways from correlation data. This method is particularly useful for our study and other studies with large metabo-

lomic datasets from untargeted platforms (16). In contrast, GSMM methods are based on a priori defined and curated pathways and have also been used in metabolomic and non-metabolomic studies (17). In addition, we used an inhouse developed GSMM tool to construct the networks. Both methods are thus complementary in their basis, i.e., without and with prior pathway knowledge. Full details regarding the methodology used to create both networks are available in the Supplementary Methods. These interactive networks can be accessed on https://tofaquih.github.io/AtlasLiver/.

4 RESULTS

4.1 Association Analyses of Metabolites and HTGC

Univariate linear regression analyses were performed to examine the associations between the 1,365 metabolites and the outcome HTGC, adjusting for sex, age, total body fat, alcohol intake, and lipid-lowering medication. In total, and after considering multiple testing correction (P-value < 6.59×10^{-5}), 118 metabolites were associated with HTGC, of which 101 were associated with higher levels of HTGC. From these metabolites, many were from the lipids and amino acid classes, as well as a few from the vitamin, nucleotide, carbohydrate classes, uncharacterized metabolites, 1 partially characterized metabolite, and 1 xenobiotic (Supp Table 1). Excluding individuals with lipid lowering medications did not alter the results.

Additional analyses after stratification for sex and menopausal status were performed, which showed complete overlap in the directions of the effects between men and women, between men above and below the age of 60, and women before and after menopause. In general, a higher number of associations present and the effect estimates were larger in women (particularly post-menopause) compared to men, with 82 and 35 metabolites associated with HTGC respectively (Figure 1). As the direction of the effects was identical between men and women for all metabolites (Figure 2), all subsequent descriptions and analyses were performed in the full set of 118 associated metabolites.

4.2 Amino acid and Carbohydrate Metabolism

In total, 37 out of the 264 measured amino acids and peptides were associated with HTGC, of which 34 were associated with higher HTGC levels—particularly BCAA and their keto forms. The amino acids and derivatives with the strongest association were glutamate [β: 2.41 (95% CI 1.93; 2.9)], 3-methyl-2-oxovalerate [β: 2.36 (95% CI 1.75; 2.97)], isoleucine [β: 2.28 (95% CI 1.75; 2.82)], tyrosine [β: 2.03 (95% CI 1.5; 2.55)], and lactoylvaline [β: 1.97 (95% CI 1.5; 2.44)] in addition to the carbohydrates, glucose [β: 1.84 (95% CI 1.21; 2.47)] and pyruvate [β: 1.44 (95% CI 0.93; 1.95)]. Sensitivity analyses adjusting for HOMA- IR were also performed for the insulin resistance related subgroup of amino acid and carbohydrate metabolites (n=264). Accordingly, 33 metabolites were associated with HTGC in this model, of which 32 overlapped with the findings in the main model, with the exception of lysine. Moreover, 10 metabolites, including pyruvate and glucose, were not associated with HTGC in the sensitivity analysis. For the overlapping associations between the two models, the effect of the metabolites on HTGC was weaker after adjusting for HOMA-IR. The findings indicated that the associations of the amino acids and carbohydrates metabolites with the levels of HTGC were overall largely independent from IR. These results are detailed and discussed in the Supplementary Materials and Supplementary Figure 1.

4.3 Lipids

Out of the 475 measured lipid-related metabolites, 62 lipid metabolites belonging to 19 lipid subclasses were associated with HTGC, including phosphatidylcholines (PCs), phosphatidylethanolamines (PEs), phosphatidylinositols (PIs), sphingomyelins, glycerolipids, corticosteroids, ceramides, and dihydroceramides.

Among these lipid metabolites. 52 were associated with higher HTGC levels. Some of these strongly associated lipids were: palmitoyl-oleoyl-glycerol (16:0/18:1) [β: 2.23 (95% CI 1.72; 2.73)], 1-palmitoyl-2-palmitoleoyl-glycerophosphorylcholine (16:0/16:1) [β: 2.16 (95% CI 1.63; 2.69)], myristoyl linoleoyl glycerol (14:0/18:2) [β: 2.12 (95% CI 1.61; 2.63)], and two isomers of diacylglycerol (14:0/18:1, 16:0/16:1) [β: 2.09 (95% CI 1.57; 2.6) and [β: 1.94 (95% CI 1.46; 2.42)]. All dihydrosphingomyelin species had positive associations with HTGC of which three metabolites had a P-value above the significance threshold, with the most salient being sphingomyelin (d18:0/18:0, d19:0/17:0) [β: 2.15 (95% CI 1.63; 2.66)].

Ten lipid metabolites were associated with lower levels of HTGC, amongst which were several ether-PC species and glycosylceramide (HCER) and lactosylceramide (LCER) species. Among those metabolites were two sphingomyelins with long fatty acyl chains—sphingomyelin (d18:2/24:1, d18:1/24:2) [β: -1.29 (95% CI -1.79; -0.78)] and sphingomyelin (d18:2/24:2) [β: -1.68 (95% CI -2.27; -1.09)]. For the HCER and LCER metabolites, the HCER glycosylceramide (d18:2/24:1, d18:1/24:2) [β: -1.31 (95% CI -1.87; -0.74)] and two LCER metabolites lactosyl-N-nervonoyl-sphingosine (d18:1/24:1) [β: -1.57 (95% CI -2.06; -1.08)] and lactosyl-N-palmitoyl-sphingosine (d18:1/16:0) [β: -1.17 (95% CI -1.68; -0.65)] reduced HTGC levels. Finally, glycerophosphorylcholine (GPC), a derivate of choline and a breakdown product of PCs, reduced HTGC levels as well [β: -1.13 (95% CI -1.68; -0.59)].

4.4 Other Metabolites

Metabolonic lactone sulfate, a partially characterized metabolite, was found to be strongly associated with higher HTGC levels [β: 2.18 (95% CI 1.73; 2.63)]. Only one xenobiotic metabolite, 4-ethylcatechol, was associated with HTGC and it was strongly associated with the reduction of HTGC levels [β: -1.63 (95% CI -2.43; -0.84]. Ten uncharacterized metabolites were associated with higher HTGC. Other metabolites included two nucleotides: xanthine and guanine. Xanthine was associated with the higher HTGC levels [β: 1.25 (95% CI 0.76; 1.75] while guanine had the opposite effect [β: -1.09 (95% CI -1.55; -0.63]. Finally, tocopherol metabolites including alpha-tocopherol [β: 1.27 (95% CI 0.77; 1.78] and gamma-tocopherol/beta-tocopherol [β: 1.18 (95% CI-0.62; 1.73] also associated with higher HTGC levels.

4.5 Correlation Based Network Analysis Using GGM

GGM networks were generated including all significantly (P-value < 6.59×10^{-5}) associated metabolites (N=118). This method identified two major clusters: amino acids and lipids (Figure 3).

The amino acids and related metabolites can be organized into 3 groups: 1) Primary amino acids e.g., glutamate; 2) Derivates of amino acids with either a carbon group (lactoylvaline), oxo- and methyl groups (e.g., 3-methyl-2-oxovalerate) or an acetyl group (e.g., N-acetyltryptophan); 3) Ketoacids, products of incomplete breakdown of amino acids (e.g., 4-methyl-2-oxopentanoate).

Overall, the network pattern shows the primary BCAA connecting to each other via the amino acids derivates and ketoacids. The common endpoint of these connections led to glutamate,

pyruvate, and glucose. Moreover, all uncharacterized metabolites related to BCAAs were intermediates for glucose, pyruvate, and alpha ketoglutarate (AKG). Some amino acids were correlated with metabolites in the lipid cluster such as the connection between leucine and N-palmitoyl-sphinganine (d18:0/16:0).

Regarding the lipids cluster, the GGM network showed strong interconnectivity between diacylglycerols, monoacylglycerols, as well as links to PIs, PCs, and PEs (https://tofaquih.github.io/AtlasLiver/Networks/LiverFatNetworks/Model2/SexAdj/network/). Moreover, all 10 of the negatively associated lipids—which included sphingomyelins, ether PC, and glucosyl- and lactosylceramide species—were connected to each other and formed a subcluster within the lipids. The partially characterized metabolite metabolonic lactone sulfate showed a positive correlation with glutamate only in the stratified analyses in men and post-menopausal women (data not shown). Alpha-tocopherol and gamma-tocopherol/beta-tocopherol were correlated and connected in the GMM network and, interestingly, alpha-tocopherol was also connected to cholesterol.

4.6 GSMM Pathway-based Analyses

To assess how and whether HTGC associated metabolites were linked biochemically to one another through metabolic pathways, a network analysis based on the genome scale metabolic model HRM2 was performed (Figure 4). The network showed several amino acids subclusters that were associated with higher HTGC, amongst which the BCAAs and their keto and 2-hydroxy form, the clusters of aromatic amino acids and intermediates, and glutamate and alpha-ketoglutarate. The lipid cluster showed the relation between the positively and negatively correlated lipids, providing insight into the enzymatic conversions which are potentially affected by HTGC. The pathway mapping of the 10 lipid metabolites that were associated with lower level HTGC reported in GGM network were also confirmed in the GSMM. Biochemical reactions were found to connect sphingomyelin species to the glucosylceramide and lactosylceramide species (labeled as LacCer in the GSMM network) via the ceramides. An additional connection not shown in the GSMM network included the reactions connecting glucosylceramide to glucose via its breakdown to ceramide and glucose.

To improve the interpretability of the network, only the main lipid classes were shown as nodes. The full network of associated metabolites is available as an interactive, single page website: https://tofaquih.github.io/AtlasLiver/Networks/LiverFatNetworks/.

5 DISCUSSION

In the present work, we establish a comprehensive atlas of metabolites associated with HTGC in a Dutch population composed of middle-aged men and women. We used an untargeted metabolomics approach to measure 1361 metabolites and ^{1}H-MRS measurement of HTGC. In total, 118 metabolites were associated with HTGC after correction for sex and other confounders, of which 101 associated with higher levels of HTGC. Stratification by sex and menopause in women revealed that a larger number of associations was significant in the women strata and the estimates of the effect sizes were higher in women, particularly post-menopause. However, the directions of the effects for the overlapping metabolite-HTGC associations were uniform in all strata. This difference was observed despite a higher average level of HTGC in men. Apparently, metabolites have strong correlations even with smaller HTGC ranges. Pathway analyses revealed 2 clusters of interrelated metabolites, one primarily involving amino acid related metabolism and the other lipid metabolism.

5.1 91. Branched Chain Amino Acids and Branched Chain Keto Acids

Our analysis showed that amino acids levels were overall associated with higher levels of HTGC. Moreover, these amino acids had clear clustering in both the GGM- and the GSMM-based pathway network analyses. These results coincided with previous literature findings that elevated levels of BCAAs associated with HTGC and NAFLD (9, 18, 19). The primary metabolites from the amino acids associated with higher HTGC were glutamate, leucine, isoleucine, and valine. In addition, untargeted metabolomics enabled us to screen metabolite derivatives and subclasses, several of which we also found to be strongly associated with the higher levels of HTGC. These derivatives form during normal or abnormal breakdown of BCAAs and include BCAAs with acetyl, lactoyl or methyl groups, as well as metabolites categorized as branched chain keto acids (BCKA), their 2-hydroxy form, and gamma-glutamyl alpha-amino acids. Furthermore, these metabolites were shown in the GGM and GSMM at the intersection of the BCAAs, glutamate, and pyruvate. The derivatives, particularly BCKAs, are an indication of an association between BCAA catabolism with HTGC levels (20, 21). It is interesting that the BCAAs and derivates, known to be associated with body fat and insulin resistance (20), remained associated with HTCG despite the adjustment for total body fat and after further adjustment for HOMA-IR in the sensitivity analysis. These results indicate that these associations were overall independent from body fat and IR. A few amino acids that were previously found to be associated with HTGC and NAFLD using ultrasonography assessment did not replicate in our study. For example, serine and glycine were not associated with HTGC in our analysis contrary to the results of previous studies (9, 22). Moreover, the proposed glutamate-serine-glycine index in those studies was also not observed in our analysis despite our larger sample size (N= 496 versus N=64 [N=20 controls]). This discrepancy could possibly be due to our study included participants from the general population with an HTGC range around 5.6% and did not focus on NAFLD or NASH patients specifically in a case-control study design. Therefore, it is possible that serine and glycine are better markers for advanced stages of HTGC, as observed in NAFLD/NASH patients, due to the stronger effect of metabolic dysregulation, but are not strong markers for general levels of HTGC.

5.2 92. Lipid Metabolites

Lipids have the largest effect estimates in the GGM and GSMM networks and reflect a strong association of lipid metabolism with HTGC. This cluster contained lipids from various subclasses of which di- and mono-glycerides were predominant. These associations with HTGC, particularly from di- and mono-glycerides, are supported by previous studies (23). The GGM network showed strong correlations between diglycerols, PCs, PEs, PIs, and other lipid subclasses, consistent with our GSMM based network and the established biological pathways (24). Overall, most of the lipids (N=52 of 62) were associated with higher HTGC levels. Notable exceptions were the group of 10 lipids comprising of ether-PC species, sphingomyelin, HCER, and LCER species, which the highest reduction effect on HTGC levels in our results. In addition, these lipids were highly connected in the GMM network and shared the same biological pathways in the GSMM network. Interestingly, the aforementioned sphingomyelins with long fatty acyl chains reduced HTGC levels, in contrast to dihydrosphingomyelins which associated with higher HTGC. Other interesting metabolite groups were the beforementioned HCER and LCER. Unlike the ceramide and dihydroceramide metabolites, which associated with higher HTGC level, both HCER and LCER had the opposite effect. In this case, the breakdown of these metabolites seems to contribute to the upregulation of ceramides and dihydroceramides, as well as glucose, specifically via HCER (25).

The associations of the ceramides and dihydroceramides with higher HTCG are analogous to previous studies on NAFLD (26-28). However, the associations of sphingomyelins and dihydrosphingomyelins with HTGC and NAFLD in humans are less studied and understood compared to the associations of ceramides (29). Since dihydroceramides are precursors for the synthesis of dihydrosphingomyelins (28, 30), it could explain their association with higher HTGC. However, a small study by Lovric et al. (N=75) (31) reported contrary pattern using the same metabolomic platform used in our study, in which higher sphingomyelins but lower dihydrosphingomyelins were associated with NAFLD. In mice, dietary sphingomyelins were associated with a decrease of HTGC in the liver (32). Regarding HCER and LCER, as stated earlier, HCER is involved in the synthesis of LCER, ceramide, and glucose (25). In addition, both LCER and HCER are involved in glycosphingolipid metabolism (33). One small study (n=28) reported a positive association HCER and LCER with NASH (29). Finally, a recent a study reported similar results of higher dihydrosphingolipid classes being associated with increased fibrosis in animal models and NAFLD patients (34).

Overall, our study presents a deeper look into these metabolite subclasses in a larger sample size in a general population. Further studies focusing on the contrasting associations of sphingomyelins versus dihydrosphingomyelins and ceramide versus HCER could elucidate their specific function and relationship to HTGC and NAFLD in humans.

5.3 Other Metabolites Strongly Associated with HTGC

In addition to the amino acids and lipids results, our analysis and the GGM network included uncharacterized and xenobiotic metabolites, not previously reported in HTGC or NAFLD studies. A particularly interesting finding was the relatively novel metabolite metabolonic lactone sulfate (formerly assigned the ID X-12063). Metabolonic lactone sulfate had a large effect estimate in the adjusted and the stratified models and was associated with higher HTGC levels. Previous studies have found this metabolite to be a biomarker and a predictor for T2D (35-37) and acute-on-chronic liver failure (38). Moreover, metabolonic lactone sulfate was associated with cardiometabolic disease (39) and was positively correlated with BMI, waist hip ratio, and HOMA-IR (40). Metabolonic lactone sulfate was also reported to be associated with the *CYP3A5*/*ZSCAN25* locus via the rs10242455 (41) and rs7808022 (42) single nucleotide polymorphisms (SNPs), respectively. This genes in this region are highly expressed in the liver and share the same regulatory promoters (43). The gene *CYP3A* is particularly expressed in the liver and encodes for the CYP3A protein, a key drug-metabolizing liver enzyme (39, 43). Although the specific functionality and underlying biological pathways of metabolonic lactone sulfate remains unelucidated, the available evidence suggests a link with the liver function, cardiometabolic diseases. This supports our findings regarding the association of metabolonic lactone sulfate with higher HTGC levels and possibly NAFLD.

4-Ethylcatechol is a xenobiotic metabolite that is primarily acquired from the ingestion of coffee beverages and products (44, 45). In our analysis, 4-Ethylcatechol was associated with lower HTGC levels and was not connected to any metabolites in the GGM or GSMM networks. Protective properties of coffee and caffeine intake against liver fibrosis have been suggested before due to its anti-fibrotic and antioxidant effects (46, 47). Several studies examining the association of coffee consumption with liver fat found similarly reduced HTGC (46, 48, 49). For instance, a recent large systematic review and meta-analysis study found that coffee consumption was negatively associated with liver fibrosis and suggested protection from severe liver fibrosis and cirrhosis (46). Results for patients with NAFLD were less conclusive with some showing that increased coffee consumption was associated with reduction of NAFLD (49).

Uncharacterized metabolites, such as X-15245, X-24295, X-19438, and X-25343, are associated with HTGC with strong connections to pyruvate, glucose, and AKG. Further analysis into the structure of these metabolites is needed to determine their identities and their biological relationships to HTGC and the other metabolites in the network.

Our analysis also shows associations with vitamin E metabolites. Briefly, vitamin E metabolites (alpha-tocopherol and gamma-tocopherol/beta-tocopherol) are positively associated with HTGC, and alpha-tocopherol was directly correlated with cholesterol in the GGM network. This finding is supported by one study that found a positive correlation between vitamin E and NAFLD (50). Moreover, vitamin E is known to bind to lipoproteins in the blood which promoted the usage of cholesterol-adjusted vitamin E in several studies as a superior measurement of vitamin E (51). However, the mechanism linking vitamin E with HTGC remains unclear and clinical trials present mixed results regarding the relationships. Some studies have shown a negative or no relation while others suggested possibly therapeutic benefits of vitamin E supplementation for NAFLD and NASH patients via the suppression of HTGC (52, 53).

5.4 Pathway Analysis

Pathway analysis was performed using two approaches. The first approach—GMM—is data-driven and calculates the partial correlations between the HTGC associated metabolites to create a network. The second approach maps metabolites to a GSMM, which consists of known functionally annotated biochemical conversions that can occur in humans. An advantage of the GGM is that all measured and associated metabolites in the study are included in the network, even the uncharacterized and xenobiotic metabolites ones. However, a GGM is data driven and does not necessarily reflect actual biological pathways, and only shows metabolites measured by the platform. That is, metabolites can be directly linked in the GGM even though they are distant in terms of intermediate biochemical reactions (16). The GSMM based network analysis, on the other hand, includes measured associated metabolites as well as relevant intermediate metabolites regardless of whether they were measured or associated with HTGC. In addition, the created network shows the directionality of the biological reactions for the biosynthesis and degradation of the metabolites and provides relevant details regarding the enzymes involved in each reaction.

Although both types of pathway analysis are complementary in their approach, we found that the network resulting from GGM had a good alignment with the metabolic reaction paths that resulted from the GSMM approach. This is in concordance with previous work (16), which showed that strongly associated metabolites generally corresponded to the same pathways.

5.5 Key Findings and Potential Targets for Genetic and Drug Research

In summary, our analyses and atlas showed interesting pathway associations for HTGC which may be relevant for fatty liver disease research. The pathways connecting glutamate, BCAAs, and derivatives of BCAAs during normal/abnormal metabolism were particularly associated with the higher HTGC levels. Among the wide variety of lipid metabolites, higher metabolite concentrations in the pathways connecting diglycerols, PCs, PEs, and PIs had a strong association with higher HTGC levels. Notably, the pathways connecting the ceramide species and their relationship with sphingomyelin species, as shown in the GMM and GSMM networks, were particularly interesting due to the contrasting associations of HTGC. Thus, it appears that the metabolomic flux between these metabolite species is associated with an overall higher HTGC levels. These metabolites and pathways are of interest to explore in future HTGC and liver fibrosis studies.

The most interesting finding was the in our study was the strong association of glutamate and the novel metabolites correlated with it. Glutamate itself is a well-known biomarker for liver fat and NAFLD (9, 22). However, our pathway analyses revealed 2 metabolites to be strongly correlated with glutamate and were also shown to be associated with HTGC. These were metabolonic lactone sulfate and the uncharacterized metabolite X-15245. These metabolites were associated with HTGC in all models and specifically interconnected with glutamate in the GGM networks for men and postmenopausal women, hence indicating that these metabolites share common biochemical pathways. As discussed earlier, metabolonic lactone sulfate has been found to be associated with several cardiometabolic related outcomes in other studies. For example, the uncharacterized metabolite X-15245 is associated with rs1260326 SNP in the *GCKR* gene (42). This particular SNP was reported to be strongly associated with NAFLD (8). Based on these findings, glutamate, metabolonic lactone sulfate and X-15245 and their potentially shared pathway are important candidates for further etiological studies on HTGC and NAFLD. Furthermore, the genetic SNPs associated with these metabolites can be used as possible new genetic marker for HTGC/NAFLD. This can be achieved similar to the approach used by Mancina et al. (54), in which SNPs associated with triglycerides were tested for their association with liver fat and subsequently used to identify a protective link between *PSD3* and HTGC. Moreover, future studies should aim to identify the uncharacterized metabolite X-15245 and elucidate the biological properties for it and metabolonic lactone sulfate.

Exploration of the aforementioned metabolites and the various other metabolites and pathways we have discussed here, in combination with metabolomic and genetic studies, can be key to identifying causal associations and possible drug targets for liver fibrosis in the future.

5.6 Strengths and Limitations

Previous literature on metabolite-HTGC associations focused mainly on a small number of well-established metabolites or metabolites involved in specific pathways. A strength of our study is the use of an untargeted metabolomics platform with over a thousand measured metabolites from 10 metabolite pathway classes in a relatively large population of middle-aged individuals in the Netherlands. Our study population was a random selection of volunteers from the Leiderdorp area and was not selected on NAFLD or NASH diagnosis. Moreover, HTCG measurements in this cohort were assessed by ^{1}H-MRS, which provides high accuracy and sensitivity in measuring HTGC even at low levels (55). Furthermore, we expanded our analysis by combining linear regression with correlation and biochemical pathway analysis methods to construct a comprehensive atlas of metabolomic profiles of HTGC and an atlas for men and women separately. A limitation of our study was the selection of only individuals of white ethnicity from a high social economic status area, which limits generalizability to other ethnicities and social status. Another limitation of our study is the lack of biological validation. Future studies with liver biopsy samples will aid in validating our findings. However, studies on NAFLD are extensive but usually focused on specific metabolites such as amino acids or lipids. Our study itself can be considered as a validation of the various results from the previous literature by taking advantage of the wide range of the metabolites from various biochemical pathway classes. The GSMM network provided useful insight into the biochemical relation between the metabolites associated with HTGC and their intermediates. The directionality of the edges in the network, however, is based on knowledge regarding the thermodynamic reversibility of the corresponding reactions and makes no assertion about causality or potential association with HTGC. Although the GSMM network provides some information regarding known gene-pathway associations, genetic analysis for some HTGC or metabolite related SNPs would have benefited our study. Finally, further investigation for sex differences in the number of

metabolite associations with HTGC levels, particularly for postmenopausal women, requires a larger longitudinal study for validation.

6 CONCLUSION

In this study, we conducted a cross sectional analysis in 496 middle aged men and women to gain insight the metabolomic profile associated with hepatic triglyceride accumulation as assessed by ^{1}H-MRS. We used a hypothesis-free approach to study a myriad of metabolites associated with HTGC using an untargeted platform that measured 1,363 metabolites. Using this platform, associations were found between 118 well-known, lesser-known, and novel metabolites with HTGC levels. These findings were combined by pathway analyses using a correlation driven network (GGM) and a biologically driven network (GSMM) to create an atlas of metabolites associated with HTGC. Analysis of these networks indicated strong associations between the BCAAs, diglycerols, ceramides, and sphingomyelins pathways with HTGC levels. These pathways were additionally found to reject the null hypothesis of the closed global test when using the FIB-4 index as the outcome. In addition, our atlas of networks, enriched with pathway knowledge, provided interesting insights regarding pathways associated with HTGC. These included the pathways connecting BCAA and BCAA derivates, the flux between the ceramide species (i.e., HCER and LCER), sphingomyelins and their dihydro forms, and the potential novel pathway linking glutamate with the novel metabolites metabolonic lactone sulfate and X-15245. Thus, our atlas is potentially essential for understanding metabolomic profiles associated with liver fat accumulation. In turn facilitating further studies to find causal links between the metabolites reported here with liver fibrosis.

7 REFERENCES

1. Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology 2016;64:73-84.
2. Angulo P. Nonalcoholic Fatty Liver Disease. 2002;346:1221-1231.
3. Stern C, Castera L. Non-invasive diagnosis of hepatic steatosis. Hepatology International 2017;11:70-78.
4. Kawano Y, Cohen DE. Mechanisms of hepatic triglyceride accumulation in non-alcoholic fatty liver disease. Journal of Gastroenterology 2013;48:434-441.
5. Byrne CD, Targher G. NAFLD: A multisystem disease. Journal of Hepatology 2015;62:S47-S64.
6. Zou B, Yeo YH, Nguyen VH, Cheung R, Ingelsson E, Nguyen MH. Prevalence, characteristics and mortality outcomes of obese, nonobese and lean NAFLD in the United States, 1999–2016. 2020;288:139-151.
7. Ballestri S, Nascimbeni F, Baldelli E, Marrazzo A, Romagnoli D, Lonardo A. NAFLD as a Sexual Dimorphic Disease: Role of Gender and Reproductive Status in the Development and Progression of Nonalcoholic Fatty Liver Disease and Inherent Cardiovascular Risk. Advances in therapy 2017;34:1291-1326.
8. Speliotes EK, Yerges-Armstrong LM, Wu J, Hernaez R, Kim LJ, Palmer CD, Gudnason V, et al. Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. PLoS Genet 2011;7:e1001324.
9. Gaggini M, Carli F, Rosso C, Buzzigoli E, Marietti M, Della Latta V, Ciociaro D, et al. Altered amino acid concentrations in NAFLD: Impact of obesity and insulin resistance. Hepatology 2018;67:145-158.
10. Fearnley LG, Inouye M. Metabolomics in epidemiology: from metabolite concentrations to integrative reaction networks. International Journal of Epidemiology 2016;45:1319-1328.
11. Alonso A, Marsal S, Julia A. Analytical methods in untargeted metabolomics: state of the art in 2015. Front Bioeng Biotechnol 2015;3:23.
12. Faquih T, van Smeden M, Luo J, le Cessie S, Kastenmüller G, Krumsiek J, Noordam R, et al. A Workflow for Missing Values Imputation of Untargeted Metabolomics Data. Metabolites 2020;10.
13. Perakakis N, Stefanakis K, Mantzoros CS. The role of omics in the pathophysiology, diagnosis and treatment of non-alcoholic fatty liver disease. Metabolism 2020;111:154320.
14. De Mutsert R, Den Heijer M, Rabelink TJ, Smit JWA, Romijn JA, Jukema JW, De Roos A, et al. The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection. European Journal of Epidemiology 2013;28:513-523.
15. Boone S, Mook-Kanamori D, Rosendaal F, Den Heijer M, Lamb H, De Roos A, Le Cessie S, et al. Metabolomics: a search for biomarkers of visceral fat and liver fat content. Metabolomics 2019;15.
16. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Systems Biology 2011;5:21.
17. Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. Current status and applications of genome-scale metabolic models. Genome Biology 2019;20:121.
18. Kalhan SC, Guo L, Edmison J, Dasarathy S, McCullough AJ, Hanson RW, Milburn M. Plasma metabolomic profile in nonalcoholic fatty liver disease. Metabolism: clinical and experimental 2011;60:404-413.
19. Sookoian S, Castaño GO, Scian R, Fernández Gianotti T, Dopazo H, Rohr C, Gaj G, et al. Serum aminotransferases in nonalcoholic fatty liver disease are a signature of liver metabolic perturbations at the amino acid and Krebs cycle level. Am J Clin Nutr 2016;103:422-434.
20. Lynch CJ, Adams SH. Branched-chain amino acids in metabolic signalling and insulin resistance. Nature Reviews Endocrinology 2014;10:723-736.
21. Shimomura Y, Honda T, Shiraki M, Murakami T, Sato J, Kobayashi H, Mawatari K, et al. Branched-chain amino acid catabolism in exercise and liver disease. J Nutr 2006;136:250s-253s.
22. Leonetti S, Herzog RI, Caprio S, Santoro N, Tricò D. Glutamate-Serine-Glycine Index: A Novel Potential Biomarker in Pediatric Non-Alcoholic Fatty Liver Disease. Children (Basel) 2020;7.

23. Puri P, Baillie RA, Wiest MM, Mirshahi F, Choudhury J, Cheung O, Sargeant C, et al. A lipidomic analysis of nonalcoholic fatty liver disease. Hepatology 2007;46:1081-1090.
24. Berg J, Tymoczko J, Stryer L. Biochemistry. New York: W H Freeman, 2002.
25. Ichikawa S, Hirabayashi Y. Glucosylceramide synthase and glycosphingolipid synthesis. Trends in Cell Biology 1998;8:198-202.
26. Régnier M, Polizzi A, Guillou H, Loiseau N. Sphingolipid metabolism in non-alcoholic fatty liver diseases. Biochimie 2019;159:9-22.
27. Carlier A, Phan F, Szpigel A, Hajduch E, Salem J-E, Gautheron J, Le Goff W, et al. Dihydroceramides in Triglyceride-Enriched VLDL Are Associated with Nonalcoholic Fatty Liver Disease Severity in Type 2 Diabetes. Cell Reports Medicine 2020;1:100154.
28. Magaye RR, Savira F, Hua Y, Kelly DJ, Reid C, Flynn B, Liew D, et al. The role of dihydrosphingolipids in disease. Cellular and Molecular Life Sciences 2019;76:1107-1134.
29. Apostolopoulou M, Gordillo R, Koliaki C, Gancheva S, Jelenik T, De Filippo E, Herder C, et al. Specific Hepatic Sphingolipids Relate to Insulin Resistance, Oxidative Stress, and Inflammation in Nonalcoholic Steatohepatitis. Diabetes Care 2018;41:1235-1243.
30. Lachkar F, Ferré P, Foufelle F, Papaioannou A. Dihydroceramides: their emerging physiological roles and functions in cancer and metabolic diseases. 2021;320:E122-E130.
31. Lovric A, Granér M, Bjornson E, Arif M, Benfeitas R, Nyman K, Ståhlman M, et al. Characterization of different fat depots in NAFLD using inflammation-associated proteome, lipidome and metabolome. Scientific reports 2018;8:14200-14200.
32. Chung RWS, Kamili A, Tandy S, Weir JM, Gaire R, Wong G, Meikle PJ, et al. Dietary Sphingomyelin Lowers Hepatic Lipid Levels and Inhibits Intestinal Cholesterol Absorption in High-Fat-Fed Mice. PLoS ONE 2013;8:e55949.
33. Mullen Thomas D, Hannun Yusuf A, Obeid Lina M. Ceramide synthases at the centre of sphingolipid metabolism and biology. Biochemical Journal 2012;441:789-802.
34. Babiy B, Ramos-Molina B, Ocaña L, Sacristán S, Burgos-Santamaría D, Martínez-Botas J, Villa-Turégano G, et al. Accumulation of dihydrosphingolipids and neutral lipids is related to steatosis and fibrosis damage in human and animal models of non-alcoholic fatty liver disease. 2022:2022.2003.2010.22271048.
35. Diboun I, Al-Mansoori L, Al-Jaber H, Albagha O, Elrayess MA. Metabolomics of Lean/Overweight Insulin-Resistant Females Reveals Alterations in Steroids and Fatty Acids. The Journal of Clinical Endocrinology & Metabolism 2020;106:e638-e649.
36. Peddinti G, Cobb J, Yengo L, Froguel P, Kravić J, Balkau B, Tuomi T, et al. Early metabolic markers identify potential targets for the prevention of type 2 diabetes. Diabetologia 2017;60:1740-1750.
37. di Giuseppe R, Koch M, Nöthlings U, Kastenmüller G, Artati A, Adamski J, Jacobs G, et al. Metabolomics signature associated with circulating serum selenoprotein P levels. Endocrine 2019;64:486-495.
38. Bajaj JS, Reddy KR, O'Leary JG, Vargas HE, Lai JC, Kamath PS, Tandon P, et al. Serum Levels of Metabolites Produced by Intestinal Microbes and Lipid Moieties Independently Associated With Acute-on-Chronic Liver Failure and Death in Patients With Cirrhosis. Gastroenterology 2020;159:1715-1730.e1712.
39. Das SK, Ainsworth HC, Dimitrov L, Okut H, Comeau ME, Sharma N, Ng MCY, et al. Metabolomic architecture of obesity implicates metabolonic lactone sulfate in cardiometabolic disease. Mol Metab 2021;54:101342.
40. Darst BF, Lu Q, Johnson SC, Engelman CD. Integrated analysis of genomics, longitudinal metabolomics, and Alzheimer's risk factors among 1,111 cohort participants. Genetic Epidemiology 2019;43:657-674.
41. Yin X, Chan LS, Bose D, Jackson AU, VandeHaar P, Locke AE, Fuchsberger C, et al. Genome-wide association studies of metabolites in Finnish men identify disease-relevant loci. Nature Communications 2022;13:1644.
42. Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J, Arnold M, et al. An atlas of genetic influences on human blood metabolites. Nat Genet 2014;46:543-550.

43. Collins JM, Wang D. Cis-acting regulatory elements regulating CYP3A4 transcription in human liver. Pharmacogenet Genomics 2020;30:107-116.
44. Miranda AM, Steluti J, Fisberg RM, Marchioni DM. Dietary intake and food contributors of polyphenols in adults and elderly adults of Sao Paulo: a population-based study. British Journal of Nutrition 2016;115:1061-1070.
45. Lang R, Mueller C, Hofmann T. Development of a Stable Isotope Dilution Analysis with Liquid Chromatography–Tandem Mass Spectrometry Detection for the Quantitative Analysis of Di- and Trihydroxybenzenes in Foods and Model Systems. Journal of Agricultural and Food Chemistry 2006;54:5755-5762.
46. Ebadi M, Ip S, Bhanji RA, Montano-Loza AJ. Effect of Coffee Consumption on Non-Alcoholic Fatty Liver Disease Incidence, Prevalence and Risk of Significant Liver Fibrosis: Systematic Review with Meta-Analysis of Observational Studies. Nutrients 2021;13:3042.
47. Dranoff JA. Coffee Consumption and Prevention of Cirrhosis: In Support of the Caffeine Hypothesis. Gene Expression 2018;18:1-3.
48. Kennedy OJ, Fallowfield JA, Poole R, Hayes PC, Parkes J, Roderick PJ. All coffee types decrease the risk of adverse clinical outcomes in chronic liver disease: a UK Biobank study. BMC Public Health 2021;21:970.
49. Chung HK, Nam JS, Lee MY, Kim YB, Won YS, Song WJ, Kim YH, et al. The increased amount of coffee consumption lowers the incidence of fatty liver disease in Korean men. Nutr Metab Cardiovasc Dis 2020;30:1653-1661.
50. Jeon D, Son M, Shim J. Dynamics of Serum Retinol and Alpha-Tocopherol Levels According to Non-Alcoholic Fatty Liver Disease Status. 2021;13:1720.
51. Thurnham DI, Davies JA, Crump BJ, Situnayake RD, Davis M. The use of different lipids to express serum tocopherol: lipid ratios for the measurement of vitamin E status. Ann Clin Biochem 1986;23 (Pt 5):514-520.
52. Nagashimada M, Ota T. Role of vitamin E in nonalcoholic fatty liver disease. 2019;71:516-522.
53. Amanullah I, Khan YH, Anwar I, Gulzar A, Mallhi TH, Raja AA. Effect of vitamin E in non-alcoholic fatty liver disease: a systematic review and meta-analysis of randomised controlled trials. Postgrad Med J 2019;95:601-611.
54. Mancina RM, Sasidharan K, Lindblom A, Wei Y, Ciociola E, Jamialahmadi O, Pingitore P, et al. PSD3 downregulation confers protection against fatty liver disease. Nature Metabolism 2022;4:60-75.
55. Springer F. Liver fat content determined by magnetic resonance imaging and spectroscopy. World Journal of Gastroenterology 2010;16:1560.

5

Table 1: Characteristics of the participants from Leiderdorp with metabolomics and HTGC measurement. Continuous variables are represented by mean (SD) unless stated otherwise; dichotomous variables are represented by percentage (%). Abbreviations: HTGC, hepatic triglyceride content; IQR, interquartile range; CVD, cardiovascular disease; HOMA1-IR, homeostatic model assessment index for insulin resistance.

	Total	Men	Women	Women	Women
		All	All	Postmenopausal	Premenopausal
n	496	233	263	159	104
Age (years)	55.8 (6)	55.9 (6.2)	55.6 (5.8)	59.4 (3.8)	49.9 (2.7)
HTGC (mean; median [IQR])	6.1; 2.74 [1.36, 6.75]	7.5; 4.18 [2.18, 9.74]	4.8; 1.79 [1.08, 4.27]	5.5; 2.09 [1.23, 5.90]	3.8; 1.36 [0.89, 3.17]
HTGC >= 5.56 (%)	153 (30.8)	96 (41.2)	57 (21.7)	43 (27.0)	14 (13.5)
BMI (kg/m^2)	25.9 (4.1)	26.6 (3.4)	25.3 (4.5)	25.5 (4.4)	25.0 (4.7)
Total Body Fat	30.6 (8.3)	24.5 (5.1)	36.1 (6.5)	35.2 (6.7)	36.6 (6.4)
Aspartate Transaminase (IU/L)	24.4 (6.6)	25.6 (6.2)	23.2 (6.6)	24.0 (5.9)	22.0 (7.5)
Alanine Aminotransferase (IU/L)	25.3 (51.4)	28.9 (11.8)	22.2 (9.2)	22.8 (7.7)	21.2 (11.0)
Platelet Count (10^9/L)	236.2 (11.0)	219.1 (47.2)	252.3 (50.2)	251.1 (492)	254.1 (51.8)
Alcohol Consumption (g/day)	14.2 (15.9)	19.5 (19.3)	9.4 (10)	9.7 (9.8)	8.9 (10.3)
Smoking (%)					
Never	202 (40.7)	89 (38.2)	113 (43.0)	59 (37.1)	54 (51.9)
Former	237 (47.8)	117 (50.2)	120 (45.6)	81 (50.9)	39 (37.5)
Current	57 (11.5)	27 (11.6)	30 (11.4)	19 (11.9)	11 (10.6)
HOMA1-IR	2.6 (2.4)	2.92 (2.9)	2.2 (1.7)	2.4 (1.7)	2.02 (1.8)
Hypertension (%)	183 (36.9)	96 (41.2)	87 (33.1)	59 (37.1)	28 (26.9)
CVD (%)	21 (4.3)	11 (4.7)	10 (3.8)	8 (5.1)	2 (1.9)
Fasting plasma glucose (mmol/L)	5.5 (1.0)	5.63 (1.2)	5.3 (0.9)	5.5 (0.9)	5.09 (0.5)
Serum triglycerides (mmol/L)	1.2 (0.8)	1.5 (0.9)	1.0 (0.6)	1.1 (0.7)	0.9 (0.5)
LDL (mmol/L)	3.6 (1.0)	3.6 (0.9)	3.6 (1.0)	3.4 (0.8)	3.4 (0.7)
HDL (mmol/L)	1.6 (0.5)	1.3 (0.3)	1.8 (0.4)	1.8 (0.4)	1.8 (0.4)
Cholesterol (mmol/L)	5.7 (1.1)	5.6 (1.0)	5.8 (1.1)	5.3 (0.8)	5.6 (0.8)
Hypertension Medication (%)	95 (19.2)	43 (18.5)	52 (19.8)	39 (24.5)	13 (12.5)
Lipid Lowering Medication (%)	41 (8.3)	26 (11.2)	15 (5.7)	15 (9.4)	0 (0.0)

Figure 1: Comparison of the point estimates and confidence intervals for the 118 metabolites associated with HTGC in the primary analysis and the sex stratified analysis. The first ring shows the effect estimates in the sex combined analysis. The second ring shows analysis in women; and the final ring is for the men stratified analysis. Hollow circles are estimates that are not significant (p-value ≤ 6.5E^{-5}). Metabolites are grouped by super pathway in each section.

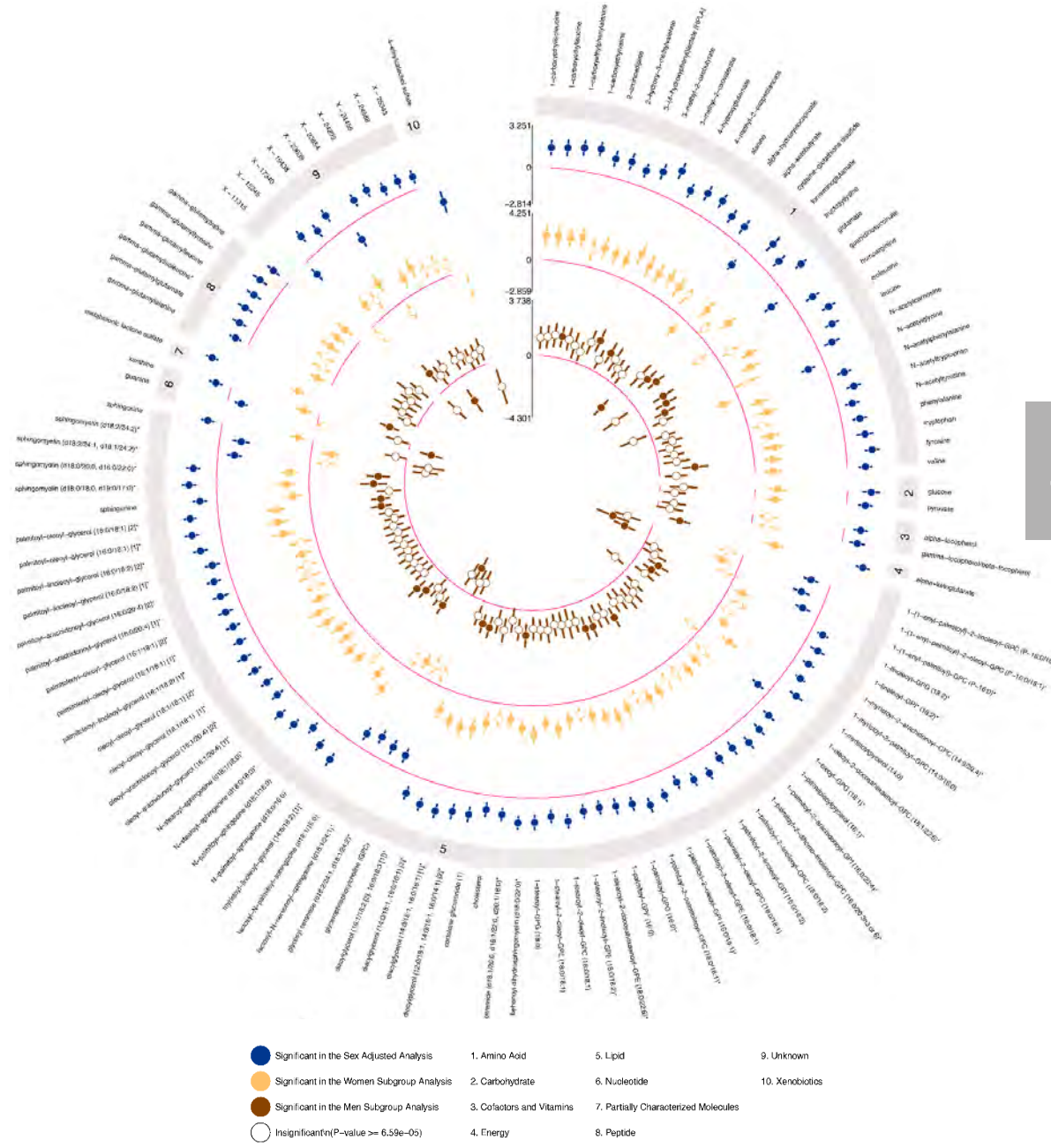


Figure 2 Beta-Beta plot comparing the effect estimates of 118 metabolites associated with HTGC. Overall, effect estimates of the metabolites were stronger in women than men.

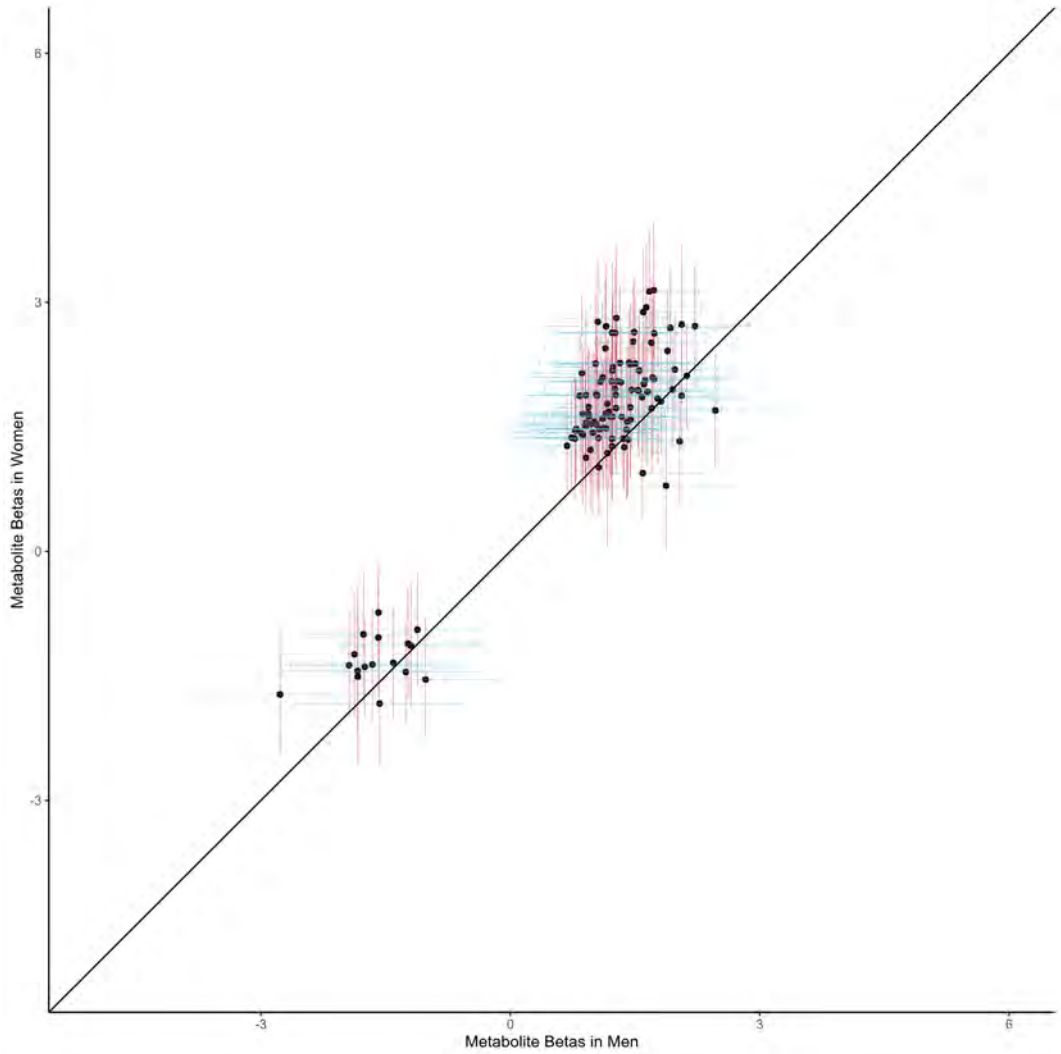


Figure 3: Gaussian Graphical Model for the sex adjusted network showing the amino acids cluster in blue and the lipids clusters in orange/yellow.

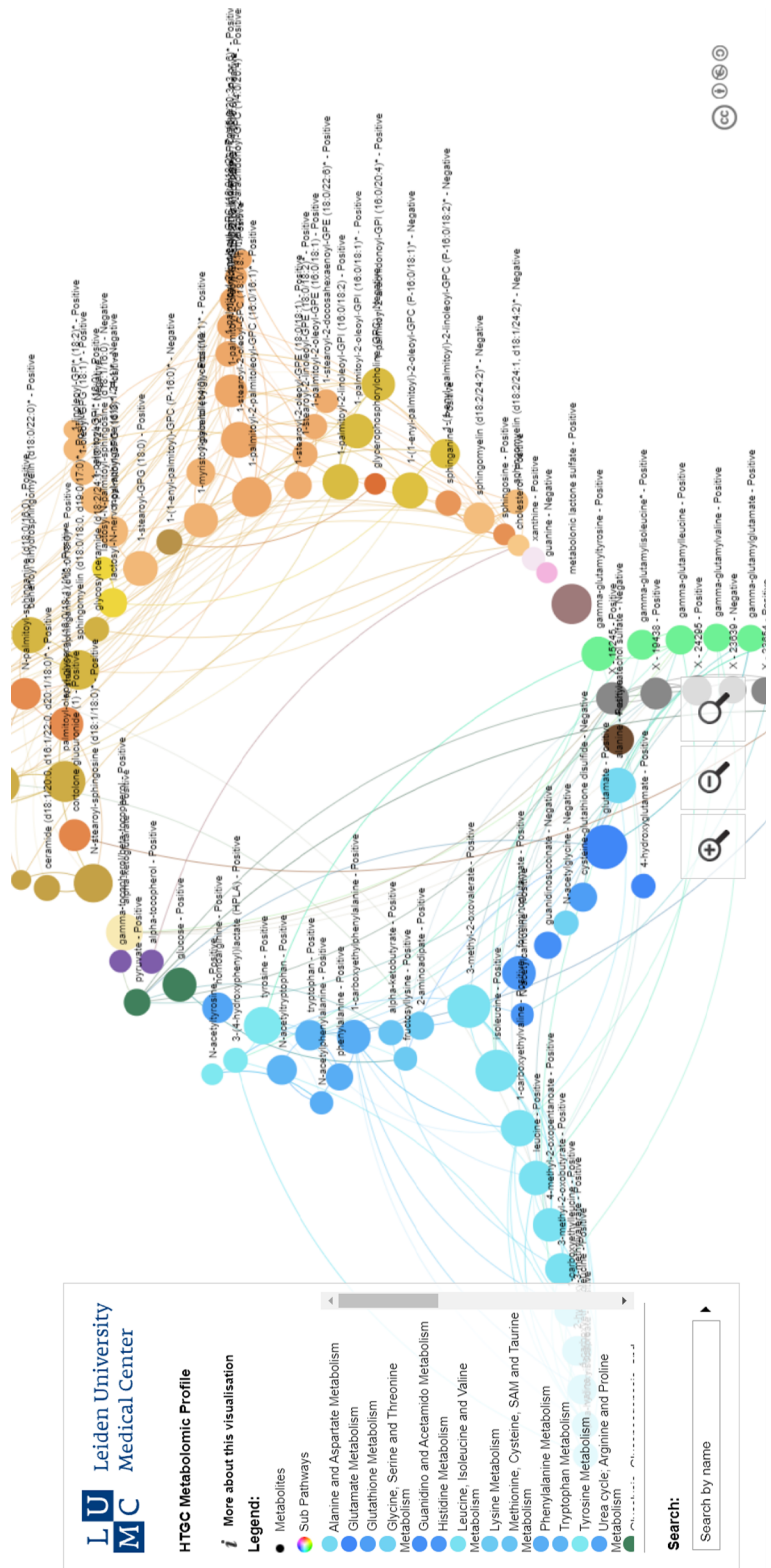
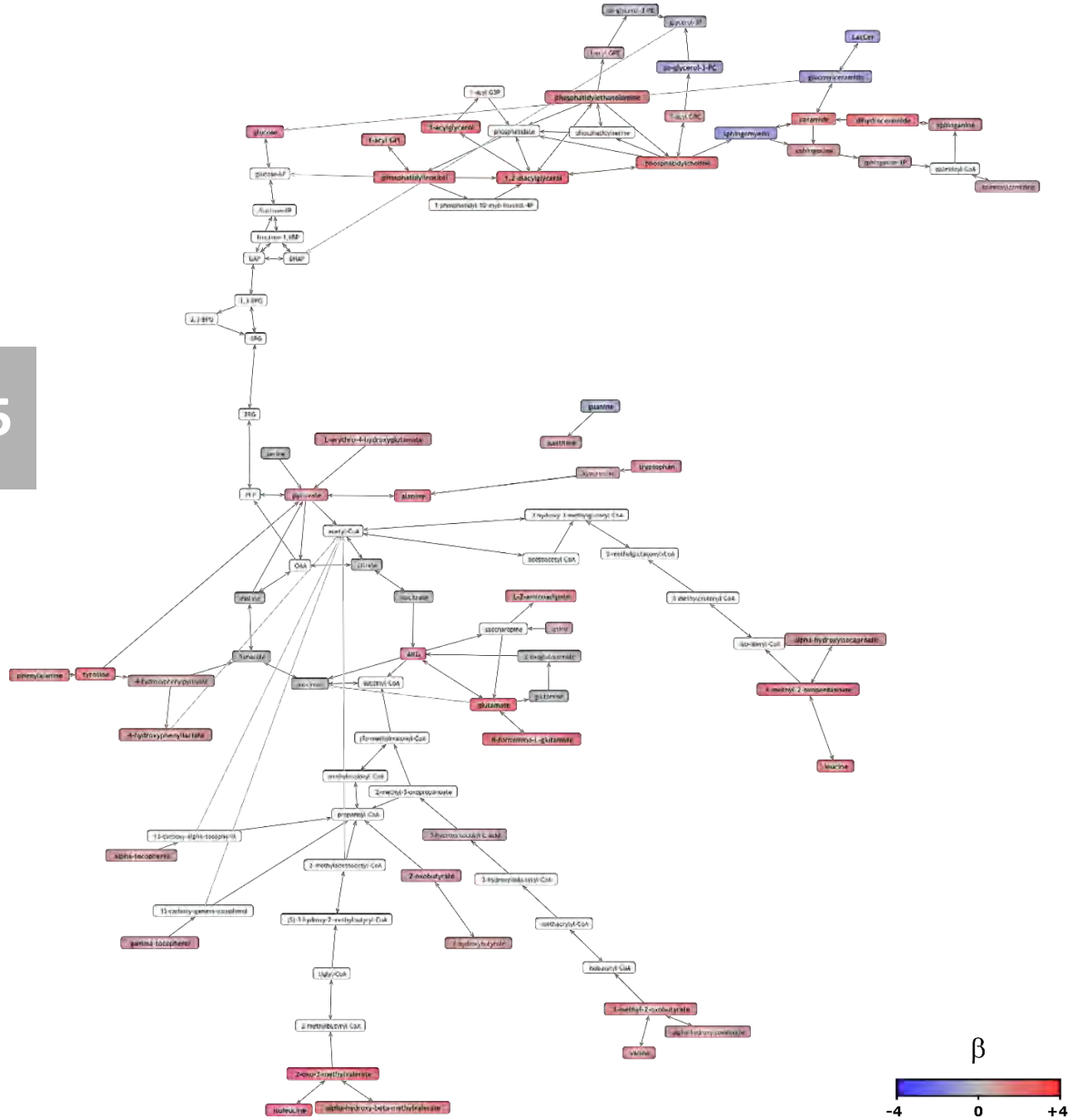


Figure 4: Genome scale metabolic model-based network showing how metabolites (nodes) that are significantly associated to HTGC are related to each other via biochemical reaction paths (edges). Metabolites are color coded according to their *b* value, where red indicates a positive *b* and blue a negative *b*. The network also contains metabolites that are not associated with HGTC but that lie in the conversion path between significant metabolites. Intermediate metabolites that have not been measured are colored white. To facilitate the interpretation of the network visualization, all intermediates of the traditional pathways of glycolysis, BCAA degradation and the Krebs cycle have been added, irrespective of whether they have been measured or not.



8 SUPPLEMENTARY MATERIALS:

8.1 Study Design

The Netherlands Epidemiology of Obesity (NEO) study is a prospective, population-based cohort study aimed at investigating the pathways leading to obesity-related conditions (1)ÿ(ÿÿ. During the period between 2008 and 2012 a total of 6671 participants aged 45–65 years, with an oversampling of overweight persons, from the Leiden greater area were included. Persons aged between 45 and 65 years with a self-reported body mass index (BMI) of ≥ 27 kg/m^2 were invited to participate through letters from their general practitioner and municipalities, as well as through local advertisements. Additionally, all inhabitants aged 45–65 years from one municipality, Leiderdorp, were invited irrespective of their BMI.

Approximately 35% (n = 2580) of the NEO participants without potential magnetic resonance imaging contra-indications were randomly selected to undergo direct assessment of the hepatic triglyceride content (HTGC) by proton magnetic resonance spectroscopy (^{1}H-MRS), which was described in detail in our previous work(1, 2).

Participants were invited for a baseline visit at the NEO study center of the Leiden University Medical Centre (LUMC) after an overnight fast. Prior to this study visit, participants completed a general questionnaire at home to report demographic, lifestyle, and clinical information. Data regarding alcohol intake (g/day) and menopausal status for women was recorded from the participants response in the questionnaire. Participants were asked to bring all medication they were using in the month preceding the baseline study visit and research nurses recorded names (e.g., lipid-lowering medication) and dosages of all medication. Participants came to the research site in the morning and completed a screening form, asking about anything that might create a health risk or interfere with MRI (most notably metallic devices, claustrophobia, or a body circumference of more than 1.70 m). All participants underwent an extensive physical examination, including anthropometry and blood sampling. The percent body fat of the participants was estimated by the Tanita bio impedance balance (TBF-310, Tanita International Division, UK) without shoes and one kilogram was subtracted to correct for the weight of clothing.

8.2 Untargeted Metabolomics Measurements

Fasting state serum samples of the Leiderdorp general subpopulation which had a normal BMI distribution (N=599 in total) from the NEO study were sent for untargeted metabolomics measurements at Metabolon Inc. (Durham, North Carolina, USA) using Metabolon™ Discovery HD4 platform. In brief, this process involves four independent ultra-high-performance liquid chromatography mass spectrometry (UHPLC-MS/MS) platforms (3, 4). Two platforms use positive ionization reverse phase chromatography, one uses negative ionization reverse phase chromatography, and one uses hydrophilic interaction liquid chromatography (HILIC) negative ionization (4). Known metabolites were annotated at Metabolon Inc. with their chemical names, super pathways, sub pathways, compound identifiers from various metabolite databases, and information regarding their biochemical properties.

8.3 Methods

8.3.1 Imputation and Scaling of Metabolite Relative Concentrations

In brief, for the endogenous and unannotated metabolites we applied the multiple imputation method and used the outcome variable, HTGC, along with a select number of correlated metabolites as auxiliary variables to generate 5 imputed datasets. Xenobiotic metabolites were imputed to zero to account for true missingness. All metabolites were natural log-transformed and scaled on their means during the imputation process (5). Following the imputation, metabolite levels were range scaled (also known as min-max scaling) by using the differences of the biological range between the maximum and minimum values for each metabolite and setting the range of all measured metabolites between -1 and 1. This method of scaling was shown to perform well for metabolomic data, particularly untargeted metabolomics data with large units of relative measurement and wide variance between the metabolites. In turn, range scaling provides equal importance to the metabolites in a more biologically relevant manner (6).

For each imputed dataset we performed the regression to assess the association between HTGC and each measured metabolite independently. Each metabolite was used as the exposure variable with the HTGC as the outcome in univariate linear regression analyses. Furthermore, we adjusted for sex, age, total body fat, alcohol intake, and lipid lowering medication. The results of the regression analyses in the imputed datasets were subsequently combined using Rubin's rules (7). All estimates represent the effect on the natural log of HTGC levels per one scaled unit of the metabolite concentration.

To adjust for multiple testing, we applied Bonferroni correction and adopted the method VeffLi estimate described by Ji and Li (8)to account for the number of independent components underlying all 1,363 metabolite measures. This method considers the strong correlations between metabolites from mutual pathways and the inclusion of multiple isomers of the same metabolites to calculate the effective number of independent variables. Accordingly, 758 independent variables were identified, and the calculated Bonferroni correction threshold was set to 0.05/758= 6.59×10^{-5}. The analysis and forest plots were all conducted using R version 3.6.1 (9) and Python 3.7.6 (10). The circular plot was made using the EpiViz R package (11, 12).

8.3.2 Gaussian Graphical Model (GGM)

Significant metabolites from the main analysis and the sex stratified analysis were used to develop the GGM networks separately. To create the network, we used the full measurements of the metabolites from our imputed dataset. Since we used multiple imputation, we could only select one imputed dataset. The correlations between the metabolites were calculated using sparse Gaussian Graphical Models with graphical LASSO using the glasso function in R (qgraph R package version 1.6.9) (13) coupled with the Extended Bayesian Information Criterium (EBIC) to tune the LASSO shrinkage factor (14). EBIC has been reported in simulation studies to perform better than the standard BIC, particularly for controlling false positivity rates. Here, we set the EBIC tuning parameter to the default value of 0.5 (wherein a value of 0 uses standard BIC) for a balanced tuning (15). Based on the correlation coefficients, edges were drawn between the correlated metabolite nodes with the width corresponding to the coefficient strength. The size of the nodes was determined by the absolute value of the effect estimate from the multivariable linear regression analysis. The metabolite nodes were colored based on their biochemical super and sub pathway classes and organized in ascending order based on their p-values. Visualization and layout of the networks were created in Gephi and exported as an interactive HTML5 using the sigmaExporter plugin (16). All GMM networks can be accessed online on https://tofaquih.github.io/AtlasLiver/.

8.3.3 Genome Scale Metabolic Model-based (GSMM) pathway analyses

We mapped the metabolite names onto a curated version of the Human Metabolic Reactions database (HMR 2) genome-scale metabolic model (GSMM) (17) and calculated which metabolite pairs were ≤ 2 reaction steps apart in the model. The GSMM-HMR2 was used to determine the biochemical interconversions between the metabolites that were measured on the Metabolon platform. All computations described in this section were performed in MATLAB R2019b. To facilitate the mapping of metabolomics data to the GSMM, we enriched the model with compound synonyms and external identifiers from the Chemical Entities of Biological Interest (ChEBI) database (18), where ChEBI identifiers and synonyms of conjugate acids and bases were also included. Of the 1067 metabolites with known identity that were quantified, 400 mapped properly to the GSMM. All reactions and compounds in HMR2 were checked for mass and redox balance and were adjusted when necessary.

Biochemical interactions between metabolites were determined by converting the GSMM into a weighted directed graph where nodes represent metabolites and edges represent reactions. Subsequently all reaction paths between the measured metabolites that involved one or two reaction steps were determined using a generic path finding algorithm that was developed *in house*. To ensure that the reaction paths represented relevant biochemical conversions, each path was checked for stoichiometric and thermodynamic consistency. In addition, only substrate-product mappings were considered that involved the transfer of carbon-based moieties, except for CO2. Therefore, half-reactions involving the transfer of electrons, amino or phosphate groups were decoupled from the main reaction in the path finding procedure. For example, in the reaction NADH + pyruvate <=> NAD^+ + lactate, only NADH and NAD^+ are linked in the graph and pyruvate and lactate are linked. Likewise, in the reaction glutamate + pyruvate <=> AKG + alanine, only glutamate and AKG are linked, and pyruvate and alanine are linked. In this way we prevented the creation of crowded, highly connected networks in which most metabolites are connected to a few hub metabolites such as H+, H2O, ATP and NADH.

The weight of the edges in the GSMM-based network was set to 1 for all reactions, except for transporter reactions that transferred compounds over the cellular membranes and (half) reactions that involved uniquely produced metabolites; both these types of reactions were assigned a weight of zero. A consequence of the second exception is that linear reaction chains in which the intermediates were not produced by other reactions were counted as a single reaction step during the path search. Finally, we added a list of intermediates from the glycolysis, Krebs cycle and BCAA degradation pathways that could not be quantified on the metabolomics platform, to prevent gaps in the traditional pathways.

This resulted in a network of connected metabolites that we integrated with biochemical reaction and pathway knowledge into an interactive HTML/JavaScript document, which can be accessed on https://tofaquih.github.io/AtlasLiver/. The results from the network analysis include measured associated metabolites as well as relevant intermediate metabolites regardless of whether they were measured in the dataset. In addition, the network shows the directionality of the biological reaction paths for the biosynthesis and degradation of the metabolites and includes details regarding the involved genes and intermediate reaction steps. Furthermore, for all metabolites and genes, also those not included in the network, hyperlinks to external databases and gene expression profiles are provided. Reaction information was enriched by importing tissue-specific gene expression from the Human Protein Atlas (HPA)(19) and Genotype-Tissue Expression (GTEx) project(20).

8.4 Supplementary Results

Amino Acids Associated with HTGC Independently after Adjustment for Insulin Resistance

Insulin resistance (IR) and HTGC have a complex bi-directional relationship and share overlapping metabolomic pathways, particularly the amino acid metabolism (21). Both of these disease are associated with an elevation and dysregulation of amino acids (22)ÿ. However, it remains uncertain if IR causes HTGC or vice versa (23). To inspect the strong and complex link between IR, HTGC, and amino acids, we performed a post hoc multivariable linear regression analysis and adjusted for IR in addition to sex, age, total body fat, alcohol intake, and lipid lowering. In this analysis we examined 43 metabolites, including amino acids, carbohydrates, metabolonic lactone sulfate, and vitamin E metabolites. Only 10 metabolites lost their association with HTGC, notably pyruvate, while a small reduction of the effect estimates occurred for the remaining metabolites. Overall, the association of the amino acid metabolites with HTGC in our main analysis remained after adjustment for IR (Supplementary Figure 1).

8.4.1 HTGC Associations with Men, Women, and Post-Menopausal Women

NAFLD has been previously shown to be a sexually dimorphic disease in mice and humans (7). Moreover, men have a higher risk of NAFLD than premenopausal women and a similar risk as postmenopausal women, indicating an increased risk of NAFLD after menopause(24-26). However, it has been shown that postmenopausal women have a higher prevalence of severe forms of fibrosis than men and premenopausal women(26). This increased risk was found to be associated with early menarche and estrogen deficiency(25-27). Sex-stratified analysis was performed to observe the general differences of metabolomic associations with HTGC in men vs women and premenopausal vs postmenopausal women (Supplementary Figure 1 and Supplementary Figure 2). We performed secondary analyses using multivariable linear regression in women stratified by menopause and adjusted for age, total body fat, alcohol intake, and lipid lowering medication (Supplementary Figure 2 and Supplementary Figure 3). We found that the number of associations and effect sizes effects were even higher in post-menopausal women while only few metabolites were even associated with outcomes HTGC in the premenopausal women. These findings are in line with previous findings of increased amino acids, such as glutamate and tyrosine, and increased lipids, such as triglycerides and phosphatidylcholines, in postmenopausal women (25, 28).

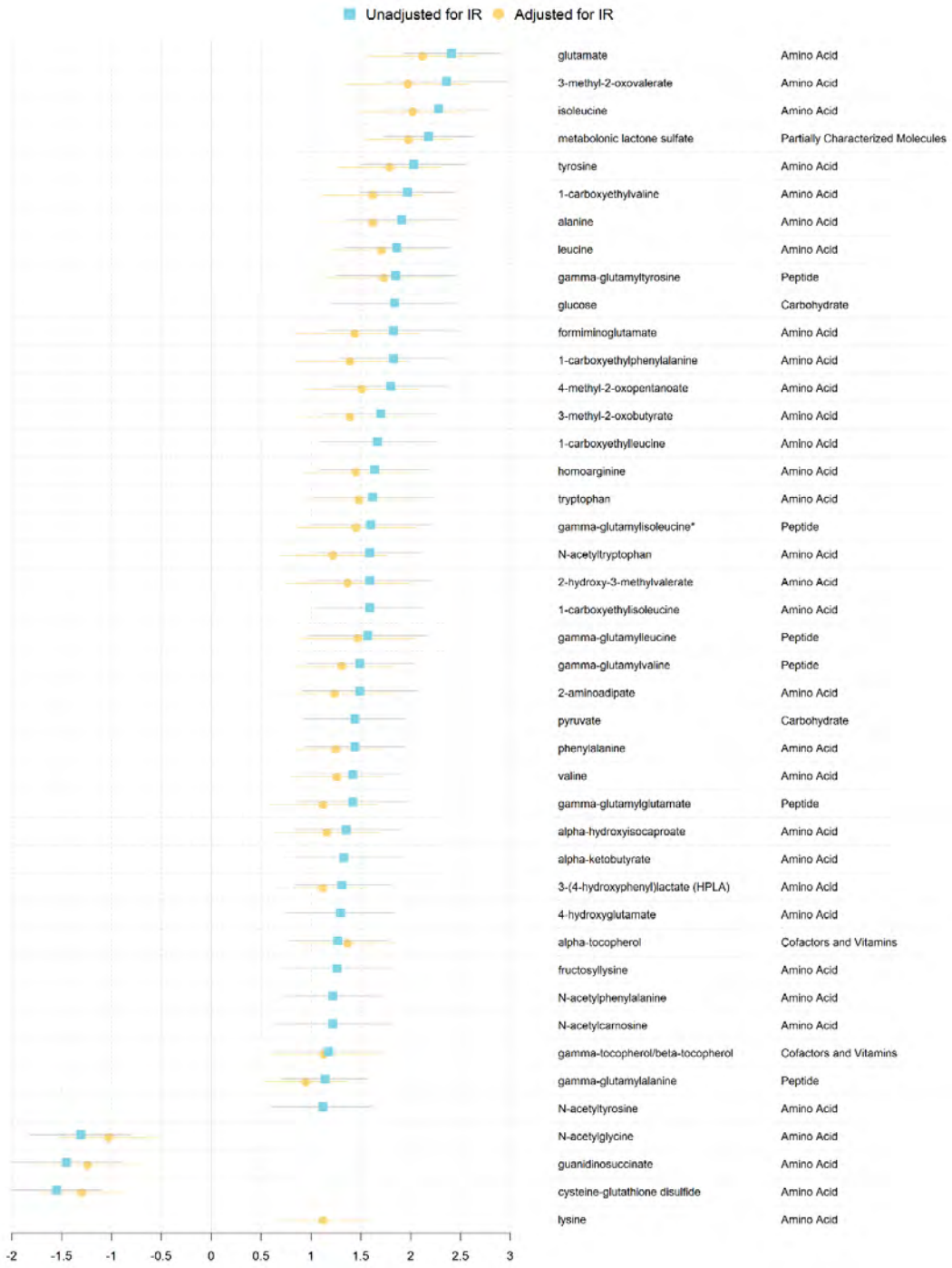
In addition to the general differences between the sex-specific metabolic profiles, several associated metabolites were not shared between men and women (Supplementary Table 1). Two metabolites of interest in men were etiocholanolone, and pregnanediol 3-O-glucuronide. Etiocholanolone glucuronide, a testosterone metabolite(29), and pregnanediol 3-O-glucuronide were negatively associated with HTGC. However, previous studies on this association show inconsistent results(26, 30-32).

Overall, directions of all associations are identical in men, pre- and postmenopausal women. The larger effect sizes in women resulted in a larger number of metabolites associated with HTGC than in men. However, in our analysis the small sample size in the sex-stratified analyses resulted in a loss of power. Thus, we did not perform any specific analysis between men and women and only describe the general observations between the two profiles and provide them as part of the GMM and GSMM networks. Our data indicate that similar pathways are associated with HTGC in men and women before and after menopause. The fact that we do not find obvious sex differences beyond effect size may be due to the homogenous and relatively healthy cohort. Further insight into the metabolite-HTGC differences between men and women, and pre and postmenopausal woman would require analysis in a larger study.

Supplementary Table 1: Metabolites associated only in men or women univariate analysis

Sex	Metabolites	P Value	Estimate (95% CI)	Pathway	Sub Pathway
Men	pregnanediol-3-glucuronide	6.48e-05	-2.98 (-4.42 – -4.42)	Lipid	Progestin Steroids
Men	etiocholanolone glucuronide	3.99e-05	-1.44 (-2.11 – -2.11)	Lipid	Androgenic Steroids
Women	3-aminoisobutyrate	5.98e-05	-1.23 (-1.83 – -1.83)	Nucleotide	Pyrimidine Metabolism, Thymine containing
Women	N-acetylleucine	2.20e-05	1.63 (0.89 – 0.89)	Amino Acid	Leucine, Isoleucine and Valine Metabolism
Women	2-hydroxyarachidate*	6.47e-06	1.78 (1.02 – 1.02)	Lipid	Fatty Acid, Monohydroxy

Supplementary Figure 1: Forest plot comparing the effect estimates of amino acids and carbohydrates on HTGC with and without the adjustment for insulin resistance (IR).



9 REFERENCES

1. De Mutsert R, Den Heijer M, Rabelink TJ, Smit JWA, Romijn JA, Jukema JW, De Roos A, et al. The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection. European Journal of Epidemiology 2013;28:513-523.
2. Boone S, Mook-Kanamori D, Rosendaal F, Den Heijer M, Lamb H, De Roos A, Le Cessie S, et al. Metabolomics: a search for biomarkers of visceral fat and liver fat content. Metabolomics 2019;15.
3. Evans A, Bridgewater B, Liu Q, Mitchell M, Robinson R, Dai H, Stewart S, et al. High resolution mass spectrometry improves data quantity and quality as compared to unit mass resolution mass spectrometry in high-throughput profiling metabolomics. 2014;4:1.
4. Rhee EP, Waikar SS, Rebholz CM, Zheng Z, Perichon R, Clish CB, Evans AM, et al. Variability of Two Metabolomic Platforms in CKD. Clinical Journal of the American Society of Nephrology 2019;14:40.
5. Faquih T, van Smeden M, Luo J, le Cessie S, Kastenmüller G, Krumsiek J, Noordam R, et al. A Workflow for Missing Values Imputation of Untargeted Metabolomics Data. Metabolites 2020;10.
6. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 2006;7:142.
7. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons, Inc., 1987.
8. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity (Edinb) 2005;95:221-227.
9. R Core Team. R: A language and environment for statistical computing. In. Vienna, Austria. URL https://www.R-project.org/. R Foundation for Statistical Computing; 2019.
10. Van Rossum G, Drake F. Python 3 Reference Manual. In. Scotts Valley, CA: CreateSpace; 2009; 2009.
11. Lee M. A MO, Hughes D, Wade K. H, Corbin L. J, McGuinness L. J, Timpson N. J. Epiviz: an implementation of Circos plots for epidemiologists. In; 2020.
12. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize Implements and enhances circular visualization in R. Bioinformatics 2014;30:2811-2812.
13. Friedman J, Hastie T, Tibshirani R. glasso: Graphical lasso-estimation of Gaussian graphical models. R package version 1.7. In; 2011.
14. Foygel R, Drton M. Extended Bayesian Information Criteria for Gaussian Graphical Models. Advances in Neural Information Processing Systems 2010.
15. Chen J, Chen Z. Extended Bayesian information critera for model selection with large model spaces. Biometrika 2008;95:759-771.
16. Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks. Proceedings of the International AAAI Conference on Web and Social Media 2009;3:361-362.
17. Mardinoglu A, Agren R, Kampf C, Asplund A, Uhlen M, Nielsen J. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nature Communications 2014;5:3083.
18. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res 2016;44:D1214-1219.
19. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, et al. Tissue-based map of the human proteome. 2015;347:1260419.
20. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics 2013;45:580-585.
21. Saponaro C, Gaggini M, Gastaldelli A. Nonalcoholic Fatty Liver Disease and Type 2 Diabetes: Common Pathophysiologic Mechanisms. Current Diabetes Reports 2015;15.
22. Hasegawa T, Iino C, Endo T, Mikami K, Kimura M, Sawada N, Nakaji S, et al. Changed Amino Acids in NAFLD and Liver Fibrosis: A Large Cross-Sectional Study without Influence of Insulin Resistance. Nutrients 2020;12:1450.

23. Valenti L, Bugianesi E, Pajvani U, Targher G. Nonalcoholic fatty liver disease: cause or consequence of type 2 diabetes? Liver International 2016;36:1563-1579.
24. Yang JD, Abdelmalek MF, Pang H, Guy CD, Smith AD, Diehl AM, Suzuki A. Gender and menopause impact severity of fibrosis among patients with nonalcoholic steatohepatitis. Hepatology 2014;59:1406-1414.
25. Ballestri S, Nascimbeni F, Baldelli E, Marrazzo A, Romagnoli D, Lonardo A. NAFLD as a Sexual Dimorphic Disease: Role of Gender and Reproductive Status in the Development and Progression of Nonalcoholic Fatty Liver Disease and Inherent Cardiovascular Risk. Advances in therapy 2017;34:1291-1326.
26. DiStefano JK. NAFLD and NASH in Postmenopausal Women: Implications for Diagnosis and Treatment. Endocrinology 2020;161.
27. Klair JS, Yang JD, Abdelmalek MF, Guy CD, Gill RM, Yates K, Unalp-Arida A, et al. A longer duration of estrogen deficiency increases fibrosis risk among postmenopausal women with nonalcoholic fatty liver disease. Hepatology 2016;64:85-91.
28. Auro K, Joensuu A, Fischer K, Kettunen J, Salo P, Mattsson H, Niironen M, et al. A metabolic view on menopause and ageing. Nat Commun 2014;5:4708.
29. Basit A, Amory JK, Prasad B. Effect of Dose and 5α-Reductase Inhibition on the Circulating Testosterone Metabolite Profile of Men Administered Oral Testosterone. Clinical and translational science 2018;11:513-522.
30. Mody A, White D, Kanwal F, Garcia JM. Relevance of low testosterone to non-alcoholic fatty liver disease. Cardiovasc Endocrinol 2015;4:83-89.
31. Yim JY, Kim J, Kim D, Ahmed A. Serum testosterone and non-alcoholic fatty liver disease in men and women in the US. 2018;38:2051-2059.
32. Hardwick RN, Ferreira DW, More VR, Lake AD, Lu Z, Manautou JE, Slitt AL, et al. Altered UDP-glucuronosyltransferase and sulfotransferase expression and function during progressive stages of human nonalcoholic fatty liver disease. Drug Metab Dispos 2013;41:554-561.

Chapter 6

Normal range CAG repeat size variations in the HTT gene are associated with an adverse lipoprotein profile partially mediated by body mass index

Tariq O. Faquih[1], N. Ahmad Aziz[2,3], Sarah L. Gardiner[4], Ruifang Li-Gao[1,5], Renée de Mutsert[1], Yuri Milaneschi[6,7,8,9], Stella Trompet[10], J. Wouter Jukema[11], Frits R. Rosendaal[1], Astrid van Hylckama Vlieg[1], Ko Willems van Dijk[12,13,14], Dennis O. Mook-Kanamori[1, 15]

[1] Department of Clinical Epidemiology, Leiden University Medical Center Leiden, The Netherlands; T.O.Faquih@lumc.nl (T.O.F.); R.Li@lumc.nl (R.L.-G.); R.de_Mutsert@lumc.nl (R.d.M); F.R.Rosendaal@lumc.nl (F.R.R.); A.van_Hylckama_Vlieg@lumc.nl (A.v.H.V.); D.O.Mook@lumc.nl (D.O.M.-K.)
[2.] Population Health Sciences, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany; Ahmad.Aziz@dzne.de (N.A.A.)
[3.] Department of Neurology, Bonn University Hospital, Bonn, Germany; Ahmad.Aziz@dzne.de (N.A.A.)
[4] Department of Neurology, Amsterdam UMC Amsterdam, The Netherlands; esel_gardiner@hotmail.com (S.L.G.)
[5] Metabolon, Inc. Morrisville, North Carolina, United State of America; R.Li@lumc.nl (R.L.-G.).
[6] Department of Psychiatry, Amsterdam UMC location Vrije Universiteit Amsterdam, Amsterdam, The Netherlands; y.milaneschi@amsterdamumc.nl (Y.M.)
[7] Amsterdam Public Health, Mental Health program, Amsterdam, The Netherlands; y.milaneschi@amsterdamumc.nl (Y.M.)
[8] Amsterdam Neuroscience, Mood, Anxiety, Psychosis, Sleep & Stress program, Amsterdam, The Netherlands; y.milaneschi@amsterdamumc.nl (Y.M.)
[9] Amsterdam Neuroscience, Complex Trait Genetics, Amsterdam, The Netherlands; y.milaneschi@amsterdamumc.nl (Y.M.)
[10] Department of Internal Medicine, Leiden University Medical Center, 2300 RC, Leiden, The Netherlands; S.Trompet@lumc.nl (S.T.)
[11] Department of Cardiology, Leiden University Medical Center, 2300 RC, Leiden, The Netherlands; j.w.jukema@lumc.nl (J.W.J.)
[12] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; K.Willems_van_Dijk@lumc.nl (K.W.v.D)
[13] Department of Internal Medicine, Division of Endocrinology, Leiden University Medical Center, Leiden, The Netherlands; K.Willems_van_Dijk@lumc.nl (K.W.v.D)
[14] Einthoven Laboratory for Experimental Vascular Medicine, Leiden University Medical Center, Leiden, The Netherlands; K.Willems_van_Dijk@lumc.nl (K.W.v.D)
[15] Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands; D.O.Mook@lumc.nl (D.O.M.-K.)

Manuscript submitted, under revision

1 ABSTRACT

Tandem CAG repeat sizes of 36 or more in the huntingtin gene (*HTT)* cause Huntington disease. Apart from neuropsychiatric complications, the disease is also accompanied by metabolic dysregulation and weight loss, which contribute to a progressive functional decline. Recent studies also reported an association between repeats below the pathogenic threshold (<36) for Huntington's disease and body mass index (BMI), suggesting that *HTT* repeat sizes in the non-pathogenic range are associated with metabolic dysregulation.

In this study we hypothesized that *HTT* repeat sizes <36 are associated with metabolite levels, possibly mediated through reduced BMI. We pooled data from three European cohorts (n=10,228) with genotyped *HTT* CAG repeat size and metabolomic measurements. All 145 metabolites were measured on the same targeted platform in all studies. Multilevel mixed-effects analysis using the CAG repeat size in *HTT* identified 67 repeat size-metabolite associations. Overall, the metabolomic profile associated with larger CAG repeat sizes in *HTT* were unfavorable—similar to those of higher risk of coronary artery disease and type 2 diabetes—and included elevated levels of amino acids, fatty acids, LDL, VLDL and IDL related metabolites whilst with decreased levels of very large HDL related metabolites. Furthermore, the associations of 50 metabolites, in particular specific very large HDL related metabolites, were mediated by lower BMI. However, no mediation effect was found for 17 metabolites related to LDL and IDL.

In conclusion, our findings indicate that large non-pathogenic CAG repeat sizes in *HTT* are associated with an unfavorable metabolomic profile despite their association with a lower BMI.

2 INTRODUCTION

Huntington disease (HD) is an autosomal dominant neurodegenerative disorder caused by the expansion of a cytosine-adenine-guanine (CAG) repeat in the first exon of the huntingtin gene (*HTT)*. The age of onset of the disease is determined by the number of CAG repeats in this exon: Full penetrance occurs when the number of repeats exceeds 36-39 units (1-4), while fewer than 36 repeats are considered non-pathogenic. However, repeat sizes ranging between 27 and 35 units are categorized as intermediate and have been associated with increased germline instability (4). Symptoms of HD include progressive motor, behavioral and cognitive deterioration, resulting in increasing functional decline and death within 15-20 years after disease onset (1). Intriguingly, HD is also characterized by a range of bio-energetic defects, including insulin resistance, increased sedentary energy expenditure and weight loss, despite increased appetite and caloric intake (5, 6).

The prevalence of HD is higher in populations of Caucasian descent than in Asian and African populations (3, 7). Recent estimates of the prevalence in Europeans vary from 9.7 to 17.3 per 100,000 (3, 4). In a study among five large European population-based cohorts (n~14,000), about 6.5% of the participants were found to have an intermediate or pathogenic number of CAG repeats within the *HTT* gene (8, 9). The pathophysiology of HD is complex and remains to be fully elucidated. Current findings suggest that somatic instability of tandem repeats, as well as disruption of transcriptional regulation, immune and mitochondrial function, protein trafficking, and post-synaptic signaling are likely to be involved (10, 11). Importantly, the rate of weight loss in HD was found to increase with larger CAG repeat sizes (10). Analysis of plasma, serum, and post-mortem brain samples of HD patients have found altered metabolite levels (12-14), reduced concentrations of branched-chain amino acids (15), phosphatidylcholines (15, 16), and reduced whole body cholesterol levels (11). Interestingly, CAG repeat size within the normal and intermediate range, which are considered non-pathogenic, have been associated with depression (17) and cognitive function (18). Metabolic dysregulation in HD patients implies that the CAG repeats in the *HTT* gene may directly affect systemic metabolism. However, the metabolomic signature of the highly polymorphic CAG repeat number variations in the *HTT* gene remains unexplored.

Here, we aimed to profile the metabolomic associations of *HTT* CAG repeat size variations in the non-pathogenic range, by utilizing a targeted nuclear magnetic resonance (1H-NMR) metabolomics platform. This platform included measurement of 145 metabolites, such as amino acids and lipoprotein measurements. To this end, we pooled 1H-NMR and genotype data from three large European cohorts (n=10,275). Given the aforementioned negative association between *HTT* CAG repeat size and BMI, we also aimed to assess to what extent the association between *HTT* CAG repeat size and metabolite levels is mediated through changes in BMI. We hypothesized that longer CAG repeat sizes in the *HTT* gene are associated with an unhealthy metabolomic profile, despite lowering BMI.

3 RESULTS

3.1 Population characteristics

We pooled the individual level datasets from the Netherlands Epidemiology of Obesity (NEO) (19), the Prospective Study of Pravastatin in the Elderly at Risk (PROSPER)(20), and the Netherlands Study of Depression and Anxiety (NESDA) (21) studies (N= 10,228). Characteristics of these studies are summarized in Table 1. The mean age was higher in the PROSPER study (76 years) than NEO (56 years) and NESDA (42 years). PROSPER was the only study to include participants outside the Netherlands, namely Scotland (n=1,808) and Ireland (n=1,448)). Sex distribution was skewed in the NESDA study (65.9% women, as expected due to oversampling of depressed subjects (22)), but nearly equal in NEO and PROSPER studies. Overall, the sex distribution was nearly even in the pooled dataset (54% women). Median CAG repeat sizes in both *HTT* alleles were equal in all studies (Figure 1; Figure 2).

3.2 Associations between *HTT* CAG repeat size variations and metabolite levels

Results from the multilevel mixed-effects linear regression analysis using the metabolite concentrations as the outcomes and *HTT* CAG repeat size, specifically of the longer allele, as exposure variable are presented in Figure 3 and Supplementary Table 1. *HTT* CAG repeat size in the long allele in the combined cohort were statistically significantly associated with the levels of 67/145 metabolites. These included concentrations of different branched and aromatic amino acids, fatty acids, ketone bodies, cholesterols, glycerides, phospholipids, as well as measurements related to different lipoprotein subfractions.

Overall, larger CAG repeat sizes in the long *HTT* allele were associated with increased concentrations of 59/67 metabolites. Conversely, the levels of 8/67 metabolites decreased with larger CAG repeat sizes in the long *HTT* allele.

Amino acids, fatty acids, and ketone bodies

Among the amino acids and branched amino acids, larger CAG repeat sizes in the long allele were associated with higher concentrations of alanine, glutamine, tyrosine, and valine levels. Those larger alleles were also associated with higher concentrations of total fatty acids (monosaturated and unsaturated), omega-3 fatty acids, and docosahexaenoic acid. In contrast, they were associated with lower concentrations of acetate and beta hydroxybutyrate.

3.2.1 Plasma total lipid levels

Larger CAG repeat sizes in the longer *HTT* allele were associated with increased overall serum total cholesterol concentrations — including esterified, remnant and free cholesterols. In line, larger repeat sizes were associated with increased apolipoprotein B (apoB), the apolipoprotein component found in LDL and VLDL. Moreover, measurements of phosphatidylcholine, total cholines, phosphoglycerides, and sphingomyelins concentrations increased by the longer CAG size. The larger CAG repeat sizes were not associated with serum total triglyceride levels.

3.2.2 VLDL-sized lipoproteins

Larger CAG repeat sizes in the longer *HTT* allele were also were also associated with increased total lipids of three VLDL subfractions. Specifically, larger repeat sizes were associated with increased levels of cholesterols (total, esters, and free cholesterols), total lipids, and phospholipids in very small VLDL, while levels of cholesterol esters increased with larger CAG repeat size

Figure 1

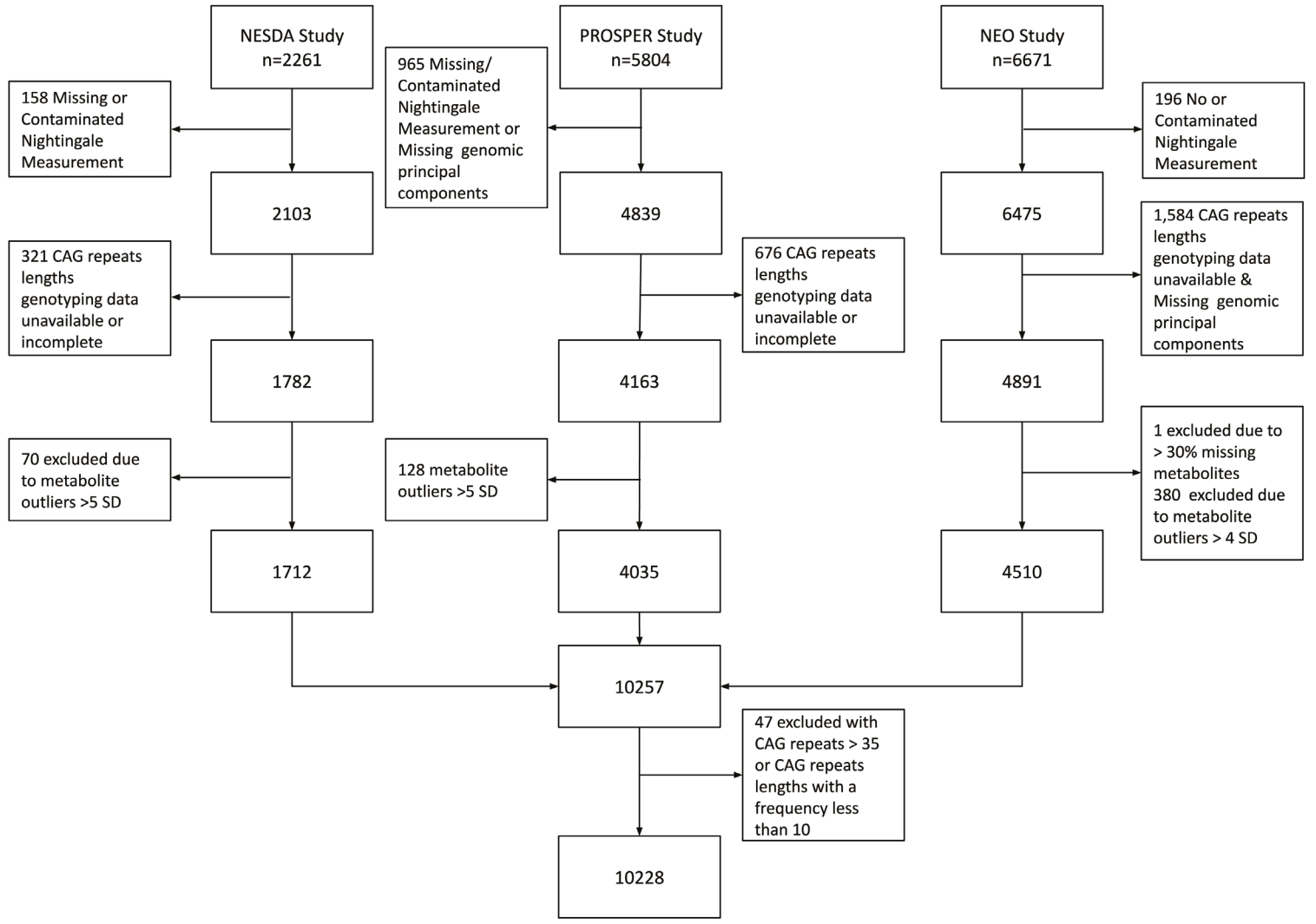
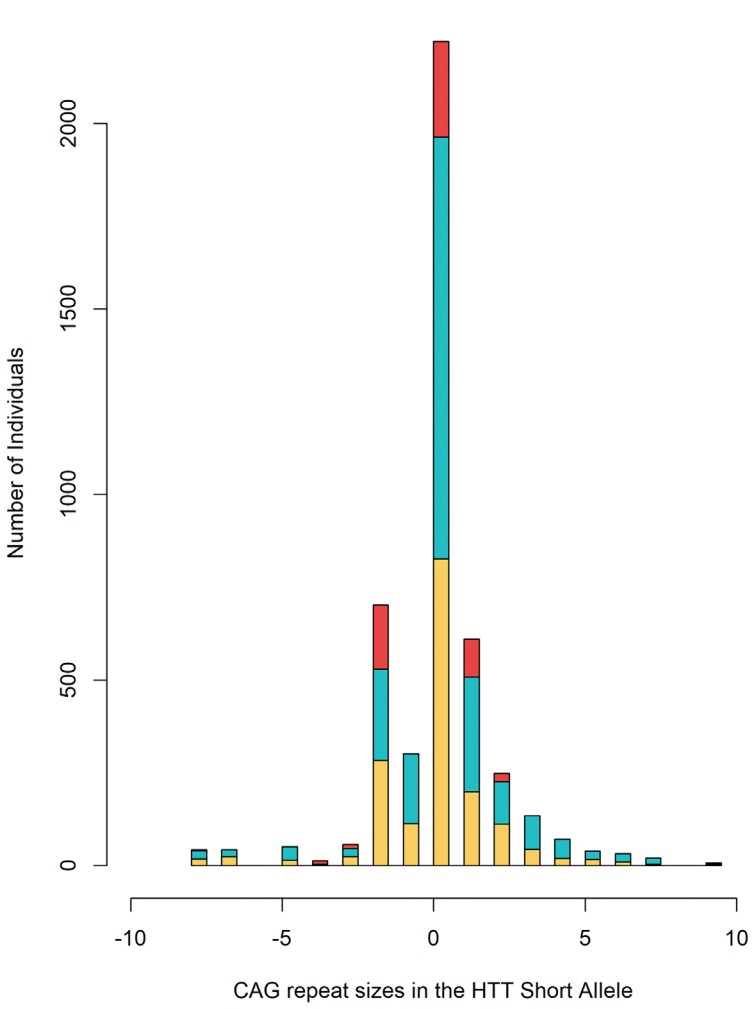


Figure 2



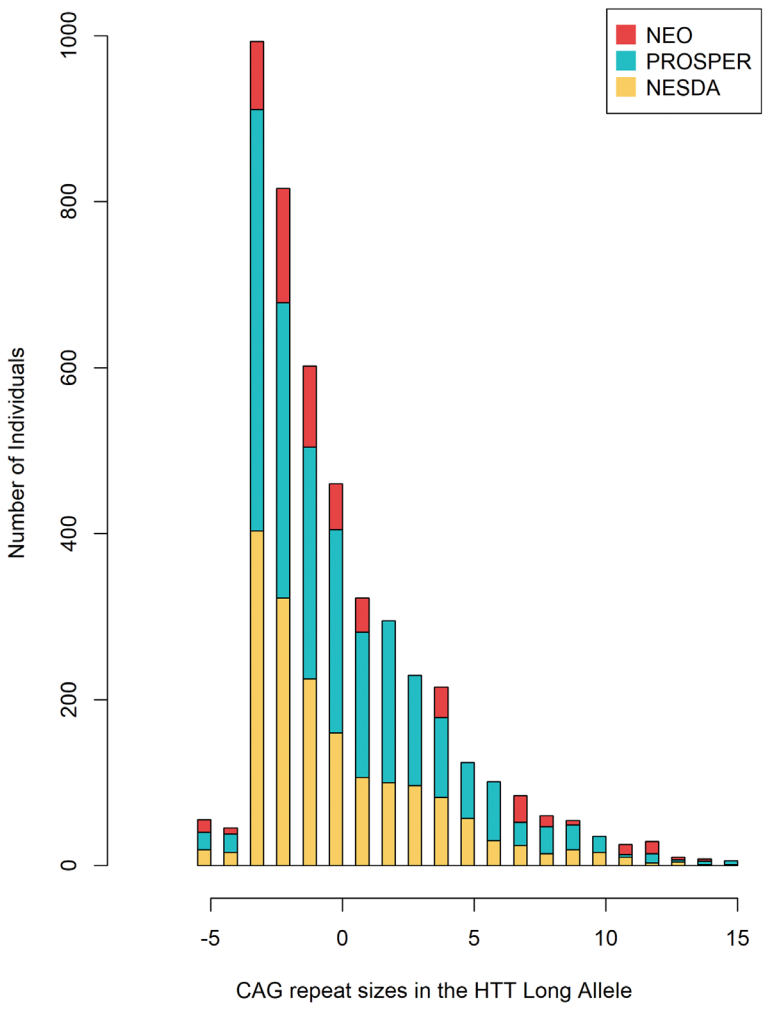
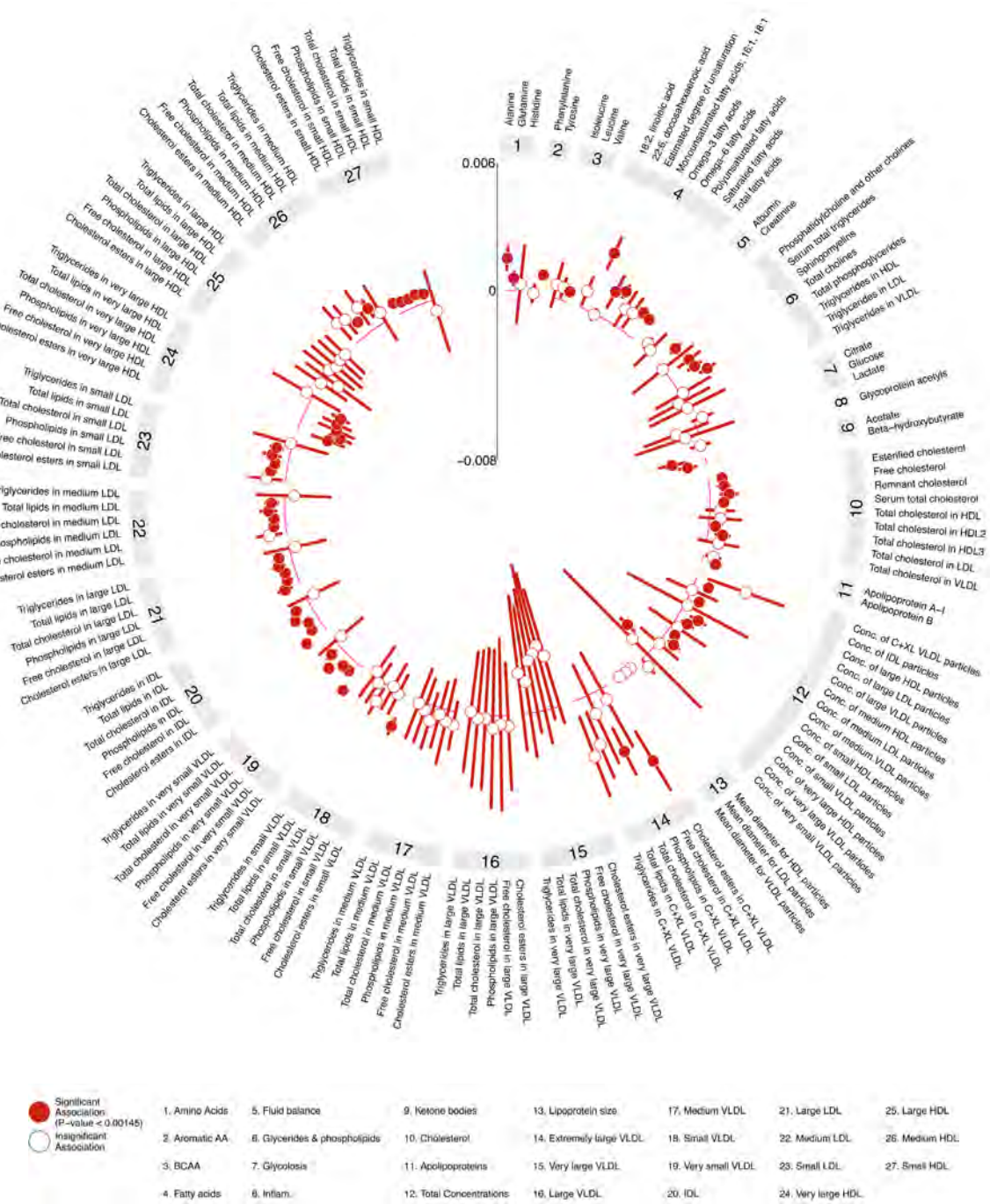


Figure 3



in small VLDL. Finally, the larger CAG repeat size was also associated with increased levels of cholesterol esters and phospholipids in extremely large (XL) VLDL.

IDL-sized lipoproteins

Total concentrations of IDL increased with larger CAG repeat size in the long *HTT* allele. Likewise, the levels of total lipids, cholesterol (total, ester, and free cholesterols), and phospholipids also increased with larger CAG repeat size.

3.2.3 LDL-sized lipoproteins

Larger CAG repeat sizes in the long allele were also associated with higher concentrations of LDL-cholesterol. This association was reflected in increased levels of cholesterols (total and free) in medium and small LDL, and cholesterols (total, ester, and free) in large LDL. Furthermore, larger *HTT* CAG repeat size was associated with increased levels of total lipids and phospholipids in all three subfractions of LDL.

3.2.4 HDL-sized lipoproteins

Larger *HTT* CAG repeat sizes were associated with increased levels of total cholesterol in HDL3, which was reflected by increased levels of small and medium HDL. In small HDL, larger CAG repeat size was associated with increased levels of cholesterol (total, ester, and free), total lipids, and phospholipids. In medium HDL, the levels total lipids and phospholipids were also increased. In contrast, larger *HTT* CAG repeat sizes were related to decreased levels of very large HDL. In addition, larger *HTT* CAG repeat sizes were associated with decreased levels of all metabolites—cholesterol (total, ester, and free), total lipids, and phospholipids—in very large HDL. No associations were present between *HTT* CAG repeat sizes and apolipoprotein A-I levels, a major component of HDL particles.

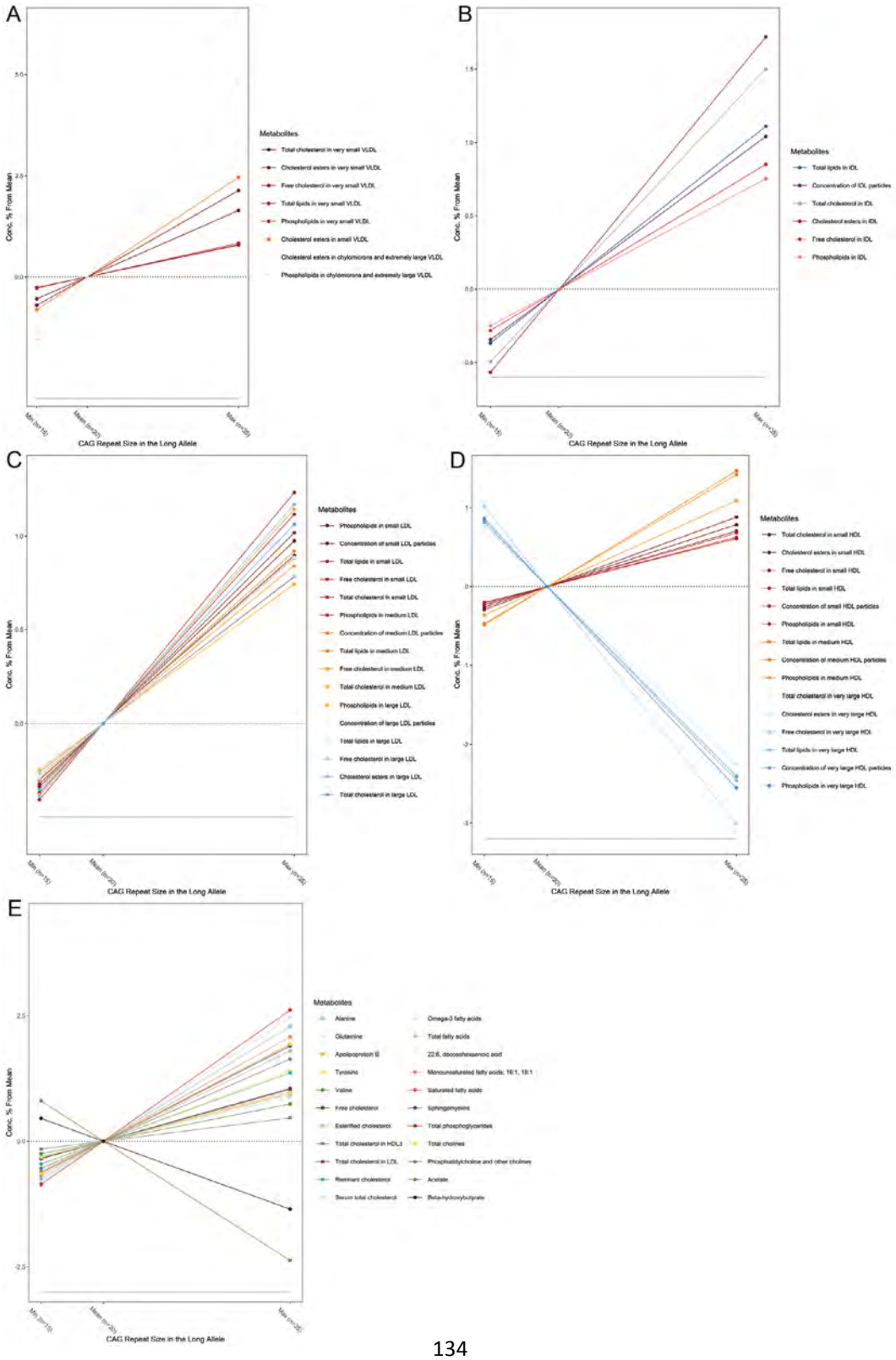
3.3 Estimation of Metabolite levels at the Largest and Smallest CAG Repeat Size

Results for the estimated percentage change for the 67 metabolites previously found in multilevel mixed-effects linear regression analysis are provided in Supplementary Table 4. Overall, at a size of 35 CAG repeats, our model predicts an increase of between 1 to 6% in the levels of all VLDL metabolites (Figure 4A). Levels of phospholipids and cholesterol esters in the XL VLDL were particularly increased to up to 5% and 6% from the mean, respectively. Levels of metabolites related to IDL, LDL, and small and medium HDL increased by 1% from the mean CAG size (Figure 4B-D). Conversely, levels of very large HDL and its lipid and cholesterol all decreased by approximately 2-3% at 35 CAG repeats compared to the mean size (Figure 4D). Amino acids, fatty acids, total cholesterols, and the other remaining metabolites, were increased between 1% and 4.5% at 35 CAG repeat. The exceptions were acetate and beta-hydroxybutyrate, which decreased by 2.4% and 1.3% at 35 CAG repeat size respectively (Figure 4E).

3.4 Nonlinear Associations

Additional sensitivity analyses were performed to assess potential interaction effects between the CAG repeat sizes in the two *HTT* alleles, as well as to assess their potential nonlinear associations with metabolite levels. In this analysis we identified 77 metabolites associated with the CAG repeat sizes, quadratic terms, or the interaction term between the two allele sizes. Of these associations, 14 CAG-metabolite associations were not previously found in the multilevel mixed-effects linear regression analysis (when the interaction and quadratic terms were not included). Ten of these 14 metabolites had an association with the quadratic terms or the inter-

Figure 4



action terms, which included citrate, apolipoprotein A-I, histidine, leucine, unsaturated fatty acid levels, mean diameters of HDL and VLDL, and measurements in large HDL and very large VLDL. Full results for these metabolites are presented in supplementary Table 2.

Among the metabolites associated with CAG repeats in the linear and nonlinear analyses (63/77), 26 had a significant association with the quadratic or interaction terms. Ten metabolites were associated with both alleles and the interaction term, and 7 metabolites had an association with the quadratic terms. These nonlinear and interaction associations were primarily with HDL, glycerides and phospholipids, fatty acids, histidine, and alanine. Overall, the associations between the linear and nonlinear models were minimal and the quadratic and interaction estimates were small.

3.5 Mediation analysis

First, we performed the multilevel mixed-effects regression between the larger *HTT* CAG repeat size as the independent variable and BMI as the outcome. Accordingly, larger CAG repeat size in the long allele was associated with lower BMI (effect estimate of -0.03 kg/m^2 per CAG repeat; 95% CI: -0.05 – -0.01; SE: 0.01).

Second, as the larger CAG repeat size were negatively associated with BMI, we performed an analysis for mediator-outcome, and calculated the mediation (indirect) and the total effect. BMI had a significant mediation effect in 50/67 of the CAG-metabolite associations. Inconsistent mediation effects, in which the direct and mediated effects were in opposing directions, were present in 74% of the associations. Metabolites without mediation effect by BMI were predominately total, free, and esterified cholesterol, total lipids, and phospholipids in LDL subfractions and IDL. The mediation effect accounted for the majority of the total effect on the concentration of very large HDL and its lipid and cholesterol content (Table 2). These measurements were also found to be strongly correlated with each other and with the total cholesterol in HDL as illustrated in the heatmap in Supplementary Figure 1. Despite the positive mediation effect on the metabolite levels, the total effect of CAG repeats on these HDL metabolites remained negative in contrast, both the mediation effect and direct effect were positive and increased the level of total cholesterol in HDL3. In summary, the overall effects of larger CAG repeat size on metabolite levels were slightly reduced after accounting for the mediation effect of BMI. Detailed results for the mediation analysis results are provided in Supplementary Table 3A and 3B.

4 DISCUSSION

We present a study on the association between CAG repeat size in the *HTT* gene—in the non-pathogenic range— and metabolite levels in more than 10,000 individuals of European ancestry. Larger *HTT* CAG repeat size in the longer allele were associated with the levels of 67 out of 145 measured metabolites. We found that the association between larger *HTT* CAG repeat sizes and total concentrations of lipid species in very large HDL remained negative despite significant mediation by lower BMI. Partial mediation by BMI was also found for 50 metabolites, wherein the larger CAG repeat sizes were associated with increased levels of lipids in small HDL and VLDL, as well as elevated levels of amino acids and fatty acids, despite inconsistent mediation by BMI. Conversely, BMI did not mediate the effects of the larger *HTT* CAG repeat sizes on the levels of 17 other metabolites, primarily consisting of cholesterol and lipids in IDL and LDL.

Overall, our findings indicate a role for tandem repeat polymorphisms in the *HTT* gene in the regulation of a diverse array of metabolites. We found that the larger size of CAG repeats of

6

the long allele was related to increased levels of small and medium HDL, and their cholesterol content, as well as increasing omega-3 fatty acids, and total cholesterol in HDL3. In this respect, the larger CAG repeat size is related to a more favorable lipoprotein profile, such as, for example, observed during weight loss (23). However, also an unfavorable metabolomic profile with increasing long allele CAG repeat size was observed: LDL, IDL, VLDL particles, apoB, remnant cholesterols, total cholesterols, valine, tyrosine, alanine, and total fatty acids were all positively associated with the larger CAG repeat size in the long allele. The associations with HDL particles of different subfractions were heterogeneous. Opposite effect directions were found for very large HDL cholesterols, lipids, and concentration in comparison to small and medium HDL. Our mediation analyses indicated an inconsistent mediation by BMI with respect to very large HDL cholesterols and lipid levels, which were highly correlated with the total levels of HDL cholesterol. On the other hand, the associations with several LDL and IDL cholesterols levels were not mediated by BMI. Moreover, the mediation effect through BMI was generally inconsistent with the direct effect and was partial or low for 50 metabolites. These findings thus suggest the existence of an alternative pathway, independent of BMI, through which *HTT* CAG repeat size variations could affect the levels of these metabolites, including LDL and IDL.

We found that larger *HTT* CAG repeat sizes were associated with a metabolic profile similar to what was recently described in people at high risk for coronary artery disease (CAD) (24). In particular an inverse association between the cholesterols in larger HDL particles—and not the small or medium HDL—with the incidence of CAD has been reported (24). Furthermore, elevated total concentrations and cholesterols levels in LDL, IDL, VLDL, triglycerides, and apoB were accompanied by a higher incidence of CAD and peripheral artery disease (PAD). ApoB in particular has recently been reported as a strong lipoprotein marker for cardiovascular risk (25). Our findings for the associations with amino acid and branched amino acids, and fatty acids – specifically for alanine, valine, tyrosine, and total fatty acids – are also indicative of a metabolic profile associated with a higher risk of CAD, type 2 diabetes (26, 27), unhealthy adiposity (23), metabolically unhealthy normal weight (28), and inactivity (29). These heterogenous metabolomic profiles are also comparable to what has been observed in HD patients (5, 10, 30, 31), in whom weight loss and increased resting estate energy expenditure are accompanied by a higher risk of CAD and type 2 diabetes.

We additionally estimated that relatively large CAG sizes can substantially decrease very large HDL-related metabolites (by up to 3%), while increasing the levels of other lipoprotein metabolites by 1-6% as compared to the mean CAG repeats size (Figure 4). Thus, our findings indicate that larger *HTT* CAG repeat sizes result in a metabolic profile reminiscent of that associated with high CAD risk, suggesting a possible role of CAG repeats in *HTT*, and potentially other genes with polymorphic CAG repeat tracts, as genetic modifiers of clinically relevant cardiometabolic traits and disorders. Indeed, *HTT* CAG tandem repeat polymorphisms may account for part of the (missing) heritability of different metabolites, and by extension, of other phenotypes, such as BMI (32) and CAD. Thus, CAG repeat size polymorphisms are promising targets for further exploration in future studies.

4.1 Strengths and Limitations

Our study has several strengths. First, the genotyping methodology used in the three cohorts was specifically designed to genotype the tandem repeat region in *HTT*. Second, we pooled the targeted metabolomics and genotyping data from 3 European cohorts for analysis, resulting in a uniquely large sample size. Third, few metabolomic studies have been conducted in HD patients. Moreover, these studies used case-control study designs with small sample sizes (33). Our study

is the largest metabolomics study thus far on the metabolomic signature associated with CAG repeat size variations in the *HTT* gene. Fourth, we found largely positive CAG-metabolite associations despite the lowering of BMI in the mediation analysis. Our study also has some potential limitations. Our study populations were at an increased risk of cardiovascular diseases and depression, which may have induced collider bias. Although, for the NEO study, we accounted for oversampling of overweight individuals, this was not possible for other characteristics, such as depression, in all studies. This was due to the unknown proportion of oversampling. However, the effect estimates were similar across studies, making it unlikely that this oversampling affected the results of our analysis. *HTT* CAG repeat size variations were also associated with the odds of depression in a previous study (17). Therefore, examining the potential role of depression may provide further insights into the mechanisms underlying the CAG-metabolite associations. However, this was beyond the scope of the current study. Finally, we could not deduce causal associations from the mediation analysis due to the difficulty verifying that no mediator-outcome confounding was present.

5 CONCLUSION

In conclusion, we examined the relationship between CAG repeat size in *HTT* with the levels of a large number of circulating metabolites. We found that non-pathogenic CAG repeat size variations in *HTT* are associated with the levels of 67 metabolites, exhibiting a heterogenous metabolomic signature. Favorable associations included positive associations with levels of cholesterol in small and medium HDL. Despite the observation that larger *HTT* CAG repeat sizes were associated with lower BMI and a favorable profile for some metabolites, we observed an additional unfavorable metabolomic profile, including associations with elevated LDL and IDL cholesterols, reduced cholesterol in very large HDL and elevated amino and fatty acids. This unfavorable profile was found to overlap with the profile seen in unhealthy adiposity, CAD, and type 2 diabetes. Based on mediation analysis, 50 metabolites showed only partial mediation and 17 metabolites—related to LDL and IDL cholesterol levels— showed no significant BMI mediation at all. Our mediation results therefore imply the potential existence of a BMI-independent mechanism underlying their association with CAG repeat size. We also found intriguing novel associations of CAG repeat size in *HTT* with metabolic dysregulation, with and without the mediation of BMI. Thus, tandem repeat polymorphisms in *HTT* and other genes may contribute to the heritability of cardiometabolic diseases and be instrumental in the elucidation of their underlying metabolomic mechanisms.

6 METHODS

6.1 Study Design

Data derived from three European cohorts were merged for pooled analyses, i.e., NEO, PROSPER, and NESDA. Details regarding the inclusion criteria of each cohort are summarized in Figure 1.

6.1.1 NEO

The NEO study is an ongoing population-based, prospective cohort study of individuals aged 45–65 years, with an oversampling of individuals with overweight or obesity. Men and women aged between 45 and 65 years with a self-reported BMI of 27 kg/m^2 or higher, living in the greater area of Leiden (in the West of the Netherlands) were eligible to participate in the NEO study. In addition, all inhabitants aged between 45 and 65 years from one municipality (Leiderdorp)

6

were invited irrespective of their BMI, allowing for a reference distribution of BMI. Recruitment of participants started in September 2008 and completed at the end of September 2012. In total, 6,671 participants have been included, of whom 5,217 with a BMI of 27 kg/m^2 or higher. Participants were invited to come to the NEO study center of the Leiden University Medical Center for a baseline study visit after an overnight fast of at least 10 hours. During the visit fasting blood samples were taken from the participants (19). The study was approved by the medical ethical committee of the Leiden University Medical Center. The sample size for this analysis was 4,510 participants of European ancestry after exclusion of participants without metabolomic data (n=99), flawed metabolomic measurements (due to high peroxide or high ethanol content) (n=97). Moreover, as the NEO study had a higher number of extreme values than the other included studies, individuals were excluded if they had metabolite measurements above 4 standard deviations (n=380), instead of the 5 standard deviations cutoff used in the PROSPER and NESDA studies. Finally, we excluded individuals without genotype data (n=1,584). In addition, one individual was excluded due to abnormally high number of missing metabolite measurements (49/145; 33%) (Figure 1).

6.1.2 PROSPER

PROSPER was a randomized, double-blind, placebo-controlled trial among 5,786 men and women between 70-82 years old with a pre-existing vascular disease or a raised risk for such a disease. Participants were recruited from three countries with 2,517 individuals from Scotland, 2,173 individuals from Ireland and 1,096 individuals from the Netherlands. Fasting blood sample were collected and stored at -80 °C for later NMR metabolomics analysis (34). The study was approved by the institutional ethics review boards of all centers and written informed consent was obtained from all participants (20). The final sample size used in this analysis was 4,035 after exclusion of participants with flawed metabolomic data (n=965), individuals with metabolite measurements above 5 standard deviations (n=128), and participants without genotype data (n=676) (Figure 1).

6.1.3 NESDA

NESDA is an ongoing longitudinal cohort study into the long-term course and consequences of depressive and anxiety disorders. The sample consists of 2,981 participants with depressive/ anxiety disorders and healthy controls recruited from the general population, general practices, and secondary mental health centers (21). Blood samples were collected after an overnight fast at the baseline visit (2004-2007). For the present analyses we initially selected data from 2,261 unrelated individuals of European ancestry identified using GWAS data. The Ethical Committees of all participating universities approved the NESDA project, and all participants provided written informed consent (35). We excluded 158 individuals without metabolomic data (n=39) or with flawed samples (n=199), and with metabolite outliers above 5 standard deviations (n=70). In addition, individuals with genotype data were also excluded (n=321). The final sample size used in the analysis was 1712 (Figure 1).

6.2 Genotyping

Due to the technical limitation of next-generation short-read sequencing to accurately call DNA repeat sequences (36), a multiplex polymerase chain reaction (PCR) method was developed using TProfessional thermocycler (Biometra, Westburg) with labelled primers to genotype the CAG repeat sizes in the two *HTT* alleles. Full details about the genotyping methodology have been described previously (17).

6.3 Metabolomics Measurements

Metabolomic profiles were measured using the Nightingale (Nightingale Health Ltd, Helsinki, Finland) NMR platform in all selected participants. Nightingale uses a targeted metabolomics approach by defining the specific metabolites to be quantitively measured in advance. This approach yields consistent and reproducible concentration measurements across studies (37). The platform measures approximately 226 metabolites and metabolite ratios, consisting predominantly of very low density (VLDL), intermediate density (IDL), low density (LDL), and high density (HDL) lipoproteins. Those lipoproteins—with the exception of IDL—are further subclassified based on their lipid composition and particle sizes(38). Accordingly, VLDL is divided into very small, small, medium, large, very large, and extremely large subfractions; HDL is divided into small, medium, large, and very large subfractions; and LDL is divided to small, medium, and large subfractions. The supplementary ratio variables calculated the ratio of various metabolites concentrations within lipoprotein subfractions, e.g., “triglycerides to total lipids ratio in IDL”. Additionally, the platform measured the concentrations of various individual metabolites beyond lipoproteins such as amino acids, free fatty acids, and ketone bodies (39). For our study we excluded the 81 ratio variables and focused on the remaining 145 metabolite concentrations that were available in all 3 cohorts. Samples in all cohorts were taken after a fasting period.

6.4 Statistical Analysis

6.4.1 Multilevel mixed-effects linear regression

We performed a joint polynomial multilevel mixed-effects linear regression using data from all three cohorts. First, for each individual, we defined the *HTT* allele with the larger CAG repeat tract as ‘long’, and the other one as ‘short’. This was done as the two alleles can have independent effects as demonstrated in previous studies (32). The number of repeats in each allele were then mean-centered to reduce multicollinearity and ease the interpretation. In order to address possible heteroscedasticity we used robust standard errors for the analysis. Influential data points (i.e., influential outliers) were accounted for by removing CAG lengths with a frequncy less than 10 in the combined cohort (n=47). Therefore, the final pooled number of particpants used in the analysis was n = 10,228 (Figure 1).

Metabolite variables were natural log-transformed and the missing values were imputed using the K-nearest neighbor imputation method described in our previous work (40, 41). In brief, for each metabolite with missingness, we selected 10 correlated metabolites with no missingness. We then used these metabolites to impute the missing values by calculating the means. We expect that this imputation method will have negligible bias and error as the number of missing values was low and sample sizes were large, as was demonstrated in the simulation results in our past work as well (41).

Since we had access to the individual level data of all three studies, we were able to perform pooled analyses, rather than meta-analyzing the effects per study. For the analysis we adjusted for age, sex, and the first 4 genetic principal components as the fixed factors. In addition, we used country and study variables as random factors in the mixed effects model. As the NEO data had an oversampling of overweight individuals, we weighted the analyses to the BMI distribution of the Dutch general population. The weight was set to 1 for the PROSPER and NESDA participants. To account for population stratification, we used the country (Netherlands, Scotland, and Ireland) and the cohort (NEO, PROSPER, and NESDA) as random effect variables.

Both alleles were included in the regression models as previous studies reported differing associations between the "long" and "short" alleles in *HTT* with different outcomes (17, 18, 32, 42). However, due to the dominant effect of the *HTT* repeat expansion in HD, we focused on the "long" allele effect estimates only. We performed the analysis for each of the 145 metabolites as the outcomes and the mean-centered number of repeats in both *HTT* alleles as the independent variables. Effects of CAG repeats have been shown to have nonlinear associations and interactions have been described between the two *HTT* alleles (17, 18, 43). Therefore, we conducted a secondary analysis to check nonlinearity and used a polynomial model by adding quadratic terms for each allele and an interaction term between the two alleles (long and short). We adjusted for age, sex, and the first 4 genetic principal components as the fixed effects and used country and study as random effects in the mixed effects model.

Data preparation and analysis were conducted with R version 4.1.0 (44). Circular plots for the effect estimates were designed using the EpiViz R package (45-47). Multilevel mixed-effects model and mediation analyses were performed by utilizing the "mixed" command in STATA/SE version 16 (StataCorp LLC) (48).

6.4.2 Multiple testing correction

To adjust for multiple testing, we used the VeffLi estimate described by Ji and Li (22). This method takes the covariance between metabolite levels into account by estimating the effective number of independent variables. Accordingly, the effective number of independent variables was 35 and the adjusted p-value cut-off was put at 0.05/35 = 0.0014.

Estimation of Metabolite levels at the Largest CAG Repeat Size

The effect estimates from the multilevel mixed-effects linear regression accounted for the effect of 1 CAG repeat size increase. By using the effect estimates per CAG repeat from the mixed linear regression model, we were able to show a simple estimation of the percentage difference from the mean of metabolites that were associated with the larger CAG repeat size from the multilevel mixed-effects linear regression analysis. We estimated the percentage change in metabolite levels at CAG repeat sizes equal to the smallest, mean, and largest CAG repeat size in the pooled dataset, corresponding to 15, 20, and 35 repeats respectively. Plots for visualizing the percentage changes were generated using the *looplot* R package (49, 50).

6.4.3 Mediation Analysis

To test for mediation by BMI of the CAG-metabolite associations, we performed three analyses as proposed by Baron & Kenny (1986) (51). First, we modelled the exposure-mediator relationship by using the multilevel mixed-effects linear regression to assess the association between the CAG repeat sizes and BMI. Second, we calculated the mediation effect using the multilevel mixed-effects linear regression for the metabolites that were associated with *HTT* CAG repeat size in the previous analysis. The natural logarithm of the metabolite levels was used as the outcome and the independent variables were the CAG repeat sizes in the short and long alleles, as well as BMI, the mediator. Third, given our large sample size, we used the simpler Sobel's test (Equation 1) instead of bootstrapping to test the mediation effect of BMI (51-53).

$$A \times B / \sqrt{(A^2 \times B_{se}^2) + (B^2 \times A_{se}^2) + (A_{se}^2 \times B_{se}^2)}$$

Equation 1: The Sobel's equation for testing mediation. A: the estimate between CAG repeat sizes in *HTT* and BMI; B: the estimate between BMI and metabolite levels; A_{se}: standard error of A; B_{se}: standard error for B; A × B is the indirect effect of BMI

Using this method, we calculated the indirect effect through BMI by multiplying the estimates of BMI from the exposure-mediator model and mediator-outcome model. We also calculated the total effect for the model by adding the direct effect, i.e. estimates of the CAG repeat sizes, to the mediation effect. Furthermore, for each allele we divided the indirect effect by the total effect to obtain the index of mediation, i.e. the percentage of the effect of CAG repeat size variations on metabolites that is mediated by BMI.

7 FUNDING

This study was supported by a VENI-grant (#91615080) from the Netherlands Organization of Scientific Research. N.A.A. is partly supported by an Alzheimer's Association Research Grant (Award Number: AARG-19-616534) and a European Research Council Starting Grant (Number: 101041677). The NEO study is supported by the participating Departments, Division, and Board of Directors of the Leiden University Medical Center, and by the Leiden University, Research Profile Area Vascular and Regenerative Medicine. DOM-K is supported by Dutch Science Organization (ZonMW-VENI Grant No. 916.14.023). T.O.F. was supported by the King Abdullah Scholarship Program and King Faisal Specialist Hospital & Research Center [No. 1012879283]. Funding for NESDA was obtained from the Netherlands Organization for Scientific Research (Geestkracht program grant 10-000-1002); the Center for Medical Systems Biology (CSMB, NWO Genomics), Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL), VU University's Institutes for Health and Care Research (EMGO+) and Neuroscience Campus Amsterdam, University Medical Center Groningen, Leiden University Medical Center, National Institutes of Health (NIH, R01D0042157-01A, MH081802, Grand Opportunity grants 1RC2 MH089951 and 1RC2 MH089995). Part of the genotyping and analyses were funded by the Genetic Association Information Network (GAIN) of the Foundation for the National Institutes of Health. Computing was supported by BiG Grid, the Dutch e-Science Grid, which is financially supported by NWO. The PROSPER study was supported by an investigator-initiated grant obtained from Bristol-Myers Squibb.

8 ACKNOWLEDGMENTS

The authors of the NEO study thank all participants, all participating general practitioners for inviting eligible participants, all research nurses for data collection, and the NEO study group: Pat van Beelen, Petra Noordijk, and Ingeborg de Jonge for coordination, laboratory, and data management. The authors are also thankful to Merel Boogaard for performing the genotyping assays.

8.1 Conflict of interest

R.L.-G. is a part-time clinical research consultant for Metabolon, Inc. All other co- authors have no conflicts of interest to declare.

9 AUTHOR CONTRIBUTIONS

T.O.F.- conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing-original draft. N.A.A.-, resources, funding acquisition, methodology, writing – review & editing. S.L.G.-, methodology, writing – review & editing. R.L.-G.- validation, writing – review & editing. R.d.M.- study design, conduct and data collection, resources, funding acquisition, writing – review & editing. Y.M.- project administration, resources, writing – review & editing. J.W.J.- resources, funding acquisition, writing – review & editing. F.R.R.- study design, funding acquisition, conceptualization. A.v.H.V. and K.W.v.D- conceptualization, supervision, writing – review & editing. D.O.M.-K.- conceptualization, supervision, funding acquisition, writing – review & editing.

10 REFERENCES

1 McColgan, P. and Tabrizi, S.J. (2018) Huntington's disease: a clinical review. *European Journal of Neurology*, **25**, 24-34.

2 Tabrizi, S.J., Flower, M.D., Ross, C.A. and Wild, E.J. (2020) Huntington disease: new insights into molecular pathogenesis and therapeutic opportunities. *Nature Reviews Neurology*, **16**, 529-546.

3 Rawlins, M.D., Wexler, N.S., Wexler, A.R., Tabrizi, S.J., Douglas, I., Evans, S.J.W. and Smeeth, L. (2016) The Prevalence of Huntington's Disease. *Neuroepidemiology*, **46**, 144-153.

4 Caron, N.S., Wright, G.E.B. and Hayden, M.R., In *GeneReviews® [Internet]: Huntington Disease*. University of Washington, Seattle, WA, USA, in press.

5 Block, R.C., Dorsey, E.R., Beck, C.A., Brenna, J.T. and Shoulson, I. (2010) Altered cholesterol and fatty acid metabolism in Huntington disease. *J Clin Lipidol*, **4**, 17-23.

6 Aziz, N.A. and Roos, R.A. (2013) Characteristics, pathophysiology and clinical management of weight loss in Huntington's disease. **3**, 253-266.

7 Pringsheim, T., Wiltshire, K., Day, L., Dykeman, J., Steeves, T. and Jette, N. (2012) The incidence and prevalence of Huntington's disease: A systematic review and meta-analysis. *Movement Disorders*, **27**, 1083-1091.

8 Evans, S.J., Douglas, I., Rawlins, M.D., Wexler, N.S., Tabrizi, S.J. and Smeeth, L. (2013) Prevalence of adult Huntington's disease in the UK based on diagnoses recorded in general practice records. *Journal of Neurology, Neurosurgery & Psychiatry*, **84**, 1156-1160.

9 Gardiner, S.L., Boogaard, M.W., Trompet, S., de Mutsert, R., Rosendaal, F.R., Gussekloo, J., Jukema, J.W., Roos, R.A.C. and Aziz, N.A. (2019) Prevalence of Carriers of Intermediate and Pathological Polyglutamine Disease-Associated Alleles Among Large Population-Based Cohorts. *JAMA neurology*, **76**, 650-656.

10 Aziz, N.A., Van Der Burg, J.M.M., Landwehrmeyer, G.B., Brundin, P., Stijnen, T. and Roos, R.A.C. (2008) Weight loss in Huntington disease increases with higher CAG repeat number. *Neurology*, **71**, 1506-1513.

11 Leoni, V., Mariotti, C., Nanetti, L., Salvatore, E., Squitieri, F., Bentivoglio, A.R., Bandettini Del Poggio, M., Piacentini, S., Monza, D., Valenza, M. *et al.* (2011) Whole body cholesterol metabolism is impaired in Huntington's disease. *Neuroscience Letters*, **494**, 245-249.

12 Cheng, M.L., Chang, K.H., Wu, Y.R. and Chen, C.M. (2016) Metabolic disturbances in plasma as biomarkers for Huntington's disease. *J Nutr Biochem*, **31**, 38-44.

13 Mastrokolias, A., Pool, R., Mina, E., Hettne, K.M., van Duijn, E., van der Mast, R.C., van Ommen, G., t Hoen, P.A., Prehn, C., Adamski, J. *et al.* (2016) Integration of targeted metabolomics and transcriptomics identifies deregulation of phosphatidylcholine metabolism in Huntington's disease peripheral blood samples. *Metabolomics*, **12**, 137.

14 Patassini, S., Begley, P., Reid, S.J., Xu, J., Church, S.J., Curtis, M., Dragunow, M., Waldvogel, H.J., Unwin, R.D., Snell, R.G. *et al.* (2015) Identification of elevated urea as a severe, ubiquitous metabolic defect in the brain of patients with Huntington's disease. *Biochem Biophys Res Commun*, **468**, 161-166.

15 Quintero Escobar, M., Pontes, J.G.D.M. and Tasic, L. (2021) Metabolomics in degenerative brain diseases. *Brain Research*, **1773**, 147704.

16 Stoy, N., Mackay, G.M., Forrest, C.M., Christofides, J., Egerton, M., Stone, T.W. and Darlington, L.G. (2005) Tryptophan metabolism and oxidative stress in patients with Huntington's disease. *Journal of Neurochemistry*, **93**, 611-623.

17 Gardiner, S.L., van Belzen, M.J., Boogaard, M.W., van Roon-Mom, W.M.C., Rozing, M.P., van Hemert, A.M., Smit, J.H., Beekman, A.T.F., van Grootheest, G., Schoevers, R.A. *et al.* (2017) Huntingtin gene repeat size variations affect risk of lifetime depression. *Translational psychiatry*, **7**, 1277.

18 Gardiner, S.L., Trompet, S., Sabayan, B., Boogaard, M.W., Jukema, J.W., Slagboom, P.E., Roos, R.A.C., van der Grond, J. and Aziz, N.A. (2019) Repeat variations in polyglutamine disease-associated genes and cognitive function in old age. *Neurobiology of aging*, **84**, 236.e217-236.e228.

19 de Mutsert, R., den Heijer, M., Rabelink, T.J., Smit, J.W., Romijn, J.A., Jukema, J.W., de Roos, A., Cobbaert, C.M., Kloppenburg, M., le Cessie, S. *et al.* (2013) The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection. *Eur J Epidemiol*, **28**, 513-523.

20 Shepherd, J., Blauw, G.J., Murphy, M.B., Bollen, E.L., Buckley, B.M., Cobbe, S.M., Ford, I., Gaw, A., Hyland, M., Jukema, J.W. *et al.* (2002) Pravastatin in elderly individuals at risk of vascular disease (PROSPER): a randomised controlled trial. *Lancet*, **360**, 1623-1630.

21 Penninx, B.W., Beekman, A.T., Smit, J.H., Zitman, F.G., Nolen, W.A., Spinhoven, P., Cuijpers, P., De Jong, P.J., Van Marwijk, H.W., Assendelft, W.J. *et al.* (2008) The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *Int J Methods Psychiatr Res*, **17**, 121-140.

22 Albert, P.R. (2015) Why is depression more prevalent in women? *Journal of psychiatry & neuroscience : JPN*, **40**, 219-221.

23 Mäntyselkä, P., Kautiainen, H., Saltevo, J., Würtz, P., Soininen, P., Kangas, A.J., Ala-Korpela, M. and Vanhala, M. (2012) Weight change and lipoprotein particle concentration and particle size: A cohort study with 6.5-year follow-up. *Atherosclerosis*, **223**, 239-243.

24 Tikkanen, E., Jägerroos, V., Holmes, M.V., Sattar, N., Ala-Korpela, M., Jousilahti, P., Lundqvist, A., Perola, M., Salomaa, V. and Würtz, P. (2021) Metabolic Biomarker Discovery for Risk of Peripheral Artery Disease Compared With Coronary Artery Disease: Lipoprotein and Metabolite Profiling of 31 657 Individuals From 5 Prospective Cohorts. *Journal of the American Heart Association*, **10**, e021995.

25 Marston, N.A., Giugliano, R.P., Melloni, G.E.M., Park, J.G., Morrill, V., Blazing, M.A., Ference, B., Stein, E., Stroes, E.S., Braunwald, E. *et al.* (2022) Association of Apolipoprotein B-Containing Lipoproteins and Risk of Myocardial Infarction in Individuals With and Without Atherosclerosis: Distinguishing Between Particle Concentration, Type, and Content. *JAMA cardiology*, **7**, 250-256.

26 Guasch-Ferré, M., Hruby, A., Toledo, E., Clish, C.B., Martínez-González, M.A., Salas-Salvadó, J. and Hu, F.B. (2016) Metabolomics in Prediabetes and Diabetes: A Systematic Review and Meta-analysis. *Diabetes care*, **39**, 833-846.

27 Ahola-Olli, A.V., Mustelin, L., Kalimeri, M., Kettunen, J., Jokelainen, J., Auvinen, J., Puukka, K., Havulinna, A.S., Lehtimäki, T., Kähönen, M. *et al.* (2019) Circulating metabolites and the risk of type 2 diabetes: a prospective study of 11,896 young adults from four Finnish cohorts. *Diabetologia*, **62**, 2298-2309.

28 Cirulli, E.T., Guo, L., Leon Swisher, C., Shah, N., Huang, L., Napier, L.A., Kirkness, E.F., Spector, T.D., Caskey, C.T., Thorens, B. *et al.* (2019) Profound Perturbation of the Metabolome in Obesity Is Associated with Health Risk. *Cell Metabolism*, **29**, 488-500.e482.

29 Kujala, U.M., Mäkinen, V.-P., Heinonen, I., Soininen, P., Kangas, A.J., Leskinen, T.H., Rahkila, P., Würtz, P., Kovanen, V., Cheng, S. *et al.* (2013) Long-term Leisure-time Physical Activity and Serum Metabolome. *Circulation*, **127**, 340-348.

30 Melkani, G.C. (2016) Huntington's Disease-Induced Cardiac Disorders Affect Multiple Cellular Pathways. *Reactive oxygen species (Apex, N.C.)*, **2**, 325-338.

31 Djoussé, L., Knowlton, B., Cupples, L.A., Marder, K., Shoulson, I. and Myers, R.H. (2002) Weight loss in early stage of Huntington's disease. *Neurology*, **59**, 1325-1330.

32 Gardiner, S.L., De Mutsert, R., Trompet, S., Boogaard, M.W., Van Dijk, K.W., Jukema, P.J.W., Slagboom, P.E., Roos, R.A.C., Pijl, H., Rosendaal, F.R. *et al.* (2019) Repeat length variations in polyglutamine disease-associated genes affect body mass index. *International Journal of Obesity*, **43**, 440-449.

33 Mastrokolias, A., Pool, R., Mina, E., Hettne, K.M., van Duijn, E., van der Mast, R.C., van Ommen, G., 't Hoen, P.A.C., Prehn, C., Adamski, J. *et al.* (2016) Integration of targeted metabolomics and transcriptomics identifies deregulation of phosphatidylcholine metabolism in Huntington's disease peripheral blood samples. *Metabolomics*, **12**, 137.

34 Delles, C., Rankin, N.J., Boachie, C., McConnachie, A., Ford, I., Kangas, A., Soininen, P., Trompet, S., Mooijaart, S.P., Jukema, J.W. *et al.* (2018) Nuclear magnetic resonance-based metabolomics identifies phenylalanine as a novel predictor of incident heart failure hospitalisation: results from PROSPER and FINRISK 1997. *Eur J Heart Fail*, **20**, 663-673.

35 de Kluiver, H., Jansen, R., Milaneschi, Y., Bot, M., Giltay, E.J., Schoevers, R. and Penninx, B.W.J.H. (2021) Metabolomic profiles discriminating anxiety from depression. *Acta Psychiatr Scand*, **144**, 178-193.

36 Treangen, T.J. and Salzberg, S.L. (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*, **13**, 36-46.

37 Soininen, P., Kangas, A.J., Würtz, P., Suna, T. and Ala-Korpela, M. (2015) Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Cardiovascular Epidemiology and Genetics. **8**, 192-206.

38 Joshi, R., Wannamethee, G., Engmann, J., Gaunt, T., Lawlor, D.A., Price, J., Papacosta, O., Shah, T., Tillin, T., Whincup, P. *et al.* (2021) Establishing reference intervals for triglyceride-containing lipoprotein subfraction metabolites measured using nuclear magnetic resonance spectroscopy in a UK population. *Annals of clinical biochemistry*, **58**, 47-53.

39 Soininen, P., Kangas, A.J., Wurtz, P., Suna, T. and Ala-Korpela, M. (2015) Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ Cardiovasc Genet*, **8**, 192-206.

40 Faquih, T. (2020). Zenodo, Zenodo, Vol. 2020.

41 Faquih, T., van Smeden, M., Luo, J., le Cessie, S., Kastenmüller, G., Krumsiek, J., Noordam, R., van Heemst, D., Rosendaal, F.R., van Hylckama Vlieg, A. *et al.* (2020) A Workflow for Missing Values Imputation of Untargeted Metabolomics Data. *Metabolites*, **10**.

42 Aziz, N.A., Jurgens, C.K., Landwehrmeyer, G.B., van Roon-Mom, W.M., van Ommen, G.J., Stijnen, T. and Roos, R.A. (2009) Normal and mutant HTT interact to affect clinical severity and progression in Huntington disease. *Neurology*, **73**, 1280-1285.

43 Gardiner, S.L., de Mutsert, R., Trompet, S., Boogaard, M.W., van Dijk, K.W., Jukema, P.J.W., Slagboom, P.E., Roos, R.A.C., Pijl, H., Rosendaal, F.R. *et al.* (2019) Repeat length variations in polyglutamine disease-associated genes affect body mass index. *International journal of obesity (2005)*, **43**, 440-449.

44 R Core Team. (2019). R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. in press.

45 Lee M. A, M.O., Hughes D, Wade K. H, Corbin L. J, McGuinness L. J, Timpson N. J. (2020), in press.

46 Gu, Z., Eils, R. and Schlesner, M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics (Oxford, England)*, **32**, 2847-2849.

47 Gu, Z., Gu, L., Eils, R., Schlesner, M. and Brors, B. (2014) circlize Implements and enhances circular visualization in R. *Bioinformatics (Oxford, England)*, **30**, 2811-2812.

48 StataCorp. (2019). Stata Press, in press., pp. 475-563.

49 Rücker, G. and Schwarzer, G. (2014) Presenting simulation results in a nested loop plot. *BMC Medical Research Methodology*, **14**, 129.

50 Kammer, M. (2022), in press.

51 Baron, R.M. and Kenny, D.A. (1986) The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, **51**, 1173-1182.

52 Aroian, L.A. (1947) The Probability Function of the Product of Two Normally Distributed Variables. *The Annals of Mathematical Statistics*, **18**, 265-271, 267.

53 MacKinnon, D.P., Lockwood, C.M., Hoffman, J.M., West, S.G. and Sheets, V. (2002) A comparison of methods to test mediation and other intervening variable effects. *Psychol Methods*, **7**, 83-104.

11 FIGURE LEGENDS

Figure 1: Flow chart of the exclusion criteria in the NEO, NESDA, and PROSPER studies before and after pooling.

Figure 2: Mean centered CAG repeat size distribution in *HTT* alleles in the pooled and individual datasets.

Figure 3: Circular plots for the 145 metabolites concentrations associated with the larger CAG repeat size in the long *HTT* allele. Each dot represents the effect estimates for the log-transformed metabolite levels. The lines crossing the circles represent the 95% confidence intervals of the estimates. Filled circles denote statistically significant estimates after adjustment for multiple testing (i.e., p<0.00145). The outer numbered rings represent the different metabolite groups.

Figure 4: Estimation of metabolite levels related to VLDL (A), IDL (B), LDL (C), HDL (D), and other metabolites (E) at 15, 20 and 35 CAG repeat sizes.

12 TABLES

Table 1. Baseline characteristics of the three included studies.

	NEO*	PROSPER	NESDA	Overall
N	4,510	4,035	1,712	10,257
Age in years (SD)	55.93 (5.9)	75.79 (3.4)	42.44 (12.9)	61.50 (14.2)
Sex = Female (%)	2,378 (52.7)	2,079 (51.5)	1,129 (65.9)	5,586 (54.5)
Country (%)				
Scotland	0	1,808 (44.8)	0	1,808 (17.6)
Ireland	0	1,448 (35.9)	0	1,448 (14.1)
Netherlands	4,510 (100.0)	779 (19.3)	1,712 (100.0)	7,001 (68.3)
BMI (SD)	26.3 (3.50)	26.8 (4.1)	25.5 (5.0)	28.0 (4.9)
CAG repeats size (median [range])				
HTT Short allele	17 [9, 26]	17 [9, 26]	17 [9, 26]	17 [9, 26]
HTT Long allele	19 [15, 35]	19 [15, 35]	19 [15, 35]	19 [15, 35]

* Means and percentages were weighted to the BMI distribution of the Dutch general.

Table 2: Largest mediation estimates by BMI in the associations between the *HTT* CAG repeats and metabolite levels.

Metabolite	Mediation Estimate	Direct Effect Estimate	Total Effect Estimate	Sobel's Test P-value
Phospholipids in very large HDL	0.0013	-0.0031	-0.0017	0.0060
Free cholesterol in very large HDL	0.0011	-0.0032	-0.0020	0.0063
Total lipids in very large HDL	0.0010	-0.0027	-0.0017	0.0064
Concentration of very large HDL particles	0.0010	-0.0027	-0.0016	0.0063
Total cholesterol in very large HDL	0.0009	-0.0025	-0.0016	0.0069
Cholesterol esters in very large HDL	0.0009	-0.0024	-0.0015	0.0072
Phospholipids in XL VLDL	-0.0021	0.0052	0.0031	0.0060
Cholesterol esters in XL VLDL	-0.0018	0.0060	0.0042	0.0071

13 ABBREVIATIONS

HD: Huntington disease; *HTT*: huntingtin gene; BMI: body mass index; CAG: cytosine-adenine-guanine; CAD: coronary artery disease; PAD: peripheral artery disease; XL-VLDL: extremely large very low density lipoprotein; VLDL: very low density lipoprotein; IDL: intermediate density lipoprotein; LDL: low density lipoprotein; HDL: high density lipoprotein; NEO: Netherlands Epidemiology of Obesity; PROSPER: Prospective Study of Pravastatin in the Elderly at Risk; NESDA: Netherlands Study of Depression and Anxiety; apoB: apolipoprotein B; HDL3: high-density lipoprotein 3 cholesterol; PCR: multiplex polymerase chain reaction.

Chapter 7

PFAS concentrations are associated with a cardio-metabolic risk profile: findings from two population cohorts

Authors: Tariq O. Faquih[1*], Elvire N. Landstra[2*], Astrid van Hylckama Vlieg[1], Ruifang Li-Gao[1,3], Renée de Mutsert[1], Frits R. Rosendaal[1], Raymond Noordam[4], Diana van Heemst[4], Dennis Mook-Kanamori[1, 5], Ko Willems van Dijk[6,7,8], Monique M.B. Breteler [2,9]

[1] Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands
[2] Population Health Sciences, German Centre for Neurodegenerative Diseases (DZNE), Bonn, Germany
[3] Metabolon, Inc. Morrisville, North Carolina, United State of America
[4] Department of Internal Medicine, Section of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands
[5] Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands
[6] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
[7] Department of Internal Medicine, Division of Endocrinology, Leiden University Medical Center, Leiden, The Netherlands
[8] Einthoven Laboratory for Experimental Vascular Medicine, Leiden University Medical Center, Leiden, The Netherlands
[9] Institute for Medical Biometry, Informatics and Epidemiology (IMBIE), Faculty of Medicine, University of Bonn, Germany
*Authors contributed equally to this project

Keywords: PFAS exposure, population cohort, metabolomics, lipoproteins, cardiovascular disease risk, meta-analysis

Corresponding author:
Monique M.B. Breteler
Monique.Breteler@dzne.de
Venusberg-Campus 1, Gebaeude 99
53127 Bonn, Germany

Acknowledgements
We would like to thank all participants and the study personnel of the Rhineland Study. The authors of the NEO study thank all participants, all participating general practitioners for inviting eligible participants, all research nurses for data collection, and the NEO study group: Pat van Beelen, Petra Noordijk, and Ingeborg de Jonge for coordination, laboratory, and data management.

Manuscript in preparation.

1 ABSTRACT

Per- and polyfluoroalkyl substances (PFAS) are widely used and persistent chemicals leading to ubiquitous and constant exposure. Although high PFAS levels have been associated with hypercholesteremia and cardiovascular disease, the levels of PFAS and associations with metabolic risk markers in general population samples are not fully characterized. We thus aimed to assess the associations between perfluorooctaneic acid (PFOA), perfluorooctane sulfonic acid (PFOS), and perfluorohexanesulfonic acid (PFHxS) and a range of metabolites as well as clinical lipid measurements. For the analysis, we used participants with complete clinical lipid, metabolite and PFAS measurements from the NEO (n= 586) and Rhineland Study (n= 1,962). The metabolites were measured with nuclear magnetic resonance by Nightingale and were mainly comprised of lipoprotein characteristics. Using linear regression analyses, we quantified age-, sex- and education-adjusted associations of PFOA, PFOS, and PFHxS (Rhineland Study only) with clinical lipid measurements and metabolites (n= 224).

In line with previous research, both studies confirmed that PFAS, particularly PFOS and PFHxS, were associated with higher clinically measured LDL and cholesterol concentrations. We uniquely showed that this was characterized by higher concentrations of total lipid, cholesterol and phospholipid content in LDL particles in particular. We also showed an interaction effect of age on the majority of associations, where the effect of PFAS was stronger in younger people (≤ 54 years). Thus, our results show that even low PFAS concentrations are associated with an unfavorable lipid profile in the general population. This emphasizes the need for further regulation of PFAS substances.

2 INTRODUCTION

Per- and polyfluoroalkyl substances (PFAS), colloquially known as the 'forever chemicals', are man-made chemicals that have been widely used in many industrial processes and products since the 1950s (1). These chemicals are persistent and resilient in nature, allowing them to circulate in water sources and become widespread around the globe (2-9). High PFAS levels may be caused by contamination in the vicinity of PFAS-producing factories, but also by the use and breakdown of PFAS-containing products—such as fire extinguishers, non-stick cooking pans, certain food packaging and pesticides (6, 10). Contact with this contaminated environment or these products may lead to exposure in humans via media such as drinking water, inhalation or dermal exposure. Exposure may also be indirect, for example through the diet (6). Given the persistence of PFAS, human exposure is continuous and ubiquitous.

Soon after the introduction of PFAS production, several health concerns were raised. Direct high exposure of PFAS has been associated with various adverse health outcomes, including obesity, kidney disease, cancer, thyroid disease, hypercholesteremia, dyslipidemia, liver damage, reduced antibody response to vaccination, and a higher risk of severe course of COVID-19 infection (4, 7-9). Previous research has associated PFAS with changes in the immune system, proteome (11), hormones (12), and metabolome (11, 12). Furthermore, a 2020 report by the European Food Safety Authority (EFSA) indicated that the risks associated with PFAS were even more severe than previously believed. They also reported that low PFAS levels, previously thought to be within the safe limits, could pose a health risk as well (8, 13, 14). However, while high PFAS exposure has been the focus of the majority of PFAS studies, much remains unclear about the effects at low exposure in the general population (6, 15).

Specific mechanisms through which PFAS exert their effects are unclear. Previous studies have shown that PFAS have been most consistently associated with changes in lipid metabolism, particularly higher cholesterol levels. Thus far, those studies have only considered traditional, composite lipid measurements, such as total, HDL and LDL cholesterol, and it remains unclear how PFAS affect the deeper metabolic and lipoprotein profiles (4).

The health concerns from PFAS exposure resulted in the classification of certain PFAS as persistent pollutants that require regulation. Specifically, perfluorooctane sulfonic acid (PFOS) and perfluorooctanoic acid (PFOA) have been subjected to growing restrictions (7, 16). However, despite the increasing scrutiny, many PFAS species remain unregulated and even PFOA and PFOS are not yet fully banned in the European Union (17). Moreover, despite efforts to reduce and discontinue PFOA and PFOS production in the Netherlands (17, 18) and Germany (19), levels and exposure remain an issue (13, 14, 20).

Whilst PFAS exposure thus continues, consequences of regular exposure to PFAS in the general population and the metabolomic effects of PFAS remain understudied. Here, we aimed to evaluate the association of PFAS levels in the general population with metabolites and lipoproteins using clinical lipid measurements and targeted metabolomics. To improve generalizability and robustness of the results, we performed the analysis in two study populations: the Netherlands Epidemiology of Obesity (NEO) study (n= 586) and the Rhineland study (n= 1,962).

3 METHODS

3.1 Study Populations

3.1.1 Netherlands Epidemiology of Obesity Study

The Netherlands Epidemiology of Obesity (NEO) study is a population-based, prospective cohort study of individuals aged 45–65 years, with an oversampling of individuals who are overweight or have obesity. Men and women aged between 45 and 65 years with a self-reported BMI of 27 kg/m2 or higher, living in the greater area of Leiden (in the West of the Netherlands) were eligible to participate in the NEO study. In addition, all inhabitants aged between 45 and 65 years from one municipality (Leiderdorp) were invited, irrespective of their BMI. Recruitment of participants started in September 2008 and was completed at the end of September 2012. In total, 6,671 participants have been included. Participants were invited to come to the NEO study center of the LUMC for one baseline study visit after an overnight fast. A blood sample of 108 mL was taken from the participants after an overnight fast of at least 10 hours (21). In the current study, only participants from the Leiderdorp sub-population with untargeted metabolomic data were included, leading to a sample size of 599 individuals. Among these individuals, 4 did not have targeted metabolomics data and 9 were excluded due to measurement errors, leading to a final sample size of n= 586 (**Figure 1**).

3.1.2 Rhineland Study

The Rhineland Study is an ongoing community-based cohort study located in two geographically defined areas in Bonn, Germany. Participation is possible by invitation only. To be able to participate, participants had to be above 30 years of age at baseline and have sufficient command in German to provide informed consent. The current analysis was based on the first 5,000 subsequent participants whose plasma was analyzed for PFAS levels (n= 4,469), Nightingale metabolites (n= 1,982) and whose education level was known, resulting in a final analytical sample of 1,962 participants (**Figure 1**). Of these, 1,805 participants had complete data on LDL, HDL, total triglyceride, and cholesterol levels, which were used for the additional analysis on clinical lipid measurements.

3.2 PFAS measurements

Relative PFAS concentrations were acquired on the untargeted Metabolon™ Discovery HD4 platform at Metabolon Inc. (Durham, North Carolina, USA). In brief, this process involves four independent ultra-high-performance liquid chromatography mass spectrometry (UHPLC-MS/MS) platforms (22, 23). It uses two positive ionization reverse phase chromatography, one negative ionization reverse phase chromatography, and one hydrophilic interaction liquid chromatography negative ionization (23). In NEO, the PFOA and PFOS levels were quantified from fasting state serum samples, while PFOA, PFOS and perfluorohexanesulfonic acid (PFHxS) concentrations were obtained from fasting plasma samples in the Rhineland Study. PFAS were quantified as relative concentrations. To ensure normality and comparability across studies, we log-transformed and Z-standardized all PFAS concentrations for the subsequent analyses.

3.3 Metabolites and lipoprotein measurements

Total serum levels of cholesterol, triglycerides, LDL, and HDL were measured in NEO as described in previous work (21). In the Rhineland Study, serum levels of cholesterol, triglycerides, LDL and HDL were measured with routine measurements at the University Hospital in Bonn.

Fasting serum (NEO) and plasma (Rhineland Study) metabolite levels were quantified using the Nightingale (Nightingale Health Ltd, Helsinki, Finland) nuclear magnetic resonance untargeted metabolomic platform. This platform quantified 224 metabolites and metabolite ratios. The data predominantly contains detailed lipoprotein lipid information. Additionally, absolute concentrations of various other metabolites, including amino acids, free fatty acids and ketone bodies, were quantified (24).

3.4 Assessment of covariates

In both the Rhineland Study and NEO, questionnaire and food frequency questionnaires were used to collect demographic and lifestyle information, including smoking history (yes/no), alcohol intake (g/day), and education (low/middle/high). In the Rhineland Study, missing values for smoking were imputed based on HD4-measured cotinine levels in the blood according to the method described by St Helen et al (25). To assure the quality of the data, alcohol values belonging to participants reporting an overall improbable caloric intake (<600 or >8,000) were excluded, as per the method of Galbete et al. (26).

The International Standard Classification of Education 2011 (ISCED) (27) was used to standardize education across both studies. Participants' education level was reclassified as low (lower secondary education or below), middle (upper secondary education to undergraduate university level) or high (postgraduate university study) in both studies. In the NEO study, the education status of two participants was missing and set to "low".

3.5 Statistical Analysis

3.5.1 Imputation of Missing Metabolite Values

A total of 224 metabolites were measured on the targeted metabolomics Nightingale platform. These metabolites had missing values ranging from 0.3% to 9%. Missing values were set to 0 where the levels were believed to be below detection. Missingness due to other reasons was imputed using a previously described pipeline (28). Accordingly, imputed datasets were generated using multiple imputation. To ensure normality, all metabolites were log-transformed after adding 1 to account for 0s.

3.5.2 Linear Regression Models and Meta-Analysis

Multiple linear regression models were performed to associate the log-transformed Nightingale and clinical lipid measurements (outcomes) with the log-transformed and Z-standardized PFAS concentrations (exposures). In model 1, we adjusted for sex (women/men), age (years) and education (low/middle/high). As PFAS accumulate during the lifespan and may have a different effect with increasing age, model 2 additionally included a multiplicative interaction term between age and the PFAS substances. We also assessed the possible difference between men and women in model 3 by including a sex-interaction term (29). If the interaction estimate was significant, a stratified analysis was performed on the basis of sex or the median age (54 years). For ease of comparability in the figures, analyses were additionally run on the Z-standardized metabolite levels after the log transformation. To test whether the associations were dependent on traditional risk behaviors, we also ran an analysis adjusting for alcohol intake, smoking, and body mass index (BMI). Lastly, we assessed the robustness of our results by performing a sensitivity analysis where we truncated extreme outlier values (>5 standard deviations) in the metabolite levels.

A meta-analysis was performed on the metabolites and stratified if the age-interaction term was significant. To improve the assessment of the confidence intervals and account for heterogeneity for the two studies, we conducted a meta-analysis using both the fixed and the random model.

3.5.3 Multiple testing correction

As the Nightingale measurements are inherently highly correlated, we used the method described by Li and Ji (30) to calculate the effective number of independent variables. Accordingly, the number of independent metabolite variables was estimated at 66 (70 including clinical lipid measurement outcomes) in both studies. Thus, we considered a p-value < 0.0007 (0.05/70) as statistically significant.

All analyses were performed using R (31) v4.1.0 (2021-05-18) in NEO and v4.0.5 (2021-03-31) in the Rhineland study. The meta-analysis was conducted using the *rmeta* package. Figures were created using the *ggplot 2*(32) R package.

4 RESULTS

4.1 Population characteristics

Participants in the Rhineland Study had a median age of 54 years (range: 30 – 89), consisted of 57% women, had an average BMI of 25.8, and were relatively highly educated (**Table 1**). NEO participants had a median age of 54 years (range: 45 – 66) and were composed of 52.6% women. Alcohol intake in the Rhineland Study was higher than in NEO. Smokers made up 13.7% and 11.3% of the Rhineland and NEO study, respectively. PFOA and PFOS was measured in all 599 participants of the NEO study. In the Rhineland Study, PFOA, PFOS, and PFHxS were measured in 1967, 1950, and 1964 out of 1967 participants, respectively.

4.2 Linear Regression

4.2.1 Overview

The role of sex and age in the association between PFAS and clinical lipid measures and metabolites was assessed by interaction analyses. No consistent difference in the associations between men and women was found. Contrastingly, we found a significant age-interaction for the majority of the associations (**Table 2, Table 3, Table 4, Table 5**).

4.2.2 PFOA

No associations were found between PFOA and clinically measured lipids in model 1 without an interaction term (**Table 2**, **Table 3**). In the age-interaction model, PFOA was associated with higher LDL levels in the Rhineland Study. The age-interaction term was negative, indicating a smaller effect with increasing age. Contrastingly, analyses after age stratification showed that the effect of PFOA on LDL levels in the Rhineland study was smaller in the younger age group (≤54 years) compared to the older group (>54 years) (**Table 3**). In NEO, PFOA was nominally associated with increased cholesterol concentrations, but not associated after adjustment for multiple testing (**Table 2**).

In model 1, we found 3/224 and 2/224 metabolites associated with PFOA levels in the NEO and Rhineland studies respectively (**Table 4**, **Table 5**). In NEO, these metabolites were related

to lipids in the small HDL particles while in the Rhineland study they were associated with cholesterol in HDL 3, sphingomyelins, and albumin. After adding the PFOA-age interaction term, 3/224 and 1/224 PFOA associations showed a significant age-interaction in the NEO and Rhineland studies respectively. In contrast to model 1, these all related to LDL and cholesterol content in NEO, while only the LDL size was associated in the Rhineland Study. In NEO, the age-interaction term indicated a stronger effect in younger people, which the age-stratified analysis confirmed, while the age-stratified analysis showed conflicting results in the Rhineland Study (**Table 4**, **Table 5**).

4.2.3 PFOS

In model 1, no associations were found between PFOS and the 4 clinical lipid measurements in either study. After the addition of the PFOS-age interaction term, we found that PFOS was associated with elevated levels of cholesterol and LDL in both NEO and the Rhineland Study, where the age-interaction term indicated a weaker effect with increasing age. This was echoed by the stratified analysis, which showed a stronger effect in the younger age group (≤54 years) in both studies (**Table 2**, **Table 3**).

In model 1, PFOS was associated with 3/224 metabolites in the NEO study, while no significant associations were found in the Rhineland Study (**Table 4**, **Table 5**). All significant associations in the NEO study were related to fatty acids. The age-interaction analysis revealed 53/224 and 80/224 associations between the PFOS-age interaction term and metabolites in the NEO and Rhineland studies, respectively (**Table 4**, **Table 5**). In both studies, these were primarily related to LDL, VLDL, IDL, apolipoprotein B (apoB), and fatty acids. In the Rhineland Study, PFOS was also associated with valine, phosphatidylcholines, albumin, phosphoglycerides and sphingomyelins. Overall, the age-interaction term indicated a weaker effect with increasing age. This was confirmed by the age-stratified analysis, which showed a stronger effect in the younger age group (≤ 54 years) compared to the older group (>54 years) (**Table 4**, **Table 5**).

4.2.4 PFHxS

PFHxS was measured only in the Rhineland study and was not associated with any of the clinical lipid measurements in model 1. In the age-interaction model, PFHxS was associated with higher cholesterol levels, the effect of which was weaker with older age. In the age-stratified analysis, the association between PFHxS and cholesterol was indeed stronger in the younger group (≤54 years) (**Table 3**).

PFHxS was associated with 40/224 of the metabolites in model 1. PFHxS was generally associated with increased levels of cholesterols, LDLs, fatty acids, albumin, and apolipoprotein A (apoA) (**Table 5**). On the other hand, PFHxS was associated with a decrease in the amino acid phenylalanine only. When adding the age-interaction term, we observed an age effect in 8/224 metabolites, namely fatty acids, cholesterols, phosphoglycerides, and phosphatidylcholines. Although the age-interaction term always showed a weakening of the effect with age, the age-stratified analysis only confirmed this for the fatty acids and phosphoglycerides. Contrastingly, the effect of PFHxS was stronger in the older group (>54 years) for cholesterols (**Table 5**).

4.3 Sensitivity analyses

When additionally adjusting for smoking, alcohol consumption, and BMI in the complete data in the NEO (n= 586) and the Rhineland studies (n= 1,733), the number of associations generally increased slightly for both the model with and without the age-interaction (**Supplementary**

Table 1). The overall results, however, remained consistent for PFOA and PFOS. Results for PFHxS in model 1 did not change substantially either. Contrastingly, we found an additional 15 significant age-interactions when adjusting for the aforementioned covariates. New associations mainly comprised of VLDL, and LDL and apoB (**Supplementary Table 1**).

A separate sensitivity analysis using truncated outliers of PFAS concentrations was also performed. In the Rhineland Study, the number of significant associations increased for PFOA (n= 9), PFOS (n= 68), and PFHxS (n= 61) (**Supplementary Table 1**). Generally, these associations spanned the categories of apolipoproteins, cholesterols, glycerides and phospholipids, HDL, IDL, LDL and VLDL. PFAS concentrations thus showed an even stronger consistent association with cholesterols and LDL concentrations and composition in particular. In the age-interaction model, we found fewer age-interactions after accounting for outliers for PFOS (n= 28). Remaining associations included apoB, cholesterols, fatty acids, LDLs and VLDLs. In the NEO study, the results of the sensitivity analyses remained largely consistent with the main results (**Supplementary Table 1**).

4.4 Meta-analysis

In the meta-analysis, we found that the heterogeneity (I2) was low in all analyses, indicating that the two populations are similar. Furthermore, the fixed and random effect model estimates and confidence intervals overlapped and were in the same direction. Hence, the fixed effect model was an appropriate method for this analysis.

In the meta-analysis for model 1, PFOA was associated with higher levels of clinically measured cholesterol and HDL, as well as a number of metabolites. The latter included the total concentrations of cholesterols, LDL, VLDL, IDL, HDL, and the lipid content of these lipoproteins. Moreover, PFOA was associated with higher levels of amino acids, fatty acids, and glycerides. Similarly, increased levels of PFOS were associated with higher clinically measured cholesterol and LDL, as well as various metabolomics measurements. Specifically, PFOS was associated with increases in the levels of LDL, IDL and VLDL, as well as their lipid content. Furthermore, it was associated with cholesterols, amino acids, fatty acids and HDL content, specifically that of the very large HDL particles (**Supplementary Table 2).**

In the age-stratified meta-analysis, we showed similar results to the study-specific analysis. Specifically, the different PFAS were associated with higher levels of clinical lipids and some metabolites (**Supplementary Table 3**). The PFOA-associated metabolites belonged to the groups of small VLDL, omega fatty acids and the size of LDL. On the other hand, PFOS was associated with valine, various fatty acids, albumin, phosphoglycerides, apoB, and cholesterols as well as the composition and concentration of IDL, LDL, and VLDL. Unlike the meta-analysis for model 1, no associations with any HDL metabolites or clinical HDL measurements were found. When comparing the different age-groups, we found more significant associations in the younger age group (age ≤54) compared to the older age group (>54). Furthermore, fixed effect estimates tended to be stronger in the younger group.

5 DISCUSSION

5.1 Summary

In this study, we investigated the association of three PFAS substances with clinically measured lipid biomarkers and a wide range of metabolites (n= 224) in the general population. By combining

7

findings from the NEO Study (n= 586) and the Rhineland Study (n= 1,962), we report common and clinically relevant effects of PFAS on lipid metabolism. In particular, PFAS molecules were associated with higher levels of clinically measured LDL and cholesterol, which was confirmed by the association with lipoprotein metabolites. Specifically, PFOS and PFHxS were associated with a metabolomic profile characterized by increased levels of apoB, phosphoglycerides, total lipids, fatty acids, and the lipid content in LDL, IDL, and VLDL. Meta-analyses showed a similar trend across the populations with small heterogeneity, further strengthening our findings. Thus, we interpret these data to indicate that even low PFAS levels in the general population have a detrimental effect on lipid metabolism.

5.2 Widespread PFAS exposure in the Netherlands and Germany

PFAS production, and thus exposure, in both Germany and the Netherlands started in the second half of the last century. In the Netherlands, production of PFAS substances began at the DuPont/ Chemours plant in Dordrecht in 1967 (33). Although the production of the so-called legacy PFAS (PFOA and PFOS) was slowly phased out in 2012 and replaced by "GenX", both surface and ground water, soil, vegetation, fish, and stock animals in the area remain highly contaminated. In addition, contamination in surface and drinking water was detected across the western regions of the Netherlands (34). The National Institute for Public Health and the Environment (RIVM) has reported that the Dutch population is likely ingesting PFAS levels above the recommended safe levels via food and water (14). Moreover, they advised against consuming fruits or vegetables grown from gardens within a 1 km radius of the Dordrecht Chemours plant and from the Westerschelde area downstream of the plant (35). Accordingly, the RIVM has concluded that the levels of PFAS in the Netherlands are highly concerning and require further research (14). In Germany, PFAS have been produced since 1968 at the Chemiepark Gendorf (CPG) (36). Therefore, contamination is widespread in the ecosystem and PFAS are detectable in drinking water (37, 38). Specifically, areas along the Rhine, Ruhr and Moehne rivers are marked as high exposure locations (37). For example, contaminated paper sludge caused high PFAS levels in the drinking water in Rastatt county in the Baden-Wuerttemberg state (37) and the town of Arnsberg in the state of North Rhine-Westphalia (39). Importantly, PFAS levels are detectable in the groundwater of most of the provinces and in all of the soil samples (19). It is thus clear that PFAS exposure is widespread in both Germany (19, 20, 37, 39) and the Netherlands (14, 34, 35, 40). Indeed, despite the production of legacy PFAS being slowly phased out over the years, PFAS levels were detectable in nearly all included participants from the NEO and the Rhineland studies.

5.3 PFAS levels associated with metabolic profile of increased risk of cardiovascular disease

Here, we showed that even low concentrations of PFAS were associated with a distinctive lipid profile. Overall, PFAS substances were associated with an increase in clinically measured total LDL and cholesterol. Further investigation using metabolomics revealed that higher PFAS levels were characterized by elevations in cholesterol and lipid content in LDLs and VLDLs of all sizes. We also found associations with apoB, fatty acids, phosphoglycerides, IDL, and phospholipids. Previously, a higher lipid content of lipoproteins was implicated in cardiovascular disease (CVD) (41), while higher levels of fatty acids and apoB were consistently associated with myocardial infarction (42). Other studies have linked a similar metabolomic profile to a higher risk of cardiometabolic diseases such as CVD (43), hypertension (44), type 2 diabetes (DM2) (45), and non-alcoholic fatty liver disease (46). Therefore, our results suggest that PFAS exposure may increase the risk of cardiometabolic outcomes by impacting the lipoprotein composition. Of note, we

found that PFHxS, one of the often-used substitutes of PFOA and PFOS (6, 47), similarly affected metabolite concentrations, as well as lipoprotein composition and concentrations.

Importantly, the abovementioned results for PFOA and PFOS were further strengthened by our meta-analysis, which showed that PFAS might have clinically relevant effects in the European population. For example, every one standard deviation increase in PFOS in the younger group was associated with a 0.03 increase in log-transformed LDL levels across the two populations, which is equal to an LDL increase of 0.1 mmol/L (CI: 0.04 – 0.20). Recommended thresholds for LDL in patients with CVD or DM2 are <1.8 mmol/L and <2.6 mmol/L, respectively. As such, we show that even general population levels of PFAS might have a clinically relevant effect.

5.4 Effects of PFAS are partially dependent on Age

We also reported a significant age-interaction for the majority of metabolite PFOA and PFOS associations, which generally showed a weakening of the effect in older individuals. On the other hand, PFHxS showed a weaker age effect. Finding a weakening of the effect of PFAS is therefore not unexpected, as other competing causes might have a higher relative importance than PFAS exposure. In people within the age range of 40-50, the absence of medication or other lifetime-accumulated exposures might therefore mean that PFAS have a higher relative impact on their lipid levels. Alternatively, the time of exposure might also impact the effect and explain the difference. Indeed, children have been reported to suffer from more severe effects than adults (48).

In the Netherlands, pre-determined cut-offs for safe levels of PFAS were recently used to conclude that the extremely high PFAS levels in the Westerschelde were no cause for concern (35). However, our results indicate that in the general population low levels of PFAS are associated with a detrimental lipid profile. Moreover, associations were robust and even increased in number in the sensitivity analysis. Taken together with previous literature, stricter regulations are required for all PFAS substances. Furthermore, due to the persistent nature of PFAS and their recirculation in the environment, there is a need to actively remove these chemicals from the environment—methods for which are under development (49).

5.5 Strengths and limitations

Our main limitation stems from the cross-sectional nature of our data. As such, we cannot establish a causal link between the PFAS exposures and the metabolites. Nonetheless, our data shows a clear link between PFAS and a cardio-metabolic risk profile across two European populations. These results are in line with previous findings, and as such, should be taken into consideration when assessing the risk and required regulation of PFAS across the whole population. Other limitations include the use of different sample media for the measurement of PFAS and Nightingale metabolites in NEO versus the Rhineland Study, as well as the use of relative, rather than absolute, PFAS concentrations. Despite these limitations, the inclusion of two European countries demonstrates a consistent and robust association with PFAS levels. Furthermore, we evaluated their relation to a detailed lipid profile and a large variety of metabolites.

6 CONCLUSION

In conclusion, our results expand on previous findings by showing a clear link between a harmful lipid profile and PFAS concentrations across different study populations, even at low PFAS levels. We report an association with increased LDL and total cholesterol as well as apoB and lipid content in LDL, VLDL, and IDL lipoproteins. The effect generally weakened with increasing age, indicating that PFAS exposure is particularly detrimental at a younger age. The combination of the well-documented persistence of PFAS and their harmful effects ensures that exposure to these substances is an enduring public health challenge. Thus, there is a clear need for further studies in general populations, as well as regulation and efforts to reduce environmental PFAS levels.

7 REFERENCES

1. Roth K, Yang Z, Agarwal M, Liu W, Peng Z, Long Z, et al. Exposure to a mixture of legacy, alternative, and replacement per- and polyfluoroalkyl substances (PFAS) results in sex-dependent modulation of cholesterol metabolism and liver injury. Environment International. 2021;157:106843.
2. Nordby GL, Luck JM. Perfluorooctanoic acid interactions with human serum albumin. The Journal of biological chemistry. 1956;219(1):399-404.
3. Buck RC, Franklin J, Berger U, Conder JM, Cousins IT, de Voogt P, et al. Perfluoroalkyl and polyfluoroalkyl substances in the environment: Terminology, classification, and origins. 2011;7(4):513-41.
4. Priestly B. Literature review and report on the potential health effects of Perfluoroalkyl compounds, mainly Perfluorooctane Sulfonate (PFOS). Monash University, Services DoHH; 2018.
5. RIVM. PFAS [Available from: https://www.rivm.nl/pfas.
6. Sunderland EM, Hu XC, Dassuncao C, Tokranov AK, Wagner CC, Allen JG. A review of the pathways of human exposure to poly- and perfluoroalkyl substances (PFASs) and present understanding of health effects. Journal of Exposure Science & Environmental Epidemiology. 2019;29(2):131-47.
7. Stockholm Convention on Persistent Organic Pollutants (POPs). The new POPs under the Stockholm Convention: Stockholm Convention on Persistent Organic Pollutants; 2001 [Available from: http://www.pops.int/TheConvention/ThePOPs/TheNewPOPs/tabid/2511/Default.aspx.
8. Schrenk D, Bignami M, Bodin L, Chipman JK, Del Mazo J, Grasl-Kraupp B, et al. Risk to human health related to the presence of perfluoroalkyl substances in food. EFSA Journal. 2020;18(9).
9. Grandjean P, Timmermann CAG, Kruse M, Nielsen F, Vinholt PJ, Boding L, et al. Severity of COVID-19 at elevated exposure to perfluorinated alkylates. PLOS ONE. 2020;15(12):e0244815.
10. Barhoumi B, Sander SG, Tolosa I. A review on per- and polyfluorinated alkyl substances (PFASs) in microplastic and food-contact materials. Environmental Research. 2022;206:112595.
11. Liu H, Sun W, Zhou Y, Griffin N, Faulkner S, Wang L. iTRAQ-based quantitative proteomics analysis of Sprague-Dawley rats liver reveals perfluorooctanoic acid-induced lipid metabolism and urea cycle dysfunction. Toxicology letters. 2022;357:20-32.
12. Shih YH, Blomberg AJ, Jørgensen LH, Weihe P, Grandjean P. Early-life exposure to perfluoroalkyl substances in relation to serum adipokines in a longitudinal birth cohort. Environ Res. 2022;204(Pt A):111905.
13. EFSA. PFAS in food: EFSA assesses risks and sets tolerable intake %J European Food Safety Authority. 2022.
14. RIVM. Te veel blootstelling aan PFAS in Nederland 2021 [updated 04-06-2021. Available from: https://www.rivm.nl/nieuws/te-veel-blootstelling-aan-pfas-in-nederland.
15. European Food Safety Authority. Outcome of a public consultation on the draft risk assessment of perfluoroalkyl substances in food. 2020;17(9):1931E.
16. Stockholm Convention on Persistent Organic Pollutants (POPs). The 16 New POPs: An introduction to the chemicals added to the Stockholm Convention as Persistent Organic Pollutants by the Conference of the Parties. 2017.
17. RIVM. Official start to ban PFAS in Europe | RIVM. 2022.
18. Zeilmaker MJJ, P. ; Versteegh, A. ; Pul, A. van ; Vries, W. de ; Bokkers, B. ; Wuijts, S. ; Oomen, A. ; Herremans, J. Risicoschatting emissie PFOA voor omwonenden: Bilthoven: National Institute for Public Health and the Environment; 2016.
19. German Environment Agency. PFAS Came to stay. What Matters. 2020:48.
20. Duffek A, Conrad A, Kolossa-Gehring M, Lange R, Rucic E, Schulte C, et al. Per- and polyfluoroalkyl substances in blood plasma – Results of the German Environmental Survey for children and adolescents 2014–2017 (GerES V). International Journal of Hygiene and Environmental Health. 2020;228:113549.
21. de Mutsert R, den Heijer M, Rabelink TJ, Smit JW, Romijn JA, Jukema JW, et al. The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection. Eur J Epidemiol. 2013;28(6):513-23.

22. Evans A, Bridgewater B, Liu Q, Mitchell M, Robinson R, Dai H, et al. High resolution mass spectrometry improves data quantity and quality as compared to unit mass resolution mass spectrometry in high-throughput profiling metabolomics. 2014;4(2):1.
23. Rhee EP, Waikar SS, Rebholz CM, Zheng Z, Perichon R, Clish CB, et al. Variability of Two Metabolomic Platforms in CKD. Clinical Journal of the American Society of Nephrology. 2019;14(1):40.
24. Soininen P, Kangas AJ, Wurtz P, Suna T, Ala-Korpela M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. Circ Cardiovasc Genet. 2015;8(1):192-206.
25. St Helen G, Novalen M, Heitjan DF, Dempsey D, Jacob P, 3rd, Aziziyeh A, et al. Reproducibility of the nicotine metabolite ratio in cigarette smokers. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2012;21(7):1105-14.
26. Galbete C, Kröger J, Jannasch F, Iqbal K, Schwingshackl L, Schwedhelm C, et al. Nordic diet, Mediterranean diet, and the risk of chronic diseases: the EPIC-Potsdam study. BMC Medicine. 2018;16(1):99.
27. UNESCO Institute for Statistics. International Standard Classification of Education ISCED 2011. Montreal, Quebec H3C 3J7 Canada: UNESCO Institute for Statistics 2012. p. 88.
28. Faquih T, van Smeden M, Luo J, le Cessie S, Kastenmüller G, Krumsiek J, et al. A Workflow for Missing Values Imputation of Untargeted Metabolomics Data. Metabolites. 2020;10(12).
29. Koskela A, Ducatman A, Schousboe JT, Nahhas RW, Khalil N. Perfluoroalkyl Substances and Abdominal Aortic Calcification. Journal of occupational and environmental medicine. 2022;64(4):287-94.
30. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity. 2005;95(3):221-7.
31. R Core Team. R: A language and environment for statistical computing. Vienna, Austria. URL https://www.R-project.org/. R Foundation for Statistical Computing; 2019.
32. Hadley W. ggplot2: Elegant Graphics for Data Analysis. 2016.
33. Municipality of Dordrecht. Gezondheid - Dordrecht: Dordrecht; 2022 [updated 2022/05/24/. Available from: https://cms.dordrecht.nl/Inwoners/Overzicht_Inwoners/Dossier_Chemours_en_DuPont/Gezondheid.
34. Gebbink WA, van Leeuwen SPJ. Environmental contamination and human exposure to PFASs near a fluorochemical production plant: Review of historic and current PFOA and GenX contamination in the Netherlands. Environment International. 2020;137:105583.
35. RIVM. PFAS in de Westerschelde: Eet zo min mogelijk zelf gevangen producten | RIVM. 2022.
36. Umwelt BLf. PFOA-Problematik im Raum Gendorf - LfU Bayern. 2022.
37. Kotthoff M, Fliedner A, Rüdel H, Göckener B, Bücking M, Biegel-Engler A, et al. Per- and polyfluoroalkyl substances in the German environment – Levels and patterns in different matrices. Science of The Total Environment. 2020;740:140116.
38. Brendel S, Fetter É, Staude C, Vierke L, Biegel-Engler A. Short-chain perfluoroalkyl acids: environmental concerns and a regulatory strategy under REACH. Environmental Sciences Europe. 2018;30(1):9.
39. Brede E, Wilhelm M, Göen T, Müller J, Rauchfuss K, Kraft M, et al. Two-year follow-up biomonitoring pilot study of residents' and controls' PFC plasma levels after PFOA reduction in public water system in Arnsberg, Germany. International Journal of Hygiene and Environmental Health. 2010;213(3):217-23.
40. van der Aa M, Hartmann J, te Biesebeek JD. Analyse bijdrage drinkwater en voedsel aan blootstelling EFSAEuropese Voedselveiligheidsautoriteit-4 PFASPoly- en perfluoralkylstoffen in Nederland en advies drinkwaterrichtwaarde. Rijksinstituut voor Volksgezondheid en Milieu. Ministerie van Volksgezondheid, Welzijn en Sport; 2021.
41. Xiao C, Dash S, Morgantini C, Hegele RA, Lewis GF. Pharmacological Targeting of the Atherogenic Dyslipidemia Complex: The Next Frontier in CVD Prevention Beyond Lowering LDL Cholesterol. Diabetes. 2016;65(7):1767-78.
42. Julkunen H, Cichońska A, Tiainen M, Koskela H, Nybo K, Mäkelä V, et al. Atlas of plasma nuclear magnetic resonance biomarkers for health and disease in 118,461 individuals from the UK Biobank. 2022:2022.06.13.22276332.

43. Tikkanen E, Jägerroos V, Holmes MV, Sattar N, Ala-Korpela M, Jousilahti P, et al. Metabolic Biomarker Discovery for Risk of Peripheral Artery Disease Compared With Coronary Artery Disease: Lipoprotein and Metabolite Profiling of 31 657 Individuals From 5 Prospective Cohorts. Journal of the American Heart Association. 2021;10(23):e021995.
44. Palmu J, Tikkanen E, Havulinna AS, Vartiainen E, Lundqvist A, Ruuskanen MO, et al. Comprehensive biomarker profiling of hypertension in 36 985 Finnish individuals. 2022;40(3):579-87.
45. Ahola-Olli AV, Mustelin L, Kalimeri M, Kettunen J, Jokelainen J, Auvinen J, et al. Circulating metabolites and the risk of type 2 diabetes: a prospective study of 11,896 young adults from four Finnish cohorts. 2019;62(12):2298-309.
46. Sliz E, Sebert S, Würtz P, Kangas AJ, Soininen P, Lehtimäki T, et al. NAFLD risk alleles in PNPLA3, TM6SF2, GCKR and LYPLAL1 show divergent metabolic effects. Human Molecular Genetics. 2018;27(12):2214-23.
47. Li J, He J, Niu Z, Zhang Y. Legacy per- and polyfluoroalkyl substances (PFASs) and alternatives (short-chain analogues, F-53B, GenX and FC-98) in residential soils of China: Present implications of replacing legacy PFASs. Environment International. 2020;135:105419.
48. Canova C, Di Nisio A, Barbieri G, Russo F, Fletcher T, Batzella E, et al. PFAS Concentrations and Cardiometabolic Traits in Highly Exposed Children and Adolescents. International journal of environmental research and public health. 2021;18(24).
49. Trang B, Li Y, Xue X-S, Ateia M, Houk KN, Dichtel WR. Low-temperature mineralization of perfluorocarboxylic acids. 2022;377(6608):839-45.

7

8 STATEMENTS AND DECLARATIONS

Funding
The NEO study is supported by the participating Departments, Division, and Board of Directors of the Leiden University Medical Center, and by the Leiden University, Research Profile Area Vascular and Regenerative Medicine. **D.O. Mook-Kanamori.** is supported by Dutch Science Organization (ZonMW-VENI Grant No. 916.14.023). **D. van Heemst.** and **R. Noordam** were supported by a grant of the VELUX Stiftung [grant number 1156]. **T.O. Faquih** was supported by the King Abdullah Scholarship Program and King Faisal Specialist Hospital & Research Center [No. 1012879283]. The Rhineland Study was supported through the "Competence Cluster Nutrition Research" funded by the Federal Ministry of Education and Research (FKZ: 01EA1809C).

Competing Interests
R. Li-Gao is a part-time clinical research consultant for Metabolon, Inc. All other authors have no relevant financial or non-financial interests to declare.

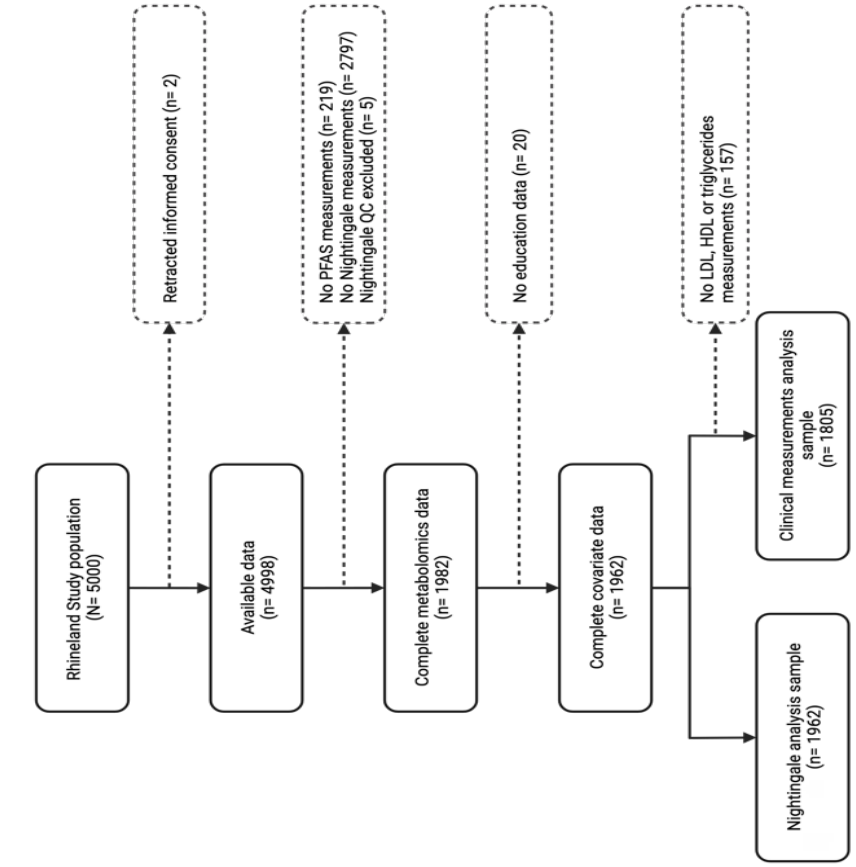
Author Contributions
T.O. Faquih: Conceptualization, Methodology, Formal Analysis, Writing – Original Draft Preparation, Visualization; **E.N. Landstra.:** Conceptualization, Methodology, Formal Analysis, Writing – Original Draft Preparation, Visualization; **R. de Mutsert**: Project Administration, Resources, Funding Acquisition, Writing – Review and Editing; **R. Noordam and D. van Heemst**: Funding acquisition, Writing – Review and Editing; **F.R. Rosendaal.**: Study design, Funding acquisition, Conceptualization; **A. van Hylckama-Vlieg and K.W. van Dijk**: Conceptualization, Supervision, Writing – Reviewing and Editing; **D.O. Mook-Kanamori:** Conceptualization, Methodology, Resources, Writing – Reviewing and Editing, Funding Acquisition, Supervision; **M.M.B. Breteler:** Conceptualization, Resources, Writing – Reviewing and Editing, Data Curation, Funding Acquisition, Supervision.
All authors read and approved the final manuscript

Data availability Rhineland Study data used for this manuscript are not publicly available due to data protection regulations. Access to data can be provided to scientists in accordance with the Rhineland Study's Data Use and Access Policy. Requests for additional information and/or access to the datasets can be send to RS-DUAC@dzne.de.All authors had full access to their respective study data and take responsibility for the integrity of the data and the accuracy of the analysis.

Ethics approval
The NEO study was approved by the medical ethical committee of the Leiden University Medical Centre (LUMC) and all participants gave their written informed consent.

The ethics committee of the medical faculty of the University of Bonn approved the undertaking of the Rhineland Study and it was carried out according to the recommendations of the International Council for Harmonisation Good Clinical Practice standards.

Consent to participate
Written informed consent was acquired from all participants per the Declaration of Helsinki in both the NEO and Rhineland Study.



Tables and Figures

Figure 1: **Overview of study population selection for NEO (A) and the Rhineland Study (B)**

Table 1: **Population characteristics** *of the NEO and Rhineland Study*

Characteristic[†]	NEO Study			Rhineland Study		
	All (n= 586)	***Women (n= 308)***	***Men (n= 278)***	***All (n= 1,962)***	***Women (n= 1,104)***	***Men (n= 858)***
Age (years (SD))	55.8 (5.9)	56.1 (6.2)	55.7 (5.7)	54.6 (14.1)	54.5 (13.7)	54 (14)
Sex (Women (%))	308 (52.6)			1104 (56.3)		
PFOA (mean (SD))	1.07 (0.50)	1.11 (0.51)	1.04 (0.49)	1.11 (0.53)	1.10 (0.62)	1.13 (0.46)
PFOS (mean (SD))	1.08 (0.56)	1.25 (0.58)	0.91 (0.49)	1.21 (0.8)	1.05 (0.91)	1.41 (1.71)
PFHxS (mean (SD))				1.09 (0.69)	0.99 (0.61)	1.22 (0.77)
Education (N (%))						
Low	99.0 (16.9)	42 (15.1)	57 (18.5)	46 (2.3)	38 (3.4)	8 (0.9)
Middle	391 (66.7)	173 (62.2)	218 (70.8)	871 (44.4)	542 (49.1)	329 (38.3)
High	96 (16.4)	63 (22.7)	33 (10.7)	1045 (53.3)	524 (47.5)	521 (60.7)
BMI (kg/m^2 (SD))	25.9 (4.0)	26.6 (3.5)	25.3 (4.4)	25.8 (4.5)	25.3 (4.7)	26.3 (4.2)
LDL (mmol/L (SD))	3.6 (0.9)	3.6 (0.9)	3.6 (1.0)	3.2 (0.9)	3.2 (0.9)	3.3 (0.,9)
HDL (mmol/L (SD))	1.6 (0.5)	1.3 (0.3)	1.8 (0.4)	1.7 (0.5)	1.9 (0.5)	1.4 (0.4)
Cholesterol (mmol/L (SD))	5.7 (1.1)	5.6 (1.0)	5.9 (1.1)	5.2 (1.0)	5.3 (1.0)	5.1 (1.0)
Triglycerides (mmol/L (SD))	1.2 (0.8)	1.4 (0.9)	1.0 (0.6)	1.3 (0.8)	1.1 (0.6)	1.5 (1.0)
Smoking (N (%))	66 (11.3)	31 (11.2)	35 (11.4)	269 (13.7)	134 (12.1)	135 (15.7)
Alcohol (g/day (SD))[‡]	14.0 (15.3)	18.9 (18.3)	9.6 (10.1)	19.4 (25.8)	15.4 (20.4)	24.7 (30.7)

Notes

[†] Number of missing values in the Rhineland Study: BMI (n= 13; 0.7%), LDL (n= 157; 8.0%), HDL (n= 157; 8.0%), Cholesterol (n= 157; 8.0%), Triglycerides (n= 157; 8.0%), Smoking (n= 0; 0%), Alcohol (n= 234; 11.9%)

[‡] In the Rhineland Study, alcohol intake of participants reporting an overall improbable caloric intake (<600 or >8,000) were excluded.

Abbreviations: Standard deviation (SD), perfluorooctanoic acid (PFOA), perfluorooctane sulfonic acid (PFOS), perfluorohexanesulfonic acid (PFHxS), body mass index (BMI), low density lipoprotein (LDL), high density lipoprotein (HDL)

Table 2: Model estimates for the log-transformed clinical lipid measurement in the NEO study

Clinical Lipid Measurement	PFAS	Model 1[†]		Model 2[‡]				Age-stratified analysis[±]			
								Age ≤ 54		Age > 55	
		β *[95%CI]*	*p-value*	β *[95% CI]*	*p-value*	β interaction term *[95% CI]*	*p-value*	β *[95%CI]*	*p-value*	β *[95% CI]*	*p-value*
LDL	PFOA	0.018 [0.007; 0.033]	0.0412	0.292 [0.124; 0.461]	0.0007	-0.005 [-0.008; -0.002]	0.0014				
	PFOS	0.023 [0.004; 0.041]	0.0015	0.325 [0.165; 0.485]	**<0.0001***	-0.005 [-0.008; -0.003]	**0.0002***	0.044 [0.016; 0.071]	0.0020	0.007 [-0.016; 0.031]	0.538
HDL	PFOA	0.016 [0.004; 0.029]	0.0011	0.055 [-0.066; 0.175]	0.3716	-0.001 [-0.003; 0.001]	0.5279				
	PFOS	0.007 [-0.006; 0.021]	0.2624	-0.027 [-0.143; 0.088]	0.6405	0.001 [-0.001; 0.003]	0.5497				
Cholesterol	PFOA	0.02 [0.007; 0.033]	0.0025	0.235 [0.11; 0.359]	**0.0002***	-0.004 [-0.006; -0.002]	0.0007				
	PFOS	0.02 [0.007; 0.034]	0.0037	0.253 [0.134; 0.372]	**<0.0001***	-0.004 [-0.006; -0.002]	**0.0001***	0.036 [0.016; 0.057]	**0.0006***	0.009 [-0.009; 0.027]	0.3141
Triglycerides	PFOA	0.013 [-0.01; 0.036]	0.2780	0.182 [-0.041; 0.404]	0.1088	-0.003 [-0.007; 0.001]	0.1340				
	PFOS	0.016 [-0.009; 0.04]	0.2076	0.298 [0.087; 0.51]	0.0058	-0.005 [-0.009; -0.001]	0.0085				

Notes

[†] Model 1 was adjusted for sex, age and education

[‡] Model 2 was adjusted for sex, age and education and additionally included an age-interaction term

[±] The age-stratified analysis was only performed if the age-interaction term was significant and was adjusted for sex, age and education

***** Associations with the significant p-value after multiple test correction (p-value < 0.0007)

Abbreviations: Beta estimate (β), confidence interval (CI), perfluorooctanoic acid (PFOA), perfluorooctane sulfonic acid (PFOS), low density lipoprotein (LDL), high density lipoprotein (HDL)

Table 3: Model estimates for the log-transformed clinical lipids in the Rhineland Study

Clinical Lipid Measurement	PFAS	Model 1†		Model 2‡				Age-stratified analysis±			
								Age ≤ 54		Age > 55	
		β *[95%CI]*	*p-value*	β *[95% CI]*	*p-value*	β interaction term *[95% CI]*	*p-value*	β *[95%CI]*	*p-value*	β *[95% CI]*	*p-value*
LDL	PFOA	0.005 [-0.006; 0.016]	0.388	0.085 [0.038; 0.132]	**0.0004***	-0.001 [-0.002; -0.001]	**0.0005***	-0.006 [-0.024; 0.013]	0.551	0.008 [-0.005; 0.021]	0.2249
	PFOS	0.007 [-0.003; 0.017]	0.1726	0.166 [0.097; 0.235]	**<0.0001***	-0.002 [-0.003; -0.001]	**<0.0001***	0.016 [-0.019; 0.051]	0.3571	0.008 [-0.002; 0.018]	0.1127
	PFHxS	0.013 [0.003; 0.023]	0.0128	0.080 [0.037; 0.123]	**0.0002***	-0.001 [-0.002; -0.000]	0.0016				
HDL	PFOA	0.007 [-0.001; 0.016]	0.0690	0.001 [-0.034; 0.035]	0.9610	0.000 [-0.000; 0.001]	0.6956				
	PFOS	0.005 [-0.002; 0.013]	0.1575	0.002 [-0.049; 0.053]	0.9310	-0.000 [-0.001; 0.001]	0.9042				
	PFHxS	0.012 [0.004; 0.019]	0.0024	0.035 [0.003; 0.067]	0.0302	-0.000 [-0.001; 0.000]	0.1347				
Cholesterol	PFOA	0.008 [-0.001; 0.016]	0.0665	0.056 [0.020; 0.092]	0.0021	-0.001 [-0.001; -0.000]	0.0066				
	PFOS	0.007 [-0.001; 0.014]	0.0837	0.117 [0.064; 0.169]	**<0.0001***	-0.002 [-0.002; -0.001]	**<0.0001***	0.011 [-0.016; 0.037]	0.4415	0.008 [0.000; 0.016]	0.0403
	PFHxS	0.012 [0.004; 0.020]	0.0023	0.079 [0.046; 0.112]	**<0.0001***	-0.001 [-0.002; -0.001]	**<0.0001***	0.008 [-0.005; 0.021]	0.2539	0.005 [-0.004; 0.014]	0.2956
Triglycerides	PFOA	-0.001 [-0.015; 0.014]	0.9362	0.035 [-0.025; 0.095]	0.2590	-0.001 [-0.002; 0.000]	0.2376				
	PFOS	-0.005 [-0.018; 0.008]	0.4502	0.025 [-0.063; 0.114]	0.5830	-0.000 [-0.002; 0.001]	0.5054				
	PFHxS	-0.006 [-0.019; 0.007]	0.3969	0.048 [-0.008; 0.103]	0.0906	-0.001 [-0.002; 0.000]	0.0517				

Notes

† Model 1 was adjusted for sex, age and education

‡ Model 2 was adjusted for sex, age and education and additionally included an age-interaction term

± The age-stratified analysis was only performed if the age-interaction term was significant and was adjusted for sex, age and education

* Associations with the significant p-value after multiple test correction (p-value < 0.0007)

Abbreviations: Beta estimate (β), confidence interval (CI), perfluorooctanoic acid (PFOA), perfluorooctane sulfonic acid (PFOS), perfluorohexanesulfonic acid (PFHxS), low density lipoprotein (LDL), high density lipoprotein (HDL)

Table 4: Number of significant associations (p-value < 0.0007) per metabolite category with each PFAS in the NEO study. The range of the effect estimates are contained within the brackets.

NEO

	Model 1: Outcome ~ PFAS + Age + Sex + Education		Model 2: model 1 + PFAS*age		
	Number of positive associations (min estimate; max estimate)	Number of negative associations (min estimate; max estimate)	Number of positive associations (min estimate; max estimate)	Number of negative associations (min estimate; max estimate)	Number of age interactions (min estimate; max estimate)
PFOA					
Clinical lipid measures (n=4)	-	-	-	-	-
Amino acids (n=3)	-	-	-	-	-
Aromatic amino acids (n=2)	-	-	-	-	-
Branched-chain amino acids (n=3)	-	-	-	-	-
Apolipoproteins (n=3)	-	-	-	-	-
Cholesterol (n=9)	-	-	1 (1.410; 1.410)	-	1 (-0.023; -0.023)
Fatty acids (n=16)	-	-	-	-	-
Fluid balance (n=2)	-	-	-	-	-
Glycerides and phospholipids (n=9)	-	-	-	-	-
Glycolysis related metabolites (n=3)	-	-	-	-	-
Inflammation (n=1)	-	-	-	-	-
Ketone bodies (n=2)	-	-	-	-	-
Lipoprotein concentrations (n=14)	1 (0.150; 0.150)	-	-	-	-
Lipoprotein particle size (n=3)	-	-	-	-	-
HDL composition (n=44)	2 (0.145; 0.150)	-	-	-	-
LDL composition (n=33)	-	-	4 (1.403; 1.500)	-	4 (-0.024; -0.025)
VLDL composition (n=66)	-	-	-	-	-
IDL composition (n=11)	-	-	-	-	-
PFOS					
Clinical lipid measures (n=4)	-	-	2 (1.518; 1.575)	-	2 (-0.025; -0.025)
Amino acids (n=3)	-	-	-	-	-
Aromatic amino acids (n=2)	-	-	-	-	-
Branched-chain amino acids (n=3)	-	-	-	-	-
Apolipoproteins (n=3)	-	-	2 (1.392; 1.702)	-	2 (-0.028; -0.023)
Cholesterol (n=9)	-	-	5 (1.412; 1.634)	-	5(-0.027; -0.024)
Fatty acids (n=16)	3 (0.157; 0.185)	-	3 (1.507; 1.663)	-	3 (-0.026; -0.028)
Fluid balance (n=2)	-	-	-	-	-
Glycerides and phospholipids (n=9)	-	-	-	-	-
Glycolysis related metabolites (n=3)	-	-	-	-	-
Inflammation (n=1)	-	-	-	-	-
Ketone bodies (n=2)	-	-	-	-	-
Lipoprotein concentrations (n=14)	-	-	5 (1.408; 1.552)	-	5 (-0.026; -0.023)
Lipoprotein particle size (n=3)	-	-	-	-	-
HDL composition (n=44)	-	-	-	-	-
LDL composition (n=33)	-	-	20 (1.504; 1.903)	5 (-1.774; -1.456)	25 (-0.031; 0.029)
VLDL composition (n=66)	-	-	7 (1.298; 1.504)	-	7 (-0.025; -0.022)
IDL composition (n=11)	-	-	5 (1.407; 1.527)	1 (-1.313)	6 (-0.025; 0.023)

Table 5: Number of significant associations (p-value < 0.0007) per metabolite category with each PFAS in the Rhineland study. The range of the effect estimates are contained within the brackets.

Rhineland Study

	Model 1: Outcome ~ PFAS + Age + Sex + Education		Model 2: model 1 + PFAS*age		
	Number of positive associations N (min estimate; max estimate)	*Number of negative associations N (min estimate; max estimate)*	*Number of positive associations N (min estimate; max estimate)*	*Number of negative associations N (min estimate; max estimate)*	*Number of age interactions N (min estimate; max estimate)*
PFOA					
Clinical lipid measures (n=4)	-	-	1 (0.402; 0.402)	-	1 (-0.006; -0.006)
Amino acids (n=3)	-	-	-	-	-
Aromatic amino acids (n=2)	-	-	-	-	-
Branched-chain amino acids (n=3)	-	-	-	-	-
Apolipoproteins (n=3)	-	-	-	-	-
Cholesterol (n=9)	1 (0.080; 0.080)	-	-	-	-
Fatty acids (n=16)	-	-	-	-	-
Fluid balance (n=2)	1 (0.135; 0.135)	-	-	-	-
Glycerides and phospholipids (n=9)	1 (0.081; 0.081)	-	-	-	-
Glycolysis related metabolites (n=3)	-	-	-	-	-
Inflammation (n=1)	-	-	-	-	-
Ketone bodies (n=2)	-	-	-	-	-
Lipoprotein concentrations (n=14)	-	-	-	-	-
Lipoprotein particle size (n=3)	-	-	-	1 (-0.382; -0.382)	1 (0.006; 0.006)
HDL composition (n=44)	-	-	-	-	-
LDL composition (n=33)	-	-	-	-	-
VLDL composition (n=66)	-	-	-	-	-
IDL composition (n=11)	-	-	-	-	-
PFOS					
Clinical lipid measures (n=4)	-	-	2 (0.708; 0.781)	-	2 (-0.010; -0.009)
Amino acids (n=3)	-	-	-	-	-
Aromatic amino acids (n=2)	-	-	-	-	-
Branched-chain amino acids (n=3)	-	-	1 (0.761; 0.761)	-	1 (-0.010; -0.010)
Apolipoproteins (n=3)	-	-	1 (0.700; 0.700)	-	1 (-0.002; -0.002)
Cholesterol (n=9)	-	-	6 (0.573; 0.856)	-	6 (-0.011; -0.007)
Fatty acids (n=16)	-	-	9 (0.563; 1.025)	-	9 (-0.013; -0.008)
Fluid balance (n=2)	-	-	1 (0.620; 0.620)	-	1 (-0.009; -0.009)
Glycerides and phospholipids (n=9)	-	-	4 (0.558; 0.725)	-	4 (-0.009; -0.007)
Glycolysis related metabolites (n=3)	-	-	-	-	-
Inflammation (n=1)	-	-	-	-	-
Ketone bodies (n=2)	-	-	-	-	-
Lipoprotein concentrations (n=14)	-	-	5 (0.688; 0.838)	-	5 (-0.011; -0.009)
Lipoprotein particle size (n=3)	-	-	-	-	-
HDL composition (n=44)	-	-	2 (0.568; 0.580)	1 (-0.542)	3 (-0.008; 0.008)
LDL composition (n=33)	-	-	21 (0.637; 0.860)	7 (-0.603; -0.759)	28 (-0.011; 0.010)
VLDL composition (n=66)	-	-	12 (0.577; 0.808)	2 (-0.674; -0.716)	14 (-0.010; 0.009)
IDL composition (n=11)	-	-	7 (0.619; 0.847)	1 (-0.616; -0.616)	8 (-0.011; 0.008)
PFHxS					
Clinical lipid measures (n=4)	-	-	1 (0,480; 0.480)	-	1 (-0.007; -0.007)
Amino acids (n=3)	-	-	-	-	-
Aromatic amino acids (n=2)	-	1 (-0.105; -0.105)	-	-	-
Branched-chain amino acids (n=3)	-	-	-	-	-
Apolipoproteins (n=3)	1 (0.088; 0.088)	-	-	-	-
Cholesterol (n=9)	6 (0.073; 0.094)	-	3 (0.372; 0.416)	-	3 (-0.006; -0.005)
Fatty acids (n=16)	9 (0.078; 0.095)	-	3 (0.336; 0.401)	-	3 (-0.006; -0.005)
Fluid balance (n=2)	1 (0.092; 0.092)	-	-	-	-
Glycerides and phospholipids (n=9)	2 (0.077; 0.078)	-	2 (0.397; 0.410)	-	2 (-0.006; -0.006)
Glycolysis related metabolites (n=3)	-	-	-	-	-
Inflammation (n=1)	-	-	-	-	-
Ketone bodies (n=2)	-	-	-	-	-
Lipoprotein concentrations (n=14)	2 (0.081; 0.086)	-	-	-	-
Lipoprotein particle size (n=3)	-	-	-	-	-
HDL composition (n=44)	1 (0.076; 0.076)	-	-	-	-
LDL composition (n=33)	11 (0.081; 0.089)	-	-	-	-
VLDL composition (n=66)	4 (0.076; 0.103)	-	-	-	-
IDL composition (n=11)	2 (0.082; 0.089)	-	-	-	-

Chapter 8
Discussion and future perspectives

1 AIMS OF THIS THESIS

In this thesis, we explored epidemiological applications and methodological challenges of genomics, proteomics, and metabolomics. Metabolomics, one of the more recent OMICs fields, was the main focus of our research. Metabolites are thought to reflect integrated genomic and proteomic influences in metabolism as well as the environmental and external effects from the individuals' exposures. Hence, metabolomics may provide novel insight in the pathophysiology of complex multifactorial diseases. Indeed, in our research, metabolomics shed light on the associations between metabolites and non-alcoholic fatty liver disease (NAFLD), metabolites and short sequence repeats in the huntingtin gene, and the associations of per-/polyfluoroalkyl substances (PFAS) chemicals in the general population with metabolites. However, methodological, technological, and statistical challenges remain in this infant field. Therefore, we explored the issues of handling missing values in metabolomic data using a simulation study based on real data. This resulted in a publicly available R script to streamline the imputation of missing values of metabolites. In addition, we showed the importance of examining the agreement between clinical measurements with protein measurements from high throughput platforms. This thesis provides tools for and perspectives of the current status and future directions of OMICs research in epidemiology.

2 MEASUREMENT METHODOLOGY AND HIGH DIMENSIONALITY

Platforms such as SOMAscan and Metabolon provide quantification of more than a 1000 proteins and metabolites, respectively. However, these platforms come with their own specific shortcomings. First, the specificity and sensitivity of the metabolite or protein measurement differs depending on the chemical properties and nature of the biomolecule. For example, binding affinity of SOMAscan's aptamers are reportedly poor with proteins with a neutral charge or those with large sizes. Second, comparing and validating measurements with "golden" standard methods have been, at least partially, neglected as is evident from **chapter 2**. The appealing prospect of measuring a large number of metabolites and proteins should not detract from validating those measurements. Third, data should be double checked for post-processing errors during the annotation of the metabolites or proteins. The possibility of human errors and software errors occurring during this process is frequently ignored. Internal validations and simulations should be an essential element of the data integration and analysis process. In addition, improvement in measurement technology and methodology are needed to enable consistent measurements of complex metabolites or proteins. Furthermore, the number of detectable metabolites and proteins is projected to increase substantially. Therefore, in addition to validating the data using golden standard methods, researchers should validate findings use their own knowledge of the chemical properties and biochemical pathways of the metabolites and proteins related to their research question. Moreover, they must keep in mind the characteristics of the population of interest used in the study, since these may confound or affect the OMICs measurements. In conclusion, researchers must be aware of the strengths and weaknesses of the selected OMICs platforms used to produce their data and consider these when interpreting the results.

3 STATISTICAL CHALLENGES AND SOLUTIONS

3.1 The N<P problem

The merit of large-scale data from high throughput OMICs acts as a double-edged sword during the statistical analysis due to the multiple testing problem. For example, if 1000 comparisons are made between metabolites and a trait, the number of false positive results given a P-value threshold of 0.05 is 50 associations. This issue is commonly addressed by reducing the P-value threshold. For example, the Bonferroni method divides the nominal P-value by the number of independent measurements. Thus, studies must have an appropriately large sample size to assess associations between traits and a large number of measurements. Ideally the number of individuals in a study should be larger than the number biochemical variables. Otherwise, the analysis would suffer from N<P problem, wherein the number of individuals (N) in the study is less than the number of predictors (P). N<P leads to bias in the analysis results as well as reduced reliability and reproducibility of the results. In the case of epidemiological studies, it is essential to have a sufficient number of individuals per outcome event. Otherwise, etiological results may lack the necessary power to confidently report any findings and prediction models may be poor, overfitted, and not generalizable (1). Therefore, sample size considerations are crucial when conducting epidemiological studies using large OMICs data. The number of cohorts with OMICs measurements across the globe has increased greatly in recent years. To name a few notable examples, Nightingale measurements are available in the UK biobank cohort (n ~ 500,000) (2) and Metabolon is planned for the Million Veteran Program (n ~ 900,000) (3). Several consortia focusing on OMICs have also been established such as the biomolecular resources and research infrastructure (BBMRI) consortium (4, 5), BBMRI-ERIC (6), and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium (7, 8). These large studies and collaborative efforts provide several advantages to overcome the limitations of large OMICs data. They provide a solution for the N<P issues by combining their data and results to achieve larger power, similar to the work in **chapter 6**. Collaborative studies provide an additional benefit by enabling the reproduction and validation of findings in different populations. Meta-analyses of genome-wide association studies at the beginning of this century were in that sense pioneering, as they performed rigorous replication and validation studies. Lack of reproducibility in scientific research remains a persistent issue and OMICs research is no different. Thus, validation of prediction models and performing meta-analyses, increases the confidence in the research findings far more than single studies (9) as demonstrated in **chapter 7**. In addition, even significant findings in a specific population or ethnicity may not be generalizable to different populations and ethnicities. Hence, OMICs research should aim for the identification, comprehension, and dissolution of these disparities. To summarize, sample size considerations are critical for OMICs studies and collaborative research is crucial to not only overcome the N>P limitations but also to improve the quality and reproducibility of the findings.

3.2 Imputation of Missing Data

As we have shown in **chapter 3**, missing values are common and remain an issue in OMICs data. Complete case analysis, in which analysis is conducted on only individuals without any missing values, can be an appropriate method but also has its disadvantages (10). As discussed earlier, maximizing the sample size for analysis in OMICs is crucial. Thus, imputing missing values is an important method to reach full sample size. However, handling missing values in metabolomic studies requires careful consideration (11) as popular methods tend to be inadequate as they may lead to biased results (12, 13). This is the case for a common method for dealing with missing values in which metabolites with missingness above a certain percentage are excluded.

Alternatively, metabolites with missing values are imputed with a single numerical value, such as the mean or half the minimum value for each respective metabolite. However, this removes valuable data, and does not consider the various reasons for missingness in metabolomic data or the different characteristics of the measured metabolite groups (12, 13). For example, we recommend in **chapter 3** that xenobiotics should be imputed with zero as they are likely truly missing. For other metabolite groups, multiple imputation and k-nearest neighbor imputation demonstrated better performance with reduced bias in the analysis (12, 13). These methods are not perfect, and their performance dwindles if the sample sizes are small with high level of missingness. Overall, all these factors must be considered for metabolomics studies.

Another issue with handling missing values in metabolomic studies is that metabolomics research lacks a uniform guideline for imputation methodologies. Thus, selecting and applying appropriate imputation methods can be challenging. Multiple imputation may be most appropriate because of reduced potential bias but are thought to be computationally and statistically challenging (13). However, recent developments have overcome many of the issues associated with multiple imputation methods. Programing packages and plugins have become readily available with full tutorials on their usage, making it easier for researchers to apply. These include resources tailored for imputing missing values in metabolomic data (for example the script presented in **chapter 3**). These software tools coupled with the rapid evolution of computer processors and hardware have facilitated the application of imputation methods on large data. Further recommendations to reduce the complexity and computational time of multiple imputation methods is selecting a reasonable number of imputations for the study at hand. For example, 5 to 10 imputed sets can be sufficient in the presence of a moderate degree of missingness (14). Another recommendation is the selection of a small number of biologically relevant “auxiliary” variables (i.e., age, sex, body mass index, etc.) to impute the missing values for the metabolites instead of using the full dataset (13).

Moving forward, guidelines—such as the Metabolomics Standards Initiative (15)—should be expanded upon to provide a uniform primer for imputing missing data using optimized methodologies for metabolomics. Achieving this would require the collaboration and agreement of the metabolomics research community and is definitely something worth pursuing.

4 REPRESENTING AND INTERPRETING METABOLOMICS AND PROTEOMICS RESULTS

The amount of data measured by OMICs platforms raises another question: how do we represent and interpret the vast amount of data from the analysis? Typical results of epidemiological studies are provided in tables and static figures within the main body of the manuscript. However, when reporting large numbers of tests for every measured metabolite, protein, or genetic variation, it becomes challenging to include all the results in a single table or figure. Moreover, scientific journals typically limit the number of tables and figures allowed in the main text. Providing these results as supplementary materials creates dozens of large, sometimes overwhelming tables and figures, which are often overlooked. This leads to selecting and reporting a handful of metabolites, proteins, or single nucleotide polymorphisms (SNPs) in the main body of scientific papers. Moreover, OMICs measurements, especially metabolites and proteins, are highly correlated and interconnected by common chemical pathways. Therefore, reporting on a select few compounds fails to capture the true depth of their biological implications, which subsequently can lead to bias in reporting.

Reporting of OMICs data can be improved in a number of ways. First, the full analysis results can be published on an online website, such as an accessible online database or interactive webpage viewed directly from an internet browser. This enables easy viewing and interaction with large tables and complex figures. In addition, a website dedicated to the results enables the inclusion of all the statistical estimates from the analysis. Examples of these sites already exist, such as the Metabolomics GWAS Server (16, 17), the Metabolome website for metabolites correlated with age (18, 19), and our website for the results regarding metabolites associated with hepatic triglyceride content, described in **chapter 5**. Other alternatives are central web databases and online analysis suites for uploading OMICs data and performing additional analyses. Examples of these resources include MetaboLights (20) and OpenGWAS (21). Second, analytical methods should be available that capture the extent of correlations and association within OMICs data. As demonstrated in **chapter 5**, methods such as gaussian graphical modeling (GGM) and genome-scale metabolic models (GSMM) provide unique insight into pathways and reactions connecting the OMICs data and the analysis results (22). Third, dynamic web figures and plots can be generated to be interactive instead of static images. This facilitates the navigation and interaction with complex results while providing the ability for both wide and pinpoint visualization.

With software and programming advancements, the tools to enable the production of such interactive figures and websites have become available for use by researchers. In addition, some tools have been designed specifically for the online representation of OMICs data. One such tool is PheWeb for the visualization and exploration of genomic study results (23). Moreover, interactive GGM and GSMM networks can be produced by common and free software. Thus, it is now feasible for researchers to create interactive figures and networks that allow full and thorough exploration and investigation of OMICs studies.

5 XENOBIOTICS, A KEY FOR EXPOSOME RESEARCH?

Exposome, the study of the effects of external environmental and lifestyle exposures on individuals' health, has gained steady momentum in recent years. Indeed, large consortia have been founded dedicated to exposome research, such as the exposome-NL consortium (24) in the Netherlands. However, one persistent limitation of these studies is the poor availability of real-life exposome measurements for the general population. Metabolomics can be a solution for this issue. Indeed, some interesting exposome variables can be detected in the xenobiotic metabolite measurements found in metabolomic platforms such as Metabolon. For example, some environmental contaminants such as pollutants in the air, soil, or water may be measured in body fluids, such as the forever chemicals (PFAS) exposure levels as demonstrated in **chapter 7**. In addition, xenobiotic metabolites such as cotinine metabolites reflect smoke exposure and other lifestyle habits. Besides environmental exposures, xenobiotic metabolites may reflect components from diet, cosmetics, and medications. Diet related metabolites can be used in unique ways in epidemiological studies. For example, some metabolites have been linked with dietary patterns such as the intake of fish and bread. Several studies have explored such biomarkers as quantitative information for dietary intake and the validation of food frequency questionnaires (FFQs) (25, 26). These FFQs can be affected by participants misremembering what they ate over short or long periods of time. This issue is called differential measurement error bias and is commonly referred to as "recall bias" (27). Therefore, quantitative diet information via metabolite measurements can address recall bias and complement the FFQs. However, achieving this goal remains a challenge that requires the identification and validation of strong and robust biomarkers for a wide number diets and food sources (26). The same principle of validating patient questionnaires using metabolites can potentially

be applied to verify other lifestyle factors such as the aforementioned tobacco smoking via cotinine metabolites (28).

Xenobiotic metabolites from medications can also provide unique epidemiological insights. Common ways to obtain medication data is from prescription and dispensing data from hospitals or pharmacies, or from the beforementioned self-reported patient questionnaires. However, this data could be scarce for the general population or not accessible for research purposes. In addition, some drugs can be obtained over the counter without a prescription and do not leave patient specific data traces. In these cases, metabolomic measurements can be useful if the medication related xenobiotic metabolites can be quantified. Nonsteroidal anti-inflammatory drugs (NSAIDs) are an example of an over-the-counter drug class that does not require prescription and are difficult to trace. This is an issue when attempting to study the negative health impact of NSAIDs overuse. Indeed, NSAIDs have been found to be associated with increased risk of heart failure (29)—particularly in hypertensive patients (30)—and increased risk of gastrointestinal bleeding (31, 32). The same principle can also be coupled with randomized controlled trials (RCTs) to collect metabolomic quantitative data regarding the patient's medication compliance. Therefore, using xenobiotic metabolites for NSAIDs and other medications can provide quantitative data of their usage in the general population. In turn, these xenobiotic measurements can be used to address various epidemiological questions such as improving the estimation of the population use of NSAIDs—or other medications—and their association with negative health outcomes or mortality.

One limitation that should be noted for the use of metabolomics to trace and quantify medication related metabolites is the half-life of metabolites in the human samples, i.e., the period of time it takes a substance (metabolite) level to decrease to half its initial concentration (33). A metabolite with a short half-life is eliminated from the body too quickly and can be difficult to measure. One possible work around for this issue is to obtain and measure multiple samples to capture metabolite levels at different time points.

As demonstrated by these examples, the quantification of xenobiotic metabolites related to environmental contamination, diet, or medication could provide tools to answer exposome related epidemiological questions. Hence, metabolomics may be a key for the expansion of exposome research.

6 THE VALUE OF CROSSING AND INTEGRATING MULTIPLE OMICS

The human biological system is incredibly complex and sophisticated. Thanks to technological advancements, OMICs have enabled the comprehensive study of different layers of this human system. OMICs studies usually focus on a single layer; however, these layers are interconnected and actively interacting. Although the genetic sequence is largely static and conserved after conception, massive diversity in gene expression occurs from epigenetic regulation. Moreover, gene expression differs in various body tissues and the degree of expression differs over time. As a consequence, the diversity and levels of proteins and metabolites in different tissues also differ over time.

This level complexity is disregarded when focusing on a single OMICs layer. Indeed, despite the important findings from genome wide association studies, examining DNA sequence data and SNPs alone is currently incapable of explain the full heritability of diseases, referred to as the missing heritability problem (34). Likewise, focusing on OMICs data such as metabolomics alone ignores genetic factors that may affect metabolism and metabolite levels.

The integration of two or more OMICs, i.e., multi-OMICs, can be a powerful approach that combines and reinforces the unique features of separate OMICs data. Indeed, multi-OMICs can expand on etiological findings by providing better understanding of disease pathophysiology. For example, the integration of genomic and metabolomic data as presented in **chapter 6,** has also been used for NAFLD research (35), and to create an atlas of genetic influences on blood metabolites (16). Another valuable addition to a multi-OMICs study is the use of transcriptomics data. Transcriptomics can be further applied to assess the dynamic expression of the genome in different cell lines in the body. Crossing over the results from genomics and transcriptomics with metabolomics or proteomics can form a network pinpointing the location and time points for gene expression and link it to the levels of proteins and metabolites. Beyond these endogenous layers, the beforementioned external xenobiotics and exposome can also be added to this network. Thus, the possible associations of environmental and lifestyle exposures with the endogenous factors can be considered. In this sense, crossing OMICs not only combines their unique features, but also compensates their individual downsides and reinforces their strengths. Therefore, integrating multi-OMICs data is an important up and coming field for understanding complex pathophysiology of human phenotypes and diseases.

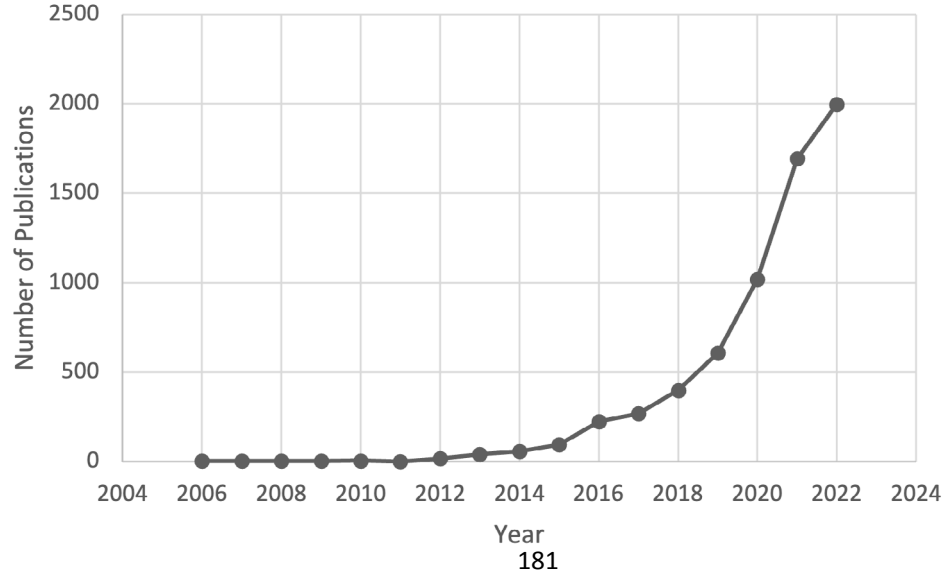
Crossing over multi-OMICs can also provide insights into causal associations. One way to achieve this is by using genomics in combination with one or more OMICs data. Since genetic variations are conserved since conception and are not affected by confounders, it is possible to use genomics to infer causality, using statistical methodologies such as Mendelian Randomization (MR) (36). On the other hand, metabolomics and proteomics are not readily applicable to assess causality in associations with diseases. However, genomics combined with other single OMICs can possibly enable inferring causality between the different single OMICs and outcomes of interest. In an MR analysis SNPs can be selected that are associated with a specific outcome and relevant metabolites or proteins biomarkers. These can then be used to estimate the causal effect of the biomarkers on the outcome. Thus, providing a method to avoid confounding if the assumptions required for causal inference in MR are maintained. A hypothetical example would be for venous thrombosis research. The search for biomarkers which are associated or predictive for VTE remains challenging (37, 38). Multi-OMICs can be utilized by using known and unknown genomic associations with VTE and its associated metabolites and proteins. Analysis of these OMICs could potentially assist in identifying novel biomarkers or causal associations related to VTE (by way of MR or other causal inference methodologies). To date, several multi-OMICs MR studies have been used to identify potential drug targets (39), inferred causality between the metabolome and obesity (40), and identified causal links between carnitine and systolic blood pressure (41).

7 MULTI-OMICS AND THE (NOT SO) LONG AND WINDING ROAD

We have described the merits of multi-OMICs and provided examples which demonstrate their potential, such as creating complex OMICs atlases for deeper etiological understanding of diseases (16, 42, 43) and in identifying causal associations (39-42). But what are the challenges facing multi-OMICs? What is needed for it to grow? How far are we in this upcoming field? For starters, the big challenge of multi-OMICs data is the sheer amount of data to combine and study. One way to address this challenge is to use the aforementioned analytical methods to construct a multi-OMICs network for the relevant biological pathways related to the outcome of interest. This selection can be done using statistical methods or prior findings from epidemiological and single OMICs studies.

Multi-OMICs studies are currently being conducted more frequently (Figure 1). However, for multi-OMICs to blossom, it would first require further advancements in the separate OMICs fields. These advancements must aim to improve the reliability and reproducibility of measurements whilst reducing costs. For genomics, we reiterate the importance of studying copy number variations, structural variations, and other variations in addition to SNPs. This may help to address the missing heritability problem and may shed light on new mechanisms underlying pathophysiology of diseases. In addition, new and efficient methods are needed to integrate and analyze the complex OMICs data. Open-source software and plugins for statistical software and programming languages, such as R and Python, can streamline these methodologies and simplify their use for multi-OMICs research. The subsequent results and findings would greatly benefit from dynamic and interactive representation to enable the full exploration by other researchers. Indeed, network analysis and interactive web tools can display multi-OMICs data and their associations efficiently. In addition, new analysis methods have been explored to process and analyze multi-OMICs data. One such method is machine learning (ML). ML techniques are increasingly applied in epidemiological and OMICs research. ML is a powerful technique for the analysis of large data and can be used for prediction modeling and even causal effect estimation (44). However, ML may suffer from several pitfalls. For one, the parameters and steps silently undertaken by the ML algorithm, to automatically learn from the data, may be nontransparent and complicated (45, 46). Moreover, ML inherently uses some level of randomness which may affect its performance and specific outcome. Changing the randomness parameter (also known as the random seed) may lead to inflated estimates of the model (45, 47) Another aspect is that ML methodologies can also be affected from the same issues as non-ML methods, such confounding, overfitting, and bias (45, 47). These aspects can be challenging when verifying and reproducing the results from ML models (45, 47). In addition, some studies reported that, for clinical prediction models, ML showed no performance benefits when compared to logistical regression (48). These problems have been addressed by combining epidemiological principles and statistical frameworks with ML and by selecting appropriate ML algorithms and properly applying them (44-46). Indeed, efforts are underway to provide guidelines for to address the aforementioned issues and help guide the development and reporting of ML prediction models (46). These aspects and solutions will be applicable for multi-OMICs as well and can enable the use of robust ML methods in OMICs and multi-OMICs research.

Figure 1: Search results of publications using the term "multi-OMICs" in PubMed from 2006 to November 2022 (49)



Ultimately, the potential of multi-OMICs research is promising and the number of studies incorporating multiple OMICs is rapidly increasing. Furthermore, statistical methodologies, such as ML, are improving as well, further enabling multi-OMICs analyses. Overall, the challenges facing multi-OMICs are being addressed and will likely be steadily resolved. The road to prevalent multi-OMICs epidemiological research appears shorter than ever before.

8 OMICS, CLINICAL CARE, AND BEYOND

OMICs and multi-OMICs are powerful tools for dissecting the pathophysiology and etiology of diseases. How can these findings be incorporated into clinical care? Physicians and clinicians rely on their medical knowledge and experience for prognosis, diagnosis, and treatment of patient's medical conditions. Based on this knowledge, they may request a screening for specific biomarkers with a strong indication for a specific outcome to verify their diagnosis. For example, c-reactive protein (CRP) is a strong biomarker for inflammation and used as a diagnostic marker of infection and inflammation (50). D-dimer, small protein fragments of fibrin, is a strong diagnostic biomarker for venous thrombosis events (51). D-dimer has a high sensitivity for the identification VTE patients but is noted to having low specificity to identify patients without VTE (52). Prohormone brain natriuretic peptide (pro-BNP) is an example of a peptide that has become regularly measured for the diagnosis, particularly in the emergency room, and prognosis of heart failure (53-55). Physicians also rely on the patient's personal and family history to order screening of specific disease biomarkers. For example, screening for the BRCA1 and BRCA2 genes has become common procedure to identify the risk of breast cancer and ovarian cancer (56). Another example is prenatal and newborn genetic screening. These tests have become routine clinical procedures to identify possible treatable but severe diseases that require swift and early action for new born babies (57). Newborn genetic screenings checks for diseases such as thyroid disorder, blood disorders, a range of metabolic disorders, and several other diseases (58). These examples show how some OMICs findings, especially genomics, have reached and enhanced some areas of routine clinical care (59, 60). In order to be implemented in clinical practice, identified biomarkers for diseases must show a robust and specific association with a disease outcome, show a strong utility for prognosis and diagnosis of an outcome, and should be readily available and easily measured. Subsequently, this may lead to the incorporation of the biomarker measurement to routine clinical application, as what was done for pro-BNP. However, challenges still remain, and progress is slower than expected for the integration of OMICs biomarkers in the clinic. As a consequence, the integration of these biomarkers into clinical use has been slow. For example despite strong evidence linking genetic variation with increased disease risk, it has taken a 25-30 year period between discovery and clinical implementation of genetic markers such BRCA1 (61) and nearly 10 years for incorporation and standardization of pro-BNP measurement in clinical care (62).

Genetic screening for patients remains expensive and time consuming. This adds another layer of complication due to limited financial coverage of these tests by health insurance companies. This financial burden discourages physicians and patients to request such tests. In addition, results for the analysis are often not easily readable by the clinician or the risk probability from the tests are not sufficient to determine a treatment course. "Why request an expensive and time-consuming test that does not add value to the standard treatment plan? Is OMICs research truly useful for clinical care?" These concerns and issues must be addressed when thinking of the future value of OMICs to clinical setting. Metabolomics and proteomics research must learn from the lesson of genomics and other successful biomarkers (such as pro-BNP) to prove their value and to expedite their integration in the clinical setting. We will discuss three potential points that can aid OMICs to reach clinical care.

First, the measurements from OMICs platform can provide better clinical relevance if they report the absolute concentration of the biomarkers instead of relative values. One of the tradeoffs of most untargeted methodologies is between the ability to measure large amounts of biomarkers and obtaining absolute quantification. One of the reasons for this is the nature of the methods and platforms used to a attain those measurements (i.e., mass spectrometry and aptamer-based methods). The large number of measured biomarkers and relative concentrations from these platforms are indeed useful and beneficial to address research questions. However, for clinical use these relative measurements can be difficult to interpret when deciding a treatment plan. This is particularly true when compared to standardized and routine clinical biomarker measurements. Of course, attaining absolute quantification is possible from some OMICs instruments and platforms, such as the Nightingale platform. However, moving forward, OMICs should aim to provide absolute concentrations on large scales. For existing absolute quantification platforms this would require expanding the range of measurable biomarkers. For relative quantification platforms this requires finding solutions such as converting relative values to absolute concentrations by including reference samples or implementing other methodologies. Alternatively, future technologies can lead the way for novel large scale untargeted absolute quantification platforms. Achieving these goals will ultimately aid in translating the findings from OMICs research into clinical relevancy.

Second, as discussed earlier, identification of causal associations from single or multi-OMICs research could be the key to finding novel and important disease biomarkers. This can be further expanded to identify potential protein and metabolite drug targets. Previous studies have successfully used MR to identify ACE2 and IFNAR2 proteins as potential drug targets for severe COVID-19 patients (63). This was in line with findings from a large randomized controlled trial (RCTs) (64). This study also showcases other benefits of MR and OMICs studies. Compared with an RCT, an MR study is much cheaper, less time consuming, and faces fewer ethical implications. In addition, many RCTs fail at the crucial phase 3 stage (65), thus wasting time, money, and potentially negatively affecting the wellbeing of patients included in the RCT. One of the reasons for this high rate of failure is designing drugs for a target without sufficient causal evidence (65, 66). An example of this is an MR study that concluded, despite previous expectations, that CETP and CETP inhibition showed no causal association with reducing cardiovascular disease risk and therefore was not a suitable drug target for prevention of CVD (67). This illustrates how failed RCTs, like the ones who targeted CETP, could have been prevented if genetic and OMICs findings were initially applied. Indeed, OMICs and multi-OMICs evidence could complement epidemiological studies and provide insight for the design of RCTs (68-70). Thus, researching this combination would be beneficial as it can potentially increase the probability of success of RCTs, while potentially reducing financial cost and reducing patient burden.

Third, biomarkers are rare, which in isolation provide strong evidence for the prognosis and diagnosis of a disease or health outcome. An alternative method may be to combine several biomarkers to generate a single score to diagnose or predict a disease. In genomics, this already has become commonplace by means of polygenic risk scores (71). A similar method was used in **chapter 4**, in which hundreds of metabolites were combined to provide a "metabolomic age". Similarly, a combination of metabolites, proteins, and genetic variations may be used to identify disease specific profiles (72). For example, these profiles could aid in the identification of patients with a high risk of cardiometabolic disease despite exhibiting a normal weight—which is typically associated with a low risk of cardiometabolic disease. Conversely, it can also identify overweight individuals who are biologically at low risk of cardiometabolic disease. These two groups are sometimes referred to as exhibiting unfavorable and favorable adiposity respectively, and have been reported to be linked with specific SNPs (73, 74) and specific SNP-metabolomic profiles

(75). Identifying these individuals without OMICs data may be challenging for physicians. In an ideal scenario, a panel of a different OMICs results associated with favorable and unfavorable adiposity is developed. The measurements from these panels can be used for the diagnosis or risk assessment of cardiometabolic disease on an OMICs level. These types of panels can also help in reducing cost and time to produce results for clinical use.

In conclusion, OMICs research is important for understanding and disentangling disease pathophysiology, discovering novel associations, revealing effects of exposures, identifying causal pathways, bridging the gap between the roles of nature and nurture, and enhance public health and clinical care. With the continuous expansion of single and multi-OMICs studies and rapid technological advancements, it is not farfetched to expect more impactful findings in the near future. Ideally, the time between discovery to clinical application will be shortened as well.

9 REFERENCES

1. Steyerberg EW. Clinical Prediction Models. 2nd ed. Cham, Switzerland: Springer International Publishing; 2019 2019.
2. UK Biobank adds the first tranche of data from a study into circulating metabolomic biomarkers to its biomedical database 2022 [updated 2022/08/26/. Available from: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/uk-biobank-adds-the-first-tranche-of-data-from-a-study-into-circulating-metabolomic-biomarkers-to-its-biomedical-database.
3. Metabolon. Metabolon to Provide Metabolomic Profiling for the US Veterans Administration Million Veteran Program 2022 [updated 2022/08/09/. Available from: https://www.metabolon.com/news/metabolon-to-provide-metabolomic-profiling-for-the-us-veterans-administration-million-veteran-program.
4. BBMRI | BBMRI 2022 [updated 2022/11/01/. Available from: https://www.bbmri.nl.
5. Onderwater GLJ, Ligthart L, Bot M, Demirkan A, Fu J, van der Kallen CJH, et al. Large-scale plasma metabolome analysis reveals alterations in HDL metabolism in migraine. 2019;92(16):e1899-e911.
6. Litton J-E. Launch of an Infrastructure for Health Research: BBMRI-ERIC. 2018;16(3):233-41.
7. Psaty BM, O'Donnell CJ, Gudnason V, Lunetta KL, Folsom AR, Rotter JI, et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circ Cardiovasc Genet. 2009;2(1):73-80.
8. CHARGE Consortium Wiki 2022 [updated 2022/10/17/. Available from: http://depts.washington.edu/chargeco/wiki/Main_Page.
9. Arya S, Kaji AH, Boermeester MA. PRISMA Reporting Guidelines for Meta-analyses and Systematic Reviews. JAMA Surgery. 2021;156(8):789-90.
10. Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. International journal of epidemiology. 2019;48(4):1294-304.
11. Hrydziuszko O, Viant MR. Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. Metabolomics. 2012;8(1):161-74.
12. Faquih T, van Smeden M, Luo J, le Cessie S, Kastenmüller G, Krumsiek J, et al. A Workflow for Missing Values Imputation of Untargeted Metabolomics Data. Metabolites. 2020;10(12).
13. Do KT, Wahl S, Raffler J, Molnos S, Laimighofer M, Adamski J, et al. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. Metabolomics. 2018;14(10):128.
14. van Buuren S. Flexible Imputation of Missing Data. Second Edition. Boca Raton, FL.: CRC Press; 2018.
15. Fiehn O, Robertson D, Griffin J, van der Werf M, Nikolau B, Morrison N, et al. The metabolomics standards initiative (MSI). Metabolomics. 2007;3(3):175-8.
16. Shin S-Y, Fauman EB, Petersen A-K, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic influences on human blood metabolites. Nature Genetics. 2014;46(6):543-50.
17. Metabolomics GWAS Server. 2022.
18. Kong SW. Metabolites correlated with age 2021 [updated 2021/05/03/. Available from: https://tom.tch.harvard.edu/supples/metabolome/age-correlation.htm.
19. Kong SW, Hernandez-Ferrer C. Assessment of coverage for endogenous metabolites and exogenous chemical compounds using an untargeted metabolomics platform. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2020;25:587-98.
20. The MetaboLights Team MetaboLights 2022 [updated 2022/08/26/. Available from: https://www.ebi.ac.uk/metabolights/studies.
21. Elsworth B, Lyon M, Alexander T, Liu Y, Matthews P, Hallett J, et al. The MRC IEU OpenGWAS data infrastructure. 2020:2020.08.10.244293.
22. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Systems Biology. 2011;5(1):21.

23. Gagliano Taliun SA, VandeHaar P, Boughton AP, Welch RP, Taliun D, Schmidt EM, et al. Exploring and visualizing large-scale genetic associations by using PheWeb. Nat Genet. 2020;52(6):550-2.
24. Exposome-NL. Exposome-NL. 2022.
25. Arab L, Tseng CH, Ang A, Jardack P. Validity of a multipass, web-based, 24-hour self-administered recall for assessment of total energy intake in blacks and whites. American journal of epidemiology. 2011;174(11):1256-65.
26. Gibbons H, Brennan L. Metabolomics as a tool in the identification of dietary biomarkers. Proceedings of the Nutrition Society. 2017;76(1):42-53.
27. Sheikh MA, Abelsen B, Olsen JA. Differential Recall Bias, Intermediate Confounding, and Mediation Analysis in Life Course Epidemiology: An Analytic Framework with Empirical Example. Frontiers in psychology. 2016;7:1828.
28. Folpmers S, Mook-Kanamori DO, de Mutsert R, Rosendaal FR, Willems van Dijk K, van Heemst D, et al. Agreement between nicotine metabolites in blood and self-reported smoking status: The Netherlands Epidemiology of Obesity study. Addictive behaviors reports. 2022;16:100457.
29. Schjerning AM, McGettigan P, Gislason G. Cardiovascular effects and safety of (non-aspirin) NSAIDs. Nature reviews Cardiology. 2020;17(9):574-84.
30. Bavry AA, Khaliq A, Gong Y, Handberg EM, Cooper-DeHoff RM, Pepine CJ. Harmful Effects of NSAIDs among Patients with Hypertension and Coronary Artery Disease. The American Journal of Medicine. 2011;124(7):614-20.
31. Straube S, Tramèr MR, Moore RA, Derry S, McQuay HJ. Mortality with upper gastrointestinal bleeding and perforation: effects of time and NSAID use. BMC Gastroenterology. 2009;9(1):41.
32. Bedene A, van Dorp ELA, Rosendaal FR, Dahan A, Lijfering WM. Risk of drug-related upper gastrointestinal bleeding in the total population of the Netherlands: a time-trend analysis. 2022;9(1):e000733.
33. Hallare J, Gerriets V. Half Life. 2022 2022 Jun 23. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing. Available from: https://www.ncbi.nlm.nih.gov/books/NBK554498/.
34. Génin E. Missing heritability of complex diseases: case solved? Human Genetics. 2020;139(1):103-13.
35. Sliz E, Sebert S, Würtz P, Kangas AJ, Soininen P, Lehtimäki T, et al. NAFLD risk alleles in PNPLA3, TM6SF2, GCKR and LYPLAL1 show divergent metabolic effects. Human molecular genetics. 2018;27(12):2214-23.
36. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ. 2018;362:k601.
37. Goldenberg NA, Everett AD, Graham D, Bernard TJ, Nowak-Göttl U. Proteomic and other mass spectrometry based “omics” biomarker discovery and validation in pediatric venous thromboembolism and arterial ischemic stroke: Current state, unmet needs, and future directions. PROTEOMICS – Clinical Applications. 2014;8(11-12):828-36.
38. Cushman M, Barnes GD, Creager MA, Diaz JA, Henke PK, Machlus KR, et al. Venous thromboembolism research priorities: A scientific statement from the American Heart Association and the International Society on Thrombosis and Haemostasis. Research and practice in thrombosis and haemostasis. 2020;4(5):714-21.
39. Russell C, Rahman A, Mohammed AR. Application of genomics, proteomics and metabolomics in drug discovery, development and clinic. Therapeutic delivery. 2013;4(3):395-413.
40. Hsu YH, Astley CM, Cole JB, Vedantam S, Mercader JM, Metspalu A, et al. Integrating untargeted metabolomics, genetically informed causal inference, and pathway enrichment to define the obesity metabolome. International journal of obesity (2005). 2020;44(7):1596-606.
41. Richard MA, Lupo PJ, Zachariah JP. Causal Inference of Carnitine on Blood Pressure and potential mediation by uric acid: A mendelian randomization analysis. Int J Cardiol Cardiovasc Risk Prev. 2021;11:200120.
42. Surendran P, Stewart ID, Au Yeung VPW, Pietzner M, Raffler J, Wörheide MA, et al. Rare and common genetic determinants of metabolic individuality and their effects on human health. Nature Medicine. 2022;28(11):2321-32.

43. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. Nature. 2018;558(7708):73-9.
44. Balzer LB, Petersen ML. Invited Commentary: Machine Learning in Causal Inference-How Do I Love Thee? Let Me Count the Ways. American journal of epidemiology. 2021;190(8):1483-7.
45. Beam AL, Manrai AK, Ghassemi M. Challenges to the Reproducibility of Machine Learning Models in Health Care. Jama. 2020;323(4):305-6.
46. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021;11(7):e048008.
47. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: Still a ways to go. Science Translational Medicine. 2021;13(586):eabb1655.
48. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology. 2019;110:12-22.
49. PubMed. multi-omics - Search Results - PubMed 2022 [updated 2022/11/14/. Available from: https://pubmed.ncbi.nlm.nih.gov/?term=multi-omics&filter=years.2006-2022&show_snippets=off.
50. Sproston NR, Ashworth JJ. Role of C-Reactive Protein at Sites of Inflammation and Infection. Frontiers in immunology. 2018;9:754.
51. Kearon C, de Wit K, Parpia S, Schulman S, Spencer FA, Sharma S, et al. Diagnosis of deep vein thrombosis with D-dimer adjusted to clinical probability: prospective diagnostic management study. 2022;376:e067378.
52. Pulivarthi S, Gurram MK. Effectiveness of d-dimer as a screening test for venous thromboembolism: an update. North American journal of medical sciences. 2014;6(10):491-9.
53. Maisel AS, Krishnaswamy P, Nowak RM, McCord J, Hollander JE, Duc P, et al. Rapid Measurement of B-Type Natriuretic Peptide in the Emergency Diagnosis of Heart Failure. 2002;347(3):161-7.
54. Nakagawa O, Ogawa Y, Itoh H, Suga S, Komatsu Y, Kishimoto I, et al. Rapid transcriptional activation and early mRNA turnover of brain natriuretic peptide in cardiocyte hypertrophy. Evidence for brain natriuretic peptide as an “emergency” cardiac hormone against ventricular overload. The Journal of Clinical Investigation. 1995;96(3):1280-7.
55. Sudoh T, Kangawa K, Minamino N, Matsuo HJN. A new natriuretic peptide in porcine brain. 1988;332(6159):78-81.
56. Toland AE, Forman A, Couch FJ, Culver JO, Eccles DM, Foulkes WD, et al. Clinical testing of BRCA1 and BRCA2: a worldwide snapshot of technological practices. npj Genomic Medicine. 2018;3(1):7.
57. Krstić N, Običan SG. Current landscape of prenatal genetic screening and testing. 2020;112(4):321-31.
58. Rijksinstituut voor Volksgezondheid en Milieu. Heel prick screening test | Prenatale en neonatale screeningen 2022 [updated 2022/09/20/. Available from: https://www.pns.nl/prenatal-and-newborn-screening/heel-prick.
59. Mamas M, Dunn WB, Neyses L, Goodacre R. The role of metabolites and metabolomics in clinically applicable biomarkers of disease. Archives of Toxicology. 2011;85(1):5-17.
60. McCarthy JJ, McLeod HL, Ginsburg GS. Genomic Medicine: A Decade of Successes, Challenges, and Opportunities. Science Translational Medicine. 2013;5(189):189sr4-sr4.
61. Gibbs RA. The Human Genome Project changed everything. Nature Reviews Genetics. 2020;21(10):575-6.
62. Semenov AG, Feygina EE. Chapter One - Standardization of BNP and NT-proBNP Immunoassays in Light of the Diverse and Complex Nature of Circulating BNP-Related Peptides. In: Makowski GS, editor. Advances in Clinical Chemistry. 85: Elsevier; 2018. p. 1-30.
63. Gaziano L, Giambartolomei C, Pereira AC, Gaulton A, Posner DC, Swanson SA, et al. Actionable druggable genome-wide Mendelian randomization identifies repurposing opportunities for COVID-19. Nature Medicine. 2021;27(4):668-76.

64. Lopes RD, Macedo AVS, de Barros ESPGM, Moll-Bernardes RJ, Feldman A, D'Andréa Saba Arruda G, et al. Continuing versus suspending angiotensin-converting enzyme inhibitors and angiotensin receptor blockers: Impact on adverse outcomes in hospitalized patients with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)--The BRACE CORONA Trial. American heart journal. 2020;226:49-59.
65. Seruga B, Ocana A, Amir E, Tannock IF. Failures in Phase III: Causes and Consequences. Clinical cancer research : an official journal of the American Association for Cancer Research. 2015;21(20):4552-60.
66. Stewart DJ, Kurzrock R. Fool's gold, lost treasures, and the randomized clinical trial. BMC cancer. 2013;13:193.
67. Blauw LL, Li-Gao R, Noordam R, de Mutsert R, Trompet S, Berbée JFP, et al. CETP (Cholesteryl Ester Transfer Protein) Concentration: A Genome-Wide Association Study Followed by Mendelian Randomization on Coronary Artery Disease. Circulation Genomic and precision medicine. 2018;11(5):e002034.
68. Ference BA, Holmes MV, Smith GD. Using Mendelian Randomization to Improve the Design of Randomized Trials. Cold Spring Harbor perspectives in medicine. 2021;11(7).
69. Henry A, Gordillo-Marañón M, Finan C, Schmidt AF, Ferreira JP, Karra R, et al. Therapeutic Targets for Heart Failure Identified Using Proteomics and Mendelian Randomization. Circulation. 2022;145(16):1205-17.
70. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Human molecular genetics. 2014;23(R1):R89-98.
71. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Medicine. 2020;12(1):44.
72. Suhre K, Zaghlool S. Connecting the epigenome, metabolome and proteome for a deeper understanding of disease. Journal of internal medicine. 2021;290(3):527-48.
73. Ji Y, Yiorkas AM, Frau F, Mook-Kanamori D, Staiger H, Thomas EL, et al. Genome-Wide and Abdominal MRI Data Provide Evidence That a Genetically Determined Favorable Adiposity Phenotype Is Characterized by Lower Ectopic Liver Fat and Lower Risk of Type 2 Diabetes, Heart Disease, and Hypertension. Diabetes. 2018;68(1):207-19.
74. Yaghootkar H, Lotta LA, Tyrrell J, Smit RAJ, Jones SE, Donnelly L, et al. Genetic Evidence for a Link Between Favorable Adiposity and Lower Risk of Type 2 Diabetes, Hypertension, and Heart Disease. Diabetes. 2016;65(8):2448-60.
75. Cirulli ET, Guo L, Leon Swisher C, Shah N, Huang L, Napier LA, et al. Profound Perturbation of the Metabolome in Obesity Is Associated with Health Risk. Cell Metabolism. 2019;29(2):488-500.e2.

Chapter 9
Summary of Findings (Dutch, English, Arabic)

SAMENVATTING

De recente ontwikkelingen in metabolomics (de studie van metabolieten), proteomics (de studie van eiwitten) en genomics (de studie van genen) geven diepgaande inzichten in de menselijke biologie en de pathofysiologie van ziekten. In dit proefschrift onderzochten we de methodologische uitdagingen waar deze OMICs-technologieën voor staan en pasten we deze toe in epidemiologische studies.

In deel één concentreerden we ons op enkele van de methodologische uitdagingen waar OMICs-onderzoek voor staat. In **Hoofdstuk 2** bestudeerden we de overeenkomst tussen proteomics metingen van biomarkers voor veneuze trombose op basis van aptameren en de gestandaardiseerde klinische metingen. Voor negen onderzochte veneuze trombose biomarkers, vonden we dat er weinig overeenkomst was tussen de twee methoden. De overeenkomst was in het bijzonder slecht voor D-dimeer, een marker die vaak gebruikt wordt in veneuze trombose diagnostiek. In **Hoofdstuk 3** hebben we onderzocht hoe om te gaan met ontbrekende waarden in metabolomics data, vooral bij gebruik van 'untargeted metabolomics'. We hebben de prestaties van twee eerder gerapporteerde imputatie methoden beoordeeld, door het simuleren van ontbrekende data in metingen van de NEO-studie, namelijk meervoudige imputatie en "*k*-nearest neighbor". Onze resultaten lieten zien dat de meervoudige imputatie methode in de meeste scenario's resulteerde in minder bias dan de *k*-nearest neighbor methode, behalve in het scenario met kleine steekproefgrootte en veel ontbrekende waarden. We hebben een openbaar beschikbaar R-script gemaakt om het imputeren van ontbrekende metabolomics data met behulp van deze twee imputatie technieken te versnellen.

Metabolieten kunnen unieke inzichten geven in de pathofysiologie van ziekten, aangezien ze de samengestelde effecten omvatten van genetische, leefstijl- en omgevingsfactoren. Daarom wordt metabolomics ook gebruikt om bijvoorbeeld biologische processen in veroudering te bestuderen. Sommige studies hebben metabolomics data gebruikt, om met behulp van voorspelmodellen de "metabolomische leeftijd" te schatten, een weerspiegeling van veroudering op het niveau van het metabolisme. In **Hoofdstuk 4** bespreken we veelvoorkomende uitdagingen bij het ontwikkelen van voorspelmodellen op basis van metabolomics data. Daarnaast hebben we zelf een model ontwikkeld om de metabolomische leeftijd te voorspellen op basis van metabolieten gemeten met het Metabolon platform. We ontwikkelden het model in de INTERVAL-studie met behulp van 'ridge regressie' en 'bootstrapping'. Deze populatie was uitermate geschikt voor het ontwikkelen van een model omdat het een relatief gezonde, grote populatie is, (n = 11.977) met zeer uiteenlopende leeftijden. Uiteindelijk was ons model goed in staat om de metabolomische leeftijd te voorspellen (gecorrigeerde R^2 = 0.83).

Het tweede deel van dit proefschrift richtte zich op verschillende epidemiologische onderzoeksvragen waarbij gebruik gemaakt wordt van genomics and metabolomics data (Metabolon en Nightingale platforms), en het toepassen van geavanceerde data-analyse methoden. In **Hoofdstuk 5** onderzochten we de associaties tussen meer dan 1300 metabolieten gemeten met het Metabolon platform en levervet. We vonden associaties voor een reeks endogene metabolieten (vooral aminozuren en lipiden), evenals voor nieuwe en ongedefinieerde metabolieten. Daarnaast hebben we twee complementaire holistische netwerken van metabolieten geconstrueerd, één op biologische basis en één op statistische basis, om de geassocieerde metabolieten toe te wijzen aan biologische 'pathways'. Deze interactieve netwerken gaven inzicht in potentiële nieuwe pathways die een relatie hebben met levervet en mogelijk met leverziekten. Deze interactieve netwerken hebben we op een online, publiek beschikbare webpagina, beschikbaar gesteld voor verdere verkenning door de onderzoeksgemeenschap.

Een type variatie in het genoom is variatie in het aantal herhaalde cytosine-adenine-guanine-nucleotidesequenties (CAG-herhalingen). Wanneer het aantal CAG-herhalingen boven een specifieke pathogene drempelwaarde ligt, en deze in of in de buurt van specifieke genen liggen, kunnen ze tot ernstige neurodegeneratieve ziekten leiden. Een voorbeeld hiervan is het aantal CAG-herhalingen in het Huntingtine-gen (*HTT*-gen) dat leidt tot het optreden van de ziekte van Huntington. Uit recent onderzoek is gebleken dat het aantal CAG-herhalingen, zelfs onder de pathogene drempelwaarde, gerelateerd is aan gezondheidsuitkomsten, zoals cognitieve functie, depressie en gewicht. In **Hoofdstuk 6** hebben we de mogelijke associaties verkend tussen variatie in het aantal CAG-herhalingen onder de pathogene drempel in het *HTT*-gen en metaboliet concentraties. Hiervoor hebben we gebruik gemaakt van de genetische data van 3 grote Europese cohort studies (n = 10,275) met metaboliet data van het Nightingale platform. We hebben gevonden dat het aantal niet-pathogene CAG-herhalingen in het *HTT*-gen geassocieerd was met een ongezond metabolieten profiel dat vergelijkbaar is met dat van patiënten met cardiometabole ziekten.

Eén van de belangrijkste voordelen van metabolomics is dat het ook mogelijk is om metabolieten te meten die oorspronkelijk uit de omgeving komen, ook wel xenobiotica genoemd, zoals uit medicatie, voedingsstofffen, en vervuilende stoffen uit het milieu. Een bekende groep xenobiotica zijn de per- en polyfluoralkylstoffen (PFAS), die door de mens zijn gemaakt. Deze stoffen worden ook wel 'forever chemicals' genoemd, omdat ze heel lang in het milieu en in het lichaam blijven en nauwelijks op natuurlijke wijze worden afgebroken. Hoge concentraties van PFAS kunnen tot ziekte leiden. Ondanks maatregelen om de productie te verminderen, blijven PFAS meetbaar in de grond en het drinkwater, alsmede ook in het bloed van inwoners van Nederland en Duitsland. In **Hoofdstuk 7**, hebben we de associaties bestudeerd tussen PFAS concentraties in het bloed, gemeten met het Metabolon platform, en metabolieten/lipoproteïnen, gemeten met het Nightingale platform in twee cohorten, één uit Nederland (NEO Studie) en één uit Duitsland (Rhineland Studie). PFAS was meetbaar in nagenoeg alle deelnemers van deze studies. Uit de meta-analyse bleek daarnaast dat PFAS concentraties geassocieerd waren met een ongezond metabolomisch profiel, met verhoogde verzuren, low-density lipoproteine en apoproteine-B concentraties. Deze associatie was het sterkst in jongere mensen. Dit profiel is zeer vergelijkbaar met het profiel dat we kennen van individuen met een verhoogd cardiometabool risico. Dit geeft aan dat PFAS een gezondheidsrisico zijn, zelfs in mensen die schijnbaar niet aan hoge concentraties PFAS blootgesteld zijn.

SUMMARY OF MAIN FINDINGS

Metabolomics, proteomics, and genomics analyses provide profound insight into human biology and disease pathophysiology In this thesis, we explored the methodological challenges facing these OMICs technologies and illustrated their applications in epidemiological studies. In part one, we focused on some of the methodological challenges facing OMICs research. **Chapter 2** examined the agreement between aptamer-based proteomics measurements of venous thrombosis (VTE) biomarkers with those measured by standardized clinical instruments. We reported that 9 venous thrombosis biomarkers showed poor agreement with the clinical measurements. The agreement was particularly poor for D-dimer, a marker often used in VTE diagnosis. In **Chapter 3**, we explored the challenges of handling missing values in metabolomics data, particularly from untargeted metabolomics platforms. We assessed the performance of two previously reported methods for imputation, namely "k-nearest neighbor" (kNN) and "multiple imputation using chained equations", by simulating missing patterns in data from the NEO study. Our findings showed that the multiple imputation method had less bias than the kNN method in most scenarios, except for the scenario with small sample size and high level of missingness. We further provided a publicly available R script to streamline the process of imputing missing metabolomics data using the two methods.

Metabolites can provide unique insight into biological pathophysiology as they encompass the cumulative effects of genetic, lifestyle, and environmental factors. Thus, metabolomics analyses have been used to study biological processes such as aging. Some studies have used metabolomics to estimate "metabolomic age", a reflection of aging on a metabolomic level, using prediction modelling methods. In **Chapter 4**, we addressed common challenges for developing metabolomics-based age prediction models and developed a model to predict metabolomic age using the Metabolon metabolomics platform measurements. We developed a model in the INTERVAL study using ridge regression and bootstrapping. This study population is a relatively healthy large population (n = 11,977) with a wide age range. Overall, after development and internal validation, our prediction models for metabolomic age demonstrated good performance (adjusted R^2 = 0.83).

The second part of the thesis addressed various epidemiological research questions by utilizing genomic data and metabolomics measurements (Metabolon and Nightingale platforms) and using advanced data analysis methods. In **Chapter 5**, we examined the associations of over 1300 metabolites, measured by the Metabolon platform, with hepatic triglycerides content (HTGC). These associations included a range of endogenous metabolites from various pathways (particularly amino acids and lipids), as well as novel and uncharacterized metabolites. Subsequently, we constructed two complementary holistic networks, one biologically based and the other statistically driven, to assign the associated metabolites to pathways. These interactive networks revealed potential novel pathways associated with HTGC and possibly fatty liver diseases. We further provided these networks on an online, publicly available webpage for further exploration by the research community.

One type of variation in the genome is variation in the number of repeated cytosine-adenine-guanine nucleotide sequences (CAG repeats). When the size of these repeats is beyond specific pathogenic thresholds in or near specific genes, they lead to the onset of several severe neurodegenerative diseases. One such case is the large CAG repeat sizes in the huntingtin (*HTT*) gene leading to the onset of Huntington's disease. Interestingly, recent studies have found that CAG repeat sizes even within the non-pathogenic range are associated with health outcomes

such as cognitive function, depression, and body weight. In **Chapter 6**, we explored the possible association between CAG repeat variation sizes below the pathogenic threshold (<36) in the *HTT* gene and the metabolite levels measured by the Nightingale platform. Multilevel mixed-effects and mediation analyses were conducted in pooled data from three large European cohorts (n = 10,275). Our results showed that large non-pathogenic CAG repeat sizes were associated with an unfavorable metabolomic profile, similar to that for an increased risk of cardiometabolic diseases.

One of the advantages of metabolomics is the capability to measure "external" metabolites, often referred to as xenobiotics, from medication use, diet, and environmental exposures and contaminations. One such group of measured xenobiotics are the man-made per- and polyfluoroalkyl substances (PFAS)—commonly known as "forever chemicals" due to their persistence in the environment and the body. In addition to their persistent nature, high concentrations of PFAS have been reported to be associated with severe health impact. Despite efforts to regulate their production, PFAS levels remain detectable in the soil, drinking water, and human blood samples in the Netherlands and Germany. In **Chapter 7**, we studied the associations between common PFAS levels in blood, as measured by the Metabolon platform, and metabolomics and lipoprotein profiles (Nightingale platform) of the general population in the Netherlands (NEO study) and Germany (Rhineland study). Overall, PFAS metabolites were detectable in nearly all participants of both studies. Furthermore, as confirmed by meta-analysis, PFAS levels were associated with an unhealthy metabolomic profile characterized by increased fatty acids, low density lipoproteins, and apolipoprotein B levels. These associations are reminiscent of the metabolomic profile for increased risk of cardiometabolic diseases and were found to be particularly stronger in younger individuals. These data indicate that PFAS are a health risk even in individuals that apparently have not been exposed to high levels of PFAS.

ملخص النتائج

توفر علوم تحليل منتجات الأيض (الميتابولوميكس / Metabolomics) وتحليل البروتينات (البروتيوميكس / proteomics) و تحليل الجينات (الجينومكس / Genomics) رؤى عميقة في علم الأحياء البشري وعلم وظائف الأعضاء المتعلقة بالأمراض وأسبابها في الفيزيولوجيا المرضية البيولوجية (Pathophysiology) على نطاق غير مسبوق. وقد استكشفنا في هذه الأطروحة التحديات البحثية التي تواجه تطبيق تقنيات الـ OMICs[1] (أوميكس) بنظرة تكاملية عميقة وأوضحنا تطبيقاتها العملية في الدراسات الوبائية (Epidemiological Studies).

وفي الجزء الأول من الأطروحة المكون من ثلاثة فصول، ركزنا على بعض تحديات طرق البحث المنهجية التي تواجه أبحاث الاوميكس. و**يتفحص الفصل الأول من الجزء الأول (Chapter 2)** من هذه الرسالة العلمية العلاقة ما بين نتائج قياسات البروتينات باستخدام "الأبتامير" (aptamer-based proteomics) للمؤشرات الحيوية للتخثر الوريدي (VTE[2]) مع طرق القياس المعتمدة التي تتم عادة في المراكز الصحية والممارسات الاكلينيكية المتعارف عليها. وقد وجدت دراستنا البحثية أن المؤشرات الحيوية التسعة التي تم قياسها للتخثر الوريدي أظهرت توافقا ضعيفا مع نتائج القياسات المعتمدة. وكانت العلاقة التوافقية ضعيفة بشكل خاص بالنسبة لـ"ثنائيات **د"** (D-dimer)، وهو مؤشر يستخدم لتشخيص داء التخثر الوريدي.

في **الفصل الثاني من الجزء الأول (Chapter 3)** من هذه الرساله ، تم دراسة التحديات المتمثلة في التعامل مع البيانات المفقودة في تحليل منتجات الأيض ، لا سيما من منصات بيانات تحليل منتجات الأيض الـ"غير مستهدفة".حيث قمنا بتقييم أداء طريقتين لاحتساب القيم المفقودة في بيانات تحليل منتجات الأيض، وهما طريقة "K-nearest neighbor" (kNN) و طريقة"Multiple imputation using chained equations" (التضمين المتعدد)، ومن خلال محاكاة الأنماط المختلفه لفقدان البيانات من الدراسة البحثية الهولندية عن السمنة [3] (NEO). أظهرت نتائجنا أن طريقة "التضمين المتعدد" كان لها تحيز أقل من طريقة kNN في معظم الحالات ، الا في الحالات ذات حجم عينة صغيره مع نسبه عاليه من البيانات المفقوده. كما قدمنا أيضا برنامجاً نصي حاسوبي بلغة برمجة R متاح للعامة لتبسيط عملية استخدام الطريقتين لاحتساب القيم المفقودة في بيانات قياسات منتجات الأيض.

بامكان منتجات الأيض (الميتابولايتس/ Metabolites) توفير نظرة ثاقبة وفريدة في الفيزيولوجيا المرضية البيولوجية لأنها تشمل تأثير العوامل الوراثية و تأثير نمط الحياة و تأثير العوامل البيئية. فلذلك تم استخدام علم تحليل منتجات الأيض لدراسة العملية الحيويه للتقدم بالعمر و الهرم. وقد استخدمت بعض الدراسات منتجات الأيض لتقدير "العمر الأيضي" ، وهو انعكاس للتقدم بالعمر على مستوى منتجات الأيض ، باستخدام طرق نمذجة التنبؤ الإحصائي. في **الفصل الثالث من الجزء الأول (Chapter 4)**، تناولنا التحديات الشائعة لتطوير نماذج التنبؤ بالعمر القائمة على منتجات الأيض

وطورنا نموذجا للتنبؤ بالعمر المبنى على ذلك باستخدام أجهزة منصة Metabolon لتحليل منتجات الأيض. بعد ذلك قمنا بتطوير نموذج في الدراسة البحثية INTERVAL[4] باستخدام أساليب الإحصاء "Ridge regression" و "Bootstrapping" .وقد اظهرت المجموعة المتضمنه في دراسة الـ INTERVAL ان عينة كبيرة من المجتمع كانت في صحة جيده نسبيا (n = 11,977) و مع نطاق عمري واسع ولذلك فان هذه المجموعه كانت مناسبه لتطوير نموذج تنبؤ العمر الأيضي. وبعد تطوير النموذج والتحقق الداخلي من صحته، أظهرت النتائج ان أداء التنبؤ الإحصائي للعمر الأيضي كان جيداً جداً (R^2 المعدل = 0.83).

اما الجزء الثاني من الأطروحة ، والمكون من ثلاثة فصول ايضا، فيتناول دراسة عدة أسئلة بحثية في علم الepidemiology من خلال استخدام البيانات الجينومية و بيانات تحليل الأيض (باستخدام منصتين Metabolon و Nightingale) بالإضافة إلى طرق تحليل البيانات المتقدمة.

في **الفصل الأول من الجزء الثاني (Chapter 5)،** قمنا بفحص ارتباطات أكثر من 1300 منتج أيضي، تم قياسها بواسطة منصة Metabolon، و محتوى الدهون الثلاثية الكبدية (HTGC). وقد تضمنت هذه الارتباطات مجموعات من منتجات الأيض تابعه لمسارات حيويه مختلفة (خاصة من مجموعات الأحماض الأمينية والدهون) ، بالإضافة إلى منتجات أيض غير معهوده. بعد ذلك، قمنا ببناء شبكتين شموليتين متكاملتين، واحدة قائمة على أساس حيوي والأخرى قائمة على أساس إحصائي ، لتعيين وربط منتجات الأيض مع المسارات الحيويه المناسبه. وقد كشفت هذه الشبكات التفاعلية عن مسارات حيويه غير مألوفه يمكن ان تكون مرتبطة بال HTGC وربما أمراض الكبد الدهنية ككل هذا وقد تم توفير هذه الشبكات على صفحة الويب المتاحة على الإنترنت للمزيد من الاستكشاف من قبل المجتمع البحثي العلمي.

اما في الفصل الثاني من الجزء الثاني (Chapter 6)، فاستكشفنا الارتباطات المحتملة لأحجام تباين تكرار CAG الأقل حجما من عتبة تسبب مرض هنتنغتون (اقل من 36 تكرار) في جين *HTT* مع مستويات الأيض المقاسة بواسطة منصة Nightingale. وتبين أن أحد أنواع الخلل الجينية في الجينوم هو خلل ناتج عن اختلاف عدد تسلسل نيوكليوتيدات (Nucleotide) السيتوزين-الأدينين-الجوانين (تكرار تسلسل CAG[5]). وعندما يتجاوز حجم هذا التكرار العتبة الغير مرضيه في جينات معينة (أو بالقرب من هذه الجينات) ، فإنها تؤدي إلى العديد من الأمراض التنكسية العصبية الشديدة. إحدى هذه الحالات هي حجم تكرار تسلسل CAG الكبيرة في جين هنتنغتين[6] *(HTT)* مما يؤدي إلى مرض هنتنغتون[7]. من والمثير للاهتمام ، أن الدراسات الحديثة وجدت أن حجم تكرار CAG حتى اذا كان داخل النطاق الغير مرضي فان له تاثير على وظائف صحية مثل الوظيفة الإدراكية والاكتئاب وانخفاض وزن الجسم وبناء على ذلك تم إجراء تحليل البيانات بطريقه تحليل "التأثيرات المختلطة"[8] و تحليل "الوساطة"[9] في بيانات تم جمعها من ثلاث دراسات أوروبية ضخمه (n = 10,275). حيث أظهرت نتائجنا أن أحجام تكرار CAG الكبيرة الغير مسببة للأمراض كانت مرتبطة بنمط أيضي غير ملائم للصحة، على غرار النمط الأيضي المتصل بزيادة خطر الإصابة بأمراض القلب والأوعية الدموية.

وتتمثل إحدى مزايا تحليل منتجات الأيض في قدرة الغطاء على قياس منتجات الأيض "الخارجية الغريبة"، والتي يشار إليها غالبا باسم Xenobitoics. وهذه المنتجات الأيضيه يكون منشاها من استخدام الأدوية والنظام الغذائي والتعرض للتلوث البيئي. وإحدى هذه الكائنات الخارجية المقاسة هي مواد "Per- and Polyfluoroalkyl substances (PFAS)" (بيفاس) الصناعيه - والمعروفة باسم "الكيماويات الأبديه" بسبب تشبثها في البيئة وجسم الكائنات الحية، و بالإضافة إلى طبيعتها التشبثيه، فقد اوضحت الدراسات ان التركيزات العالية من الPFAS مرتبطة بمشاكل صحيه شديده في الانسان. وانه على الرغم من الجهود المبذولة لتنظيم إنتاجها، لا تزال مستويات PFAS قابلة للاكتشاف في التربة ومياه الشرب وعينات الدم البشري في كل من هولندا وألمانيا.

فلذلك قمنا في **الفصل الثالث من الجزء الثاني (Chapter 7)** ، بدراسه العلاقة ما بين مستويات PFAS في الدم ، التى تم قياسها بواسطة منصة Metabolon ، و مستويات المنتجات الأيضيه و البروتين الدهني التي تم قياسها بواسطة منصة Nightingale في مجموعة من عامة سكان هولندا (دراسة NEO) وألمانيا (دراسة راينلاند[10]). ووجدنا ان مستوى المنتجات الأيضيه للPFAS قابل للقياس في جميع الافراد تقريبا في كلتا الدراستين، علاوة على ذلك وكما أكد تحليل الميتا[11] فقد ارتبطت مستويات PFAS بنمط أيضي غير صحي يتميز بزيادة مستوى "الأحماض الدهنية" و"البروتينات الدهنية منخفضة الكثافة" ومستويات "البروتين الشحمي **(ب)**"[12]. وكانت هذه الارتباطات متشابهة بالنمط الأيضي الذي يزيد خطر الإصابة بأمراض القلب والأوعية الدموية .ووجدنا ايضا أن هذا النمط شديد بشكل خاص لدى الأفراد الأصغر سنا. هذه النتائج تشير الى أن الPFAS يمثل خطرًا صحيًا حتى في الأفراد الذين لم يتعرضوا على ما يبدو لمستويات عالية من PFAS.

[1] "OMICs" هو مجال في العلم الحديث الذي يتطلع إلى الفهم التام للحياة والعمليات الحيوية عبر التحليل الجزيئي للعناصر المختلفة التي تشكل الحياة، مثل الجينات والبروتينات والأملاح والمركبات الكيميائية الأخرى حيث يتم تحليل هذه العناصر عبر عدة انماط "أوميكس" مختلفة، مثل "جينوميكس" (تحليل الجينات) و"بروتوميكس" (تحليل البروتينات). يستخدم هذا المجال في التحقيقات العلمية لفهم أفضل للعمليات الحيوية وللعثور على حلول للمشاكل الصحية المختلفة.

[2] Venous Thrombosis

[3] The Netherlands Epidemiology of Obesity Study (NEO)

[4] تم إنشاء دراسة INTERVAL بالتعاون بين جامعتي كامبريدج وأوكسفورد و التعاون ايضا مع دائرة الدم والتبرعات الصحية في المملكة المتحدة

[5] Cytosine-adenine-guanine nucleotide sequences

[6] Huntingtin gene

[7] Huntington's disease

[8] Multilevel mixed-effects model

[9] Mediation analysis

[10] Rhineland study

[11] Meta analysis

[12] Apolipoprotein B

Appendix

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my parents for their unwavering love, support, and encouragement throughout my life and academic journey. Their love and belief in me have been, and always will be, a constant source of inspiration and motivation. I also thank my sister Dr. Nada Faquih for her support and for assisting in translating the summary of findings of this thesis to Arabic.

I am also deeply grateful to my supervisors: My promotor Prof.dr. Ko Willemas van Dijk taught me valuable lessons on conducting research and writing scientific papers. He was always there for me and made time to aid in any matter I had, big or small; My co-supervisor Dr. Dennis O. Mook-Kanamori was my first contact for this PhD position. Without him I would have never had this life changing opportunity. He provided me the guidance and the creative freedoms to progressively become an independent epidemiological researcher; My co-supervisor Dr. Astid van Hylckama Vlieg for her supervision, support and her invaluable guidance and epidemiological knowledge. In these past 4 years I have learned so much from them about epidemiology, "OMICs" and conducting scientific research. Their expertise, dedication, support and kindness have been instrumental in the completion of this PhD thesis. I can truly say, thanks to them, I have transformed to a real epidemiologist and found my true calling for research. They are the best supervisors I could have asked for.

I would like to extend my thanks to the members of the Clinical Epidemiology department for their support and assistance throughout my studies. I want to particularly thank Dr. Ruifang Li-Gao for her feedback and suggestions, Prof.dr.Saskia le Cessie for always sparing the time to provide her insightful statistical knowledge, Dr. Renée de Mutsert for her support and input for all my NEO study related projects. Special thanks to Yvonne Souverein for her help in all organizational matters and Ingeborg de Jonge for providing the necessary NEO data for my projects. I also want to thank Prof.dr.Olaf Dekkers, Dr. Raymond Noordam, Dr. Jan van Klinken, Dr. Maarten van Smeden, Dr. Jelle Goeman for sharing their expertise that aided the completion of my projects.

Finally, I want to extend a special thanks to all my colleagues in the past four years for their encouragement and friendship. Not only were valuable colleagues, they were true friends. I am grateful for the memories we have shared together and the valuable lessons that I have learned from them.

This dissertation would not have been possible without the support of all of you. Thank you from the bottom of my heart.

CURRICULUM VITAE

Tariq Faquih was born on the 4th of September 1989 (4th Ṣafar 1410 Hijri) in Riyadh, Saudi Arabia. He received a scholarship from the King Abdullah Scholarship program to attend the faculty of Biological Sciences in the University of Leeds (United Kingdom) from 2007 to 2011. After he received a BSc degree in Human Genetics, he attended the Biochemistry department at Georgetown University (Washington DC, United States) from 2012-2013 and obtained a MSc degree in Bioinformatics. He subsequently worked at the King Faisal Specialist Hospital and Research Center (Riyadh, Saudi Arabia) as a bioinformatician with the Saudi Human Genome Project team from 2013-2018.

In 2018 he received a scholarship from the King Faisal Specialist Hospital and the King Abdullah Scholarship program to begin his PhD at the Clinical Epidemiology Department at the Leiden University Medical Center (LUMC). There, he was under the supervision of Prof.dr.ir. J.A.P. Willems van Dijk, Dr. Dennis O. Mook-Kanamori, and Dr. Astrid van Hylckama Vlieg. His work focused on the epidemiological applications of multi-OMICs in the study of a variety of diseases and outcomes. These OMICs included genomics, metabolomics, and proteomics. During his PhD he worked briefly as a visiting researcher at the University of Cambridge (Department of Public Health and Primary care) under the co-supervision of Dr. Praveen Surendran. He also collaborated with the Rhineland Study, DZNE (Germany), Pravastatin in elderly individuals at risk of vascular disease (PROSPER) study, the Thrombophilia, Hypercoagulability and Environmental Risks in Venous Thromboembolism (THE-VTE) study, and with the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium.

In January 2023 he began working as a postdoctoral research fellow in Medicine in the Division of Sleep and Circadian Disorders, Department of Medicine, Brigham and Women's Hospital (BWH) and Research Fellow at the Harvard Medical School, Harvard University (Boston, United States).

LIST OF PUBLICATIONS

Published articles included in this thesis

1. **Faquih T**, van Smeden M, Luo J, le Cessie S, Kastenmüller G, Krumsiek J, Noordam R, van Heemst D, Rosendaal FR, van Hylckama Vlieg A, Willems van Dijk K, Mook-Kanamori DO. **A Workflow for Missing Values Imputation of Untargeted Metabolomics Data.** Metabolites. 2020 Nov 26;10(12):486. doi: 10.3390/metabo10120486. PMID: 33256233; PMCID: PMC7761057.
2. **Faquih T**, Mook-Kanamori DO, Rosendaal FR, Baglin T, Willems van Dijk K, van Hylckama Vlieg A. **Agreement of aptamer proteomics with standard methods for measuring venous thrombosis biomarkers.** Res Pract Thromb Haemost. 2021 May 4;5(4):e12526. doi: 10.1002/rth2.12526. PMID: 34013156; PMCID: PMC8110437.
3. **Faquih TO**, Aziz NA, Gardiner SL, Li-Gao R, de Mutsert R, Milaneschi Y, Trompet S, Jukema JW, Rosendaal FR, Hylckama Vlieg A, Dijk KW, Mook-Kanamori DO. **Normal range CAG repeat size variations in the HTT gene are associated with an adverse lipoprotein profile partially mediated by body mass index.** Hum Mol Genet. 2023 Jan 30:ddad020. doi: 10.1093/hmg/ddad020. Epub ahead of print. PMID: 36715614.

Other publications

1. Shrine N, …, **Faquih T**, et al. **Multi-ancestry genome-wide association study improves resolution of genes, pathways and pleiotropy for lung function and chronic obstructive pulmonary disease.** medRxiv 2022.05.11.22274314; doi: https://doi.org/10.1101/2022.05.11.22274314 8.
2. Bedene A, van Dorp ELA, **Faquih T**, Cannegieter SC, Mook-Kanamori DO, Niesters M, van Velzen M, Gademan MGJ, Rosendaal FR, Bouvy ML, Dahan A, Lijfering WM. **Causes and consequences of the opioid epidemic in the Netherlands: a populationbased cohort study.** Sci Rep. 2020 Sep 17;10(1):15309. doi: 10.1038/s41598-020-72084-6. PMID: 32943678; PMCID: PMC7499208.
3. Loef M, **Faquih TO**, von Hegedus JH, Ghorasaini M, Ioan-Facsinay A, Kroon FPB, Giera M, Kloppenburg M. **The lipid profile for the prediction of prednisolone treatment response in patients with inflammatory hand osteoarthritis: The HOPE study.** Osteoarthr Cartil Open. 2021 Apr 22;3(4):100167. doi: 10.1016/j.ocarto.2021.100167. PMID: 36474761; PMCID: PMC9718086.
4. DiCorpo D, LeClair J, Cole JB, Sarnowski C, Ahmadizar F, Bielak LF, Blokstra A, Bottinger EP, Chaker L, Chen YI, Chen Y, de Vries PS, **Faquih T**, Ghanbari M, Gudmundsdottir V, Guo X, Hasbani NR, Ibi D, Ikram MA, Kavousi M, Leonard HL, Leong A, Mercader JM, Morrison AC, Nadkarni GN, Nalls MA, Noordam R, Preuss M, Smith JA, Trompet S, Vissink P, Yao J, Zhao W, Boerwinkle E, Goodarzi MO, Gudnason V, Jukema JW, Kardia SLR, Loos RJF, Liu CT, Manning AK, Mook-Kanamori D, Pankow JS, Picavet HSJ, Sattar N, Simonsick EM, Verschuren WMM, Willems van Dijk K, Florez JC, Rotter JI, Meigs JB, Dupuis J, Udler MS. **Type 2 Diabetes Partitioned Polygenic Scores Associate With Disease Outcomes in 454,193 Individuals Across 13 Cohorts.** Diabetes Care. 2022 Mar 1;45(3):674-683. doi: 10.2337/dc21-1395. PMID: 35085396; PMCID: PMC8918228.

Articles under preparation/Revision

1. **Faquih T**, Willems van Dijk K, van Hylckama Vlieg A., Mook-Kanamori DO., et al. **Hepatic triglyceride content is intricately associated with numerous metabolites and biochemical pathways.** Under revision. 2022.

2. **Faquih T**, Willems van Dijk K, van Hylckama Vlieg A., Mook-Kanamori DO., et al. **Robust metabolomic age prediction based on a wide selection of metabolites.** Under submission.

3. **Faquih T**[1], Landstra, E.N.*, et al. **PFAS concentrations are associated with an unfavorable cardio-metabolic risk profile: findings from two population cohorts.** Under submission.

4. **Faquih, T***, Imtiaz, M.A.*, et al. **Genome Wide Association on Missing Metabolite Measurements.** Under submission.

1 Shared first co-authorship

PHD PORTFOLIO

	Year	Hours
Mandatory courses		
- Basic Methods and Reasoning in Biostatistics (done)	2019	1.5
- PhD Introductory Meeting (including the workshop Scientific Conduct for PhDs) (done)	2019	5
- BROK Course (exempted)		
Generic/disciplinary courses		
- Epidemiology “An Introduction”	2018	3
- Meta-analyse 2019	2019	1
- ONLINE - Survival analysis (Advanced Biostatistics) 2021	2021	1.5
- Analysis of Repeated Measurements	2021	1.5
- Statistical Aspects of Clinical Trials 2022	2022	1
- Writing An Excellent Grant Proposal	2022	1
- International Course on Clinical Epidemiology	2019	2
- Prediction modelling and Intervention research	2019	3
- Causal Inference	2020	3
- Genetics in Drug Development	2022	0.5
Attended lectures, LUMC presentations, participation in meetings		
- Human Genetics work discussion	2019	
Congress attendance and poster or oral presentations		
- WEON 2022 conference ‘The art of Epidemiology’	2022	1.5
- Multiomics to Mechanisms - Challenges in Data Integration Symposium	2019	0.5
- The Dutch Society for Sleep-Wake Research (NSWO) and The Slaapgeneeskunde Vereniging Nederland (SVNL)	2021	0.5
TEACHING ACTIVITIES		
Lecturing, lab assistance, student supervision		
- AWV2: Academic and Scientific Training 2nd year medical students	2019	0.5
- AWV2: Academic and Scientific Training 2nd year medical students	2020	0.5
- Mendelian Randomization practicum/workshop	2019	0.5
- CRIP Clinical Research in Practice	2020	0.5
- CRIP Clinical Research in Practice	2019	0.5
- Critical Appraisal of a Topic Project	2020	1
- Design and Analysis of Biomedical Studies	2020	0.5
- Design and Analysis of Biomedical Studies	2021	0.5
TOTAL number of hours		**31**