Statistical modelling strategies in molecular epidemiology, with an application to attention-deficit hyperactivity disorder

Ville Karhunen

Thesis submitted for the degree of Doctorate of Philosophy

Department of Epidemiology and Biostatistics

Imperial College London

September 2020

Declaration of originality

I hereby state that the work presented in this thesis is my own. Any other work conducted by collaborators or other scientists is appropriately detailed or referenced.

Copyright

Abstract

Since 1990s, the technological developments in measuring molecular data have been instrumental in advancing molecular epidemiology. A consequential challenge is to integrate the evidence from multiple molecular datasets for a better understanding of the biological mechanisms underlying complex traits.

In this thesis, I applied different statistical modelling approaches to investigate the biological background of attention-deficit/hyperactivity disorder (ADHD), a common neurodevelopmental disorder with an early onset, high persistence and a notable impact on the global burden of disease. I examined the putative impact of exposure to maternal smoking during pregnancy on the risk of ADHD and other adverse outcomes in the offspring related to epigenetic modifications and other molecular changes. The potential causality between ADHD and obesity was analysed using genetically informative methods. The link between systemic chronic inflammation, which can be triggered by smoking or obesity, and common psychiatric outcomes was also investigated.

The main dataset used was the Northern Finland Birth Cohort 1986 (N = 6,728 for ADHD symptoms, N = 432 for phenotype and full omics data available), with complementary data from other European cohorts and publicly available summary statistics. I used integrative statistical approaches that leverage evidence from different omics datasets. Regression modelling and Mendelian Randomisation techniques were applied throughout, and a recently published network method based on sparse canonical correlation analysis was also used.

The results showed evidence for a long-term impact of intrauterine smoke exposure on offspring DNA methylation, and some indication that DNA methylation mediates the effect of the smoke exposure on offspring later life health outcomes. There was also suggestive bidirectional causality between ADHD and obesity, and evidence for an inflammatory component in the aetiology of psychiatric outcomes. This thesis adds to the literature by a thorough investigation of different omics datasets and integrative statistical approaches applied to ADHD and other psychiatric outcomes.

Acknowledgements

This work was carried out within CAPICE (Childhood and Adolescence Psychopathology: unravelling the complex etiology by a large Interdisciplinary Collaboration in Europe) EU project under the Marie Skłodowska-Curie grant agreement no. 721567. Thanks to the funding body, the fellow PhD students within this project for the shared experiences, and to all other members of this team. I have been priviledged to be a part of this group.

I want to express my sincere thanks to my supervisors Prof Marjo-Riitta Järvelin, Dr Marina Evangelou and Prof Alina Rodriguez for their kindness, patience and expertise in supervising me. They have taught me valuable lessons on conducting science and working in academia, for which I will always be extremely grateful.

I want to thank all participants in the study cohorts and the corresponding data management teams, obviously without them this scientific work would have never been possible. Thanks to my earlier colleagues in Oulu who encouraged me to pursue a career in science, the colleagues and staff at Imperial who welcomed me in the department, and the office mates in VC11/12 for the enjoyable working environment.

I want to thank all the co-authors that I have collaborated with. Special thanks to Petri, and fellow PhD students Andrea, Tom, and Dipender, who have motivated me to be a better scientist.

Finally, I want to thank Charlotte, my friends – you know who you are – and my family for their support. Kiitos.

Funding

This work was carried out within CAPICE EU project under the Marie Skłodowska-Curie grant agreement no. 721567.

Northern Finland Birth Cohorts 1966 and 1986 have received core support from multiple funders. The data generation, curation and manpower concerning current work are primary supported by the following grants: the Academy of Finland: Grant no. 285547 EGEA; University Hospital Oulu, Finland: Grant no. 75617, the EU FP5 EURO-BLCS, QLG1-CT-2000-0164 (genetic/epigenetic data generation); NIH, USA: NIHM MH063706 (for Smalley and Järvelin for ADHD data generation); ERDF European Regional Development Fund Grant no. 539/2010 A31592 (epigenetic data generation); the EU H2020–PHC-2014 DynaHEALTH action: Grant no. 633595; EU-H2020 LifeCycle Action: Grant no. 733206; EU-H2020 EDCMET: Grant no. 825762; the Medical Research Council, UK: Grant no. MR/M013138/1, MRC/BBSRC MR/S03658X/1 (JPI HDHL H2020).

Contents

List of Tables

List of Figures

Abbreviations

ADHD	Attention-deficit/hyperactivity disorder
AHRR	Aryl-hydrocarbon receptor repressor
ALSPAC	Avon longitudinal study of parents and children
ApoBtoApoA1	Ratio of apolipoprotein B to apolipoprotein A1
ARIES	Accessible resource for integrated epigenomic studies
ASD	Autism spectrum disorder
BFP	Body fat percentage
BMI	Body mass index
BMR	Basal metabolic rate
BPD	Bipolar disorder
CCA	Canonical correlation analysis
CI	Confidence interval
CPACOR	Control probe adjustment and reduction of global correlation
CpG	Cytosine-phosphate-guanine
CV	Cross-validation
DAG	Directed acyclic graph
DNA	Deoxyribonucleic acid
DSM-5	Diagnostic and statistical manual of mental disorders, 5th edition
EWAS	Epigenome-wide association study
FDR	False discovery rate
FWER	Family-wise error rate
GIANT	Genetic investigation of antrhopometric traits
GLM	Generalised linear model
GNG12	Guanine nucleotide binding protein, gamma 12
GTEx	Genotype-tissue expression
GWAS	Genome-wide association study
HRC	Haplotype reference consortium
HWE	Hardy-Weinberg equilibrium
IBD	Identical by descent
InSIDE	Instrument strength independent of direct effect
IOWBC	Isle of Wight birth cohort
IQR	Interquartile range
IVW	Inverse-variance weighted
KEGG	Kyoto encyclopedia of genes and genomes

LD	Linkage disequilibrium
MAF	Minor allele frequency
MDD	Major depressive disorder
ML	Maximum likelihood
MR	Mendelian randomisation
NFBC	Northern Finland birth cohort
NMR	Nuclear magnetic resonance
OR	Odds ratio
PC	Principal component
PCA	Principal component analysis
PCR	Principal component regression
PGC	Psychiatric Genomics Consortium
PMD	Penalised matrix decomposition
PRESSO	Pleiotropy residual sum and outlier
PRS	Polygenic risk score
QQ	Quantile-quantile
QTL	Quantitative trait loci
RCT	Randomised controlled trial
RMSE	Root mean squared error
RNA	Ribonucleic acid
RSS	Residual sum of squares
sCD40L	soluble cluster of differentiation 40 ligand
SCOPA	Software for correlated phenotype analysis
SD	Standard deviation
SDQ	Strengths and difficulties questionnaire
SE	Standard error
SES	Socioeconomic status
sICAM1	soluble vascular cell adhesion molecule-1
SmCCNet	Sparse multiple canonical correlation network analysis
SNP	Single nucleotide polymorphism
SWAN	Strengths and weaknesses of ADHD symptoms and normal behaviour
WC	Waist circumference
WHR	Waist-hip-ratio
YFS	Cardiovascular risk in young Finns study

Chapter 1

Introduction, overall aim and scope

Mental illnesses are large contributors to the global burden of disease, and they are estimated to account for a third of total years lived with disability worldwide (Vigo et al., 2016). A large contributor to this burden is attention-deficit/hyperactivity disorder (ADHD), a neurodevelopmental disorder with early onset and high persistence (Faraone et al., 2015). Improving the understanding of the aetiology of ADHD and other mental disorders is of great importance from a public health perspective.

Technological developments since 1990s have enabled a cost-effective production of large sets of molecular data. In addition, the increasing trend of global and collaborative scientific effort has led to a notable increase in sample sizes that are available. These developments have revolutionalised traditional epidemiological research and enabled large-scale studies to identify biomarkers associated with health outcomes.

The current overwhelming abundance of data highlights the importance of adequate use of statistical methods for the research questions of interest. Different statistical modelling strategies are required for different research questions. The correct use and interpretation of the methods and approaches applied ensures valid interpretation and inference of the results. Moreover, integrative approaches that combine data and results across multiple molecular datasets are crucial for improving the understanding of disease aetiology.

The overall aim of this thesis was to examine the molecular aetiology of ADHD via different statistical modelling strategies in high-dimensional datasets. Specific attention is given to analytical approaches across large molecular datasets and causal inference based on observational data. ADHD serves as a phenotype for applying these approaches in psychiatric epidemiology, and is discussed on a more generic level.

In Chapter 2, I provide background information on molecular epidemiology, give a short narrative literature review on the epidemiology of ADHD, and outline the research gaps and main objectives. Chapter 3 describes the datasets used in this thesis. The statistical methods and analytical approahces used are presented in Chapter 4. I present the main results for the research objectives in Chapters 5, 6, 7, 8 and 9. Overall discussion and conclusions are given in Chapter 10.

Chapter 2

Background

This chapter outlines the main background information for this thesis. In Sections 2.1 and 2.2, I introduce the main concepts for genetic and other molecular epidemiology, and briefly outline the opportunities and challenges in analysing molecular datasets.

In Section 2.3, a short narrative literature review on the epidemiology of ADHD is provided. Research gaps and the main objectives are stated in sections 2.4 and 2.5, respectively.

2.1 Genetic epidemiology

Genetic epidemiology is the study of the role of genetic factors, both on their own and together with environmental factors, on health outcomes in populations. The field emerged in the 1950s, and up to 1990s, the focus in the field was in family-based studies and linkage studies (Morton, 2006; Fallin et al., 2016). Family-based studies do not necessarily require *genotyping*, i.e. the full or partial measurement of an individual's DNA sequence. These studies are still largely applied, and they are able to estimate the *heritability* of an outcome, that is, the proportion of variance in the outcome explained by genetic factors. However, without genotyping, the specific genes involved in the aetiology of the outcome cannot be inferred. The first studies using genotyping were linkage studies, in which a limited number of genetic variants are genotyped, and genetic loci that are associated with the outcome of interest are inferred based on familial relationships (Dawn Teare & Barrett, 2005).

Towards the end of 20th century, genetic epidemiology started to recognise the utility of

genetic association studies using unrelated individuals (Risch & Merikangas, 1996). The first such approach was the candidate gene study design, where genetic variants selected for genotyping were based on *a priori* knowledge or hypothesis of the involvement of a gene in the aetiology of the outcome of interest (Tabor et al., 2002). Both linkage and candidate gene studies had limited success in finding specific genes affecting *complex traits*, i.e. phenotypes not conforming to Mendelian inheritance patterns (Hirschhorn & Daly, 2005).

Since the late 1990s, the development of cost-effective high-throughput technologies have paved the way for the emergence of *omics* fields – the comprehensive assessment of a specific set of molecules – including genomics, that is, the study of the complete set of genetic material (genome). A big breakthrough was the sequencing of the full human genome in 2001 (Lander et al., 2001). Consequently, the rise of genomics has shifted the focus of genetic epidemiology from family-based linkage studies and candidate gene studies to genome-wide studies of unrelated individuals, usually genotyped using single nucleotide polymorphism (SNP) arrays. These SNP arrays are designed to primarily detect common genetic variants, i.e. SNPs with minor allele frequency (MAF) over 1% or 5%. The need for large sample sizes was also recognised, which led to an increase of consortium framework in genetic epidemiology (Fallin et al., 2016).

The standard approach to analyse associations between a genotype and a phenotype of interest on a genome-wide level has been genome-wide association studies (GWAS), where a large set – typically up to millions – of SNPs are independently analysed for association with the phenotype. The main principles of GWAS are discussed in Section 4.2.1.1.

GWAS have identified thousands of common genetic variants associated with complex traits (Visscher et al., 2017). For example, GWAS catalog – an open-access online repository for GWAS summary statistics (`https://www.ebi.ac.uk/gwas/`) – reported "71,673 variant-trait associations" across 5,687 GWAS summary statistics as of September 2018 (Buniello et al., 2018).

What has been evident from results across all genetic studies of complex traits is their polygenicity, and that the effect sizes of invididual variants are typically small compared to the effect sizes for monogenic diseases. Moreover, for most complex traits, there is still a gap between the GWAS-identified variants and the underlying biology. Accordingly, the current 'post-GWAS' era of genetic epidemiology is increasingly interested in downstream gene products, discussed in the next section (Gallagher & Chen-Plotkin, 2018).

2.2 Other molecular epidemiology

Besides genomics, other omics datasets include epigenomics, transcriptomics, proteomics and metabolomics. Analyses of these omics datasets aim to improve the understanding of how gene expression, genetic regulation and gene products interact with each other and the environment and how they are mechanistically linked to health outcomes.

Figure 2.1 shows the interactive nature of the omics datasets. The genetic information saved in DNA is transmitted to RNA in transcription. Epigenetic changes may affect the transcription. The information carried by RNA transcripts is then translated into proteins. Metabolites are products of cellular metabolic functions. The omics layers are dynamic across each other, the environment (*exposome*) and the set of all phenotypes, that is, the *phenome* (Haas et al., 2017; Vrijheid, 2014; Wild, 2005). There is also gene–environment correlation, that is, an association between individual's genotype and their environmental conditions.

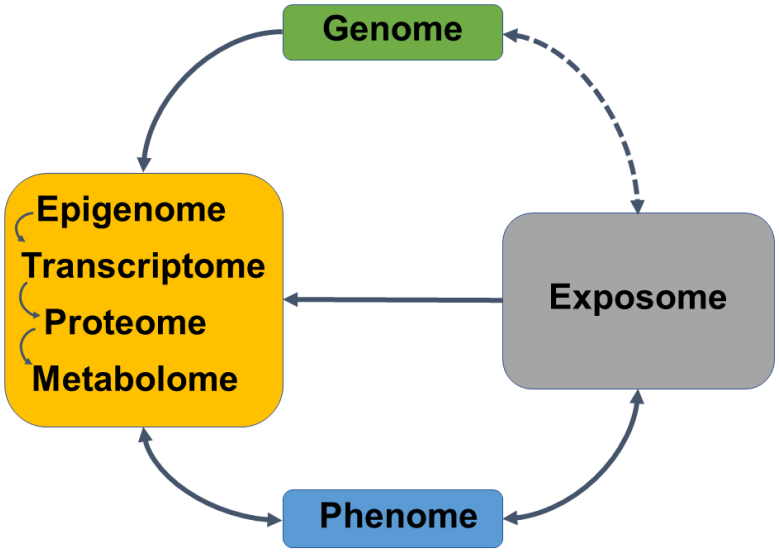

Figure 2.1: Schematic figure of the interplay of different omics. Thick lines represent the interplay between different omics, exposome and phenome. Thin lines represent the flow of genetic information within molecular omics. Dashed line represents the correlation between genetic and environmental factors.

2.2.1 Epigenomics

Epigenomics is the study of epigenetics, i.e. reversible modifications of DNA (or DNA-associated proteins), on a genome-wide scale. Epigenetic mechanisms include DNA methylation, histone modification and various RNA-related processes. Epigenetic processes are thought to play a role in gene expression (Gibney & Nolan, 2010), and are suggested to be a potential mechanism by which genetic and/or environmental factors affect disease outcomes (Relton & Davey Smith, 2010). The epigenome depends on both genetic and environmental factors (Bernstein et al., 2007). Epigenetic patterns are also tissue-specific and potentially dynamic over time (Lokk et al., 2014). Thus, in contrast to genetic sequence variation, epigenetic risk can be not only modifiable, but also reversible.

The most studied epigenetic mechanism is DNA methylation, where a methyl group is attached to either cytosine or adenine in the DNA strand. The more common of these is the methylation of cytosines at cytosine-phosphate-guanine (CpG) sites as a mechanism of gene regulation. In this thesis, DNA methylation is the only considered epigenetic mechanism.

Due to the available technology, differential DNA methylation at CpG sites across the full epigenome has become increasingly measured in population-based studies. In particular, methods based on bisulfite conversion have proven to be robust and accurate methods for epigenome-wide quantification of the proportion of methylated cells at each CpG site (Sandoval et al., 2011; Teschendorff & Relton, 2018). Analogously to GWAS, epigenome-wide association studies (EWAS) have emerged as a popular way to analyse the association between epigenetic variation (mainly DNA methylation) and phenotypes, on an epigenome-wide scale (Rakyan et al., 2011).

2.2.2 Transcriptomics, proteomics and metabolomics

Transcriptomics, proteomics and metabolomics are other omics layers that, in addition to genomics and epigenomics, complement the molecular landscape of human biology. Transcriptomics examines RNA transcripts from DNA on a genome-wide level. Proteomics is a study of all proteins expressed by a genome (Sun & Hu, 2016). Proteins are encoded by the genome and are responsible for performing many biological functions of living organisms. Compared to transcriptomics, proteomics data are expected to be more proximal to disease mechanisms (Hasin et al., 2017).

Metabolites are molecular products of cellular metabolic functions, and they serve as direct signatures of biochemical activity (Patti et al., 2012). Metabolomics is the study of a comprehensive set of metabolites. Metabolome is the first omics layer not directly encoded in the genome, and thus presents the omics layer that is the most proximal to the phenotype (Haas et al., 2017).

2.2.3 Opportunities and challenges

The technological advances that allowed the quantification of different omics have enabled a great opportunity to improve the knowledge on complex trait aetiology. There is now a possibility for enhanced biomarker discovery and a chance to conduct detailed analysis on putative disease mechanisms.

The vast amount of omics data also imposes analytical challenges. The omics datasets are typically high-dimensional, and 'agnostic' in the sense that the omics data are measured without prior knowledge in relation to the phenotype(s) of interest. In addition, it is important to be able to combine the information from different omics levels in a meaningful way for a convergence of a better explanation of complex trait biology. The essential methods for analysing omics data and integrative approaches for combining information across multiple omics datasets are presented in Section 4.2.

Another key challenge for omics data, that also applies to any data in general, is how to separate correlation from causation. Eliciting a true cause of a phenotype from a mere correlate is fundamental especially in aetiological research. Associations between omics variables and the phenotype(s) of interest may arise from complex biological and environmental phenomena that are not necessarily on the same causal pathway. In the abscence of causality, interventions for non-causal omics correlates to prevent an undesirable phenotypic outcome are likely to be futile. The challenge of causal inference, especially using observational data, is tackled in detail in Section 4.3.

These opportunities and challenges for omics data analysis, integrative approaches and causal inference are universal and can be applied to virtually any phenotype of interest. The next section gives an introduction to attention-deficit/hyperactivity disorder, the main phenotype used throughout this thesis.

2.3 Attention-deficit/hyperactivity disorder (ADHD)

Attention-deficit/hyperactivity disorder (ADHD) is described by DSM-5 (The Diagnostic and Statistical Manual of Mental Disorders, 5th edition (American Psychiatric Association, 2013)) as a disorder characterised by persistent patterns of inattention, impulsivity and/or hyperactivity, which interfere with functioning or development. Inattention refers to a poor attention span not due to defiance or lack of comprehension, hyperactivity indicates excessive and inappropriate motor activity, and impulsivity refers to impetuous actions without forethought (Table 2.1). For an ADHD diagnosis, the symptoms are required to have been persistent for at least six months to an extent which is abnormal for the developmental level and harmful for social and educational activities, prior to age of 12 years.

For the narrative literature review in this section, I used PubMed and Google Scholar databases with the following search terms: “ADHD”, “BMI”, “chronic inflammation”, “comorbidity”, “continuous trait”, “early life”, “epidemiology”, “genetic correlation”, “genetics”, “global burden”, “GWAS”, “heritability”, “inflamm*”, “maternal”, “mortality”, “neuroinflamm*”, “obesity”, “outcomes”, “parental”, “prevalence”, “psychiatric disorders”, “sex difference”, “smok*”, “subthreshold”. The focus was on reviews and meta-analyses from 2000s, the quality of which was judged based on the reporting guidelines for meta-analysis (Moher et al., 2009). The list was amended with other key relevant papers, based on the reporting guidelines of observational studies (von Elm et al., 2007) that were conducted on a large population-wide scale using registry data or of sufficient sample size (>5,000), or using Northern Finland Birth Cohort 1986 (Section 3.1).

Table 2.1: Key symptoms of ADHD.

Inattention	**Hyperactivity/impulsivity**
Does not give close attention to details, makes careless mistakes	Fidgets with or taps hands or feet, or squirms in seat
Has difficulty in sustaining attention on tasks or activities	Leaves seat when staying seated is expected
Does not seem to listen when spoken to directly	Runs about or climbs when not appropriate (for adolescents and adults: feeling restless)
Does not follow through instructions or finish homework, chores or workplace duties	Unable to play or engage in leisure activities quietly
Has difficulty organising tasks or activities	Is constantly 'on the go' or acting as if 'driven by a motor'
Avoids, dislikes or is reluctant to do tasks requiring sustained mental effort	Talks excessively
Loses things that are needed for tasks or activities	Blurts out an answer before a question has been finished
Is easily distracted	Has difficulty waiting for their turn
Is forgetful in daily activities	Interrupts or intrudes on others

2.3.1 Prevalence, screening and burden of disease

In a meta-analysis by Polanczyk et al. (2007), the prevalence of ADHD in children and adolescents was estimated at around 5% worldwide. A follow-up analysis (Polanczyk et al., 2014) found no evidence of increase in the worldwide prevalence since mid-1980s. In general, there is high variability in the prevalence estimates for ADHD (Polanczyk et al., 2007). The esimated prevalence among adolescents (mean age 16 years) in the Northern Finland Birth Cohort 1986 – the main data source in this thesis, Section 3.1 – was 8.5% (95% confidence interval [CI] 4.5% to 15.1%) (Smalley et al., 2007).

The prevalence of ADHD decreases with age, and is estimated at around 2.5% within adults according to a meta-analysis by Simon et al. (2009). Although the majority of childhood and adolescent ADHD cases do not fulfill the diagnostic criteria in adulthood, the persistence of symptoms is common (Biederman et al., 2000; Faraone et al., 2006).

The gold standard to diagnose ADHD is by a thorough clinical interview conducted by a professional. Despite the unavoidably categorical nature of the diagnosis, evidence suggests that ADHD can be considered as the tail of a continuous range of symptoms (Balázs & Keresztény, 2014; Demontis et al., 2019; Hawi et al., 2015; Larsson et al., 2012). In population-based studies, questionnaires such as Rutter (Rutter, 1967), SDQ (Strengths and Difficulties Questionnaire) (Goodman, 1997) or SWAN (Strengths and Weaknesses of ADHD-symptoms and Normal-Behaviour) (Swanson et al., 2012) are used as screening in-

struments for ADHD symptoms. These questionnaires are known for their association with clinical diagnoses (Goodman, 1997), and they reflect the continuous nature of the symptoms underlying the ADHD diagnosis. The symptoms can be assessed by parents or other informants (including relatives and teachers) in childhood and additionally via self-report in adolescence (Faraone et al., 2015).

ADHD has a profound impact on the affected individual and their families (Sayal et al., 2018). Erskine et al. (2014) measured the global burden of ADHD and report the disorder contributing to almost 500,000 disability-adjusted life years as a measure of healthy life years lost. Le et al. (2014) investigated societal costs of ADHD in the Netherlands and estimated them to be around 1 billion euros per year for a country with a population of around 16 million.

2.3.2 Genetics

ADHD is highly heritable, with heritability estimates from twin studies at around 75% (Faraone et al., 2005; Franke et al., 2012; Lichtenstein et al., 2010). The candidate gene studies of ADHD mostly involved genes from dopaminergic or adrenergic pathways targeted by ADHD medications (Faraone & Larsson, 2019). In their meta-analysis of candidate gene studies on ADHD, Gizer et al. (2009) detected evidence for association with several candidate genes across the investigated pathways. However, the results from candidate gene studies should be taken with caution, as there has been a lack of reliable replication, especially for behavioural and psychiatric traits (Munafò, 2006).

The early GWAS on ADHD – summarised by Franke et al. (2009) – did not discover any genome-wide significant variants. The largest GWAS on ADHD to date with 20,183 ADHD cases and 35,191 controls was conducted by Demontis et al. (2019), and the authors report 12 genome-wide significant loci for ADHD, and SNP heritability (i.e. heritability based on additive effects of common SNPs) at around 20%.

Results from family-based studies show evidence for shared genetic aetiology between ADHD other psychiatric traits, such as autism spectrum disorder (ASD) (Rommelse et al., 2010), bipolar disorder (BPD) (Faraone et al., 2012), major depressive disorder (MDD) (J. Cole et al., 2009) and schizophrenia (Larsson et al., 2013). Population-based studies have reported ADHD having genetic correlations with smoking, anthropometric traits (Demontis et al.,

2019), and other psychiatric traits (Anttila et al., 2018; P. H. Lee et al., 2019).

ADHD is more prevalent among males than females, with an estimated male-to-female ratio of prevalences at about 2.5 to 1 (Polanczyk et al., 2007). However, the ratio is smaller among adults (Rucklidge, 2010), and therefore the sex difference in ADHD is suggested to be partially due to differences in manifesting ADHD symptoms and the diagnosis process (Mowlem et al., 2019).

2.3.3 Parental and early-life factors

ADHD has been associated with a range of parental factors. Two largely replicated maternal exposures are maternal smoking during pregnancy and maternal obesity. A meta-analysis of exposure to intrauterine cigarette smoke exposure and ADHD by Huang et al. (2018) report an odds ratio (OR) of 1.60 (95% CI 1.45 to 1.76) for maternal smoking during pregnancy and offspring ADHD. In their systematic review and meta-analysis for maternal obesity and offspring ADHD, Li et al. (2020) report an increased risk of offspring ADHD for mothers with obesity, compared to normal-weight mothers (risk ratio 1.92, 95% CI 1.84 to 2.00).

Other parental or early life factors that have been associated with ADHD are maternal stress (Manzari et al., 2019), parental socioeconomic status (Russell et al., 2016), preterm birth (Serati et al., 2017), and low birth weight (Franz et al., 2018).

2.3.4 Smoking, obesity and inflammation

ADHD is associated with a range of adverse health conditions and negative later-life outcomes. Individuals diagnosed with ADHD are more likely to smoke (van Amsterdam et al., 2018) and have higher prevalence of obesity than non-diagnosed (Cortese & Tessari, 2017; Cortese, 2019).

Both smoking and obesity are known to trigger systemic chronic inflammation (Furman et al., 2019; Gonçalves et al., 2011; Gregor & Hotamisligil, 2011). Inflammation is a natural response by the immune system to an infection or irritation. Normal inflammatory response is acute, and the affected tissues eventually restore to their normal structural and functional state. In systemic chronic inflammation, this response fails to resolve (Nathan & Ding, 2010). Circulating levels of some proteins, such as cytokines, can be used as biomarkers for chronic

inflammation (Liu et al., 2017).

Systemic chronic inflammation has been associated with a range of social, biological and psychological stressors as well as a plethora of adverse health outcomes (Furman et al., 2019). There is evidence for comorbidity of ADHD with inflammatory and autoimmune disorders (G. A. Dunn et al., 2019).

2.3.5 Other comorbidities and later-life outcomes

There is also comorbidity between ADHD and other psychiatric disorders (Katzman et al., 2017). In a systematic review and meta-analysis by Erskine et al. (2016), the authors report positive associations between ADHD and other mental disorders, substance use disorders and criminality, and negative associations between ADHD and academic achievement and employment. In a nationwide Danish register study, Dalsgaard et al. (2015) showed that ADHD diagnosis was associated with premature mortality, driven by accidents and other unnatural causes. Subthreshold levels of ADHD symptoms are also associated with adverse outcomes (Balázs & Keresztény, 2014).

2.4 Research gaps

The putative molecular mechanisms underlying the associations between ADHD and its risk factors, comorbidities and consequences are largely unclear (Thapar & Cooper, 2016). Here, I discuss and summarise the main research gaps that are tackled in this thesis.

2.4.1 Causality of maternal smoking and obesity on offspring ADHD

The causality of maternal smoking and obesity on offspring ADHD is questioned. It is suggested that the fetal exposure to adverse intrauterine environment predisposes the offspring to chronic illnesses, including adverse mental health outcomes, known as the *developmental origins of health and disease* hypothesis (O'Donnell & Meaney, 2017).

Causality of both exposures is supported by animal models (Alkam et al., 2017; Menting et al., 2019; Zhu et al., 2012). The results from human studies are indecisive, as the results may

be due to common causes, such as genetic predisposition, for the maternal exposures and ADHD. In systematic reviews of family-based genetically informed studies on prenatal smoke exposure and ADHD by Rice et al. (2018) and Thapar & Rice (2020), the authors conclude that the evidence is not consistent with a causal explanation. On the other hand, Sourander et al. (2019) examined the association with maternal cotinine levels – a known biomarker for smoking – and offspring ADHD and found evidence for a dose–response relationship.

The association between maternal obesity and offspring ADHD is also suggested to be due to common familial factors (Huang et al., 2018; Li et al., 2020). However, a recent study on the associations between parental body mass index (BMI) and offspring ADHD showed that the associations were specific to maternal BMI (S. L. Robinson et al., 2020).

2.4.2 DNA methylation

In addition to an increased risk of ADHD, exposure to maternal smoking is associated with many adverse neurodevelopmental, respiratory and cardiometabolic outcomes later in life (Cupul-Uicab et al., 2012; Doherty et al., 2009; Hofhuis et al., 2003; Ng & Zelikoff, 2007; Power et al., 2010). One suggested mechanism for these associations is DNA methylation (Knopik et al., 2012; Nielsen et al., 2016; Parmar et al., 2018). Smoking is known to modify DNA methylation profile (Joehanes et al., 2016), and maternal smoking during pregnancy is associated with differential offspring DNA methylation in cord blood (Joubert et al., 2016).

There is evidence that DNA methylation mediates the effect of maternal smoking during pregnancy on offspring birth weight (Küpers et al., 2015). Such molecular mediation may underlie the risk of other offspring outcomes related to prenatal smoke exposure.

DNA methylation is suggested to play a role in the aetiology of mental health outcomes (Barker et al., 2018; Guintivano & Kaminsky, 2016), including ADHD (Hamza et al., 2019). There is a growing number of large-scale human studies examining the associations between epigenome-wide DNA methylation and ADHD (Mooney et al., 2020; van Dongen et al., 2019; Walton et al., 2017).

Questions remain whether maternal smoking related DNA methylation changes detected in cord blood are sustained into offspring's adolescence and adulthood. If these DNA methylation changes are persistent, then DNA methylation would be a plausible candidate for the mechanism leading to adverse health outcomes in the offspring.

Moreover, the link of these DNA methylation changes to other omics datasets is not known. Leveraging information across multiple omics datasets has the potential to reveal more subtle and complex relationships between the omics measurements and the risk of ADHD.

2.4.3 Causality between obesity and ADHD

The comorbidity of ADHD and obesity is well-known (Cortese, 2019). The double burden of both conditions is estimated to account for nearly a two-fold increase in the cost of care compared to individuals without either condition (Libutzki et al., 2019).

Unclarity remains whether there is a causal association between the traits. Based on their systematic review, Cortese & Tessari (2017) concluded that the evidence from observational studies supports a bidirectional relationship. However, similarly as for the intergenerational association between maternal obesity and offspring ADHD, shared familial causes – such as genetics – may also explain the comorbidity between the traits on an individual level (Q. Chen et al., 2013; Huang et al., 2018; Li et al., 2020).

A Mendelian Randomisation (Section 4.3.3) study by Martins-Silva et al. (2019) showed evidence for a causal effect of high BMI on an increased risk of ADHD, but not vice versa. However, the only obesity-related variable used was BMI, which does not fully reflect body fat or abdominal obesity, the latter being an important risk factor for disease development (Y. Chen et al., 2018), especially in children (Savva et al., 2000).

2.4.4 Inflammation and psychiatric outcomes

Accumulating evidence supports a neuroinflammatory component in psychiatric illnesses (Najjar et al., 2013). There is evidence for associations between inflammatory biomarkers and common neuropsychiatric outcomes (Yuan et al., 2019), such as ADHD (G. A. Dunn et al., 2019), ASD (Xu et al., 2015), BPD (Muneer, 2016), depression (Dowlati et al., 2010) and schizophrenia (Na et al., 2014).

However, smoking, obesity and medication have all been identified as potential common causes for the associations between inflammatory biomarkers and psychiatric disorders in observational studies (Jeon et al., 2019; Na et al., 2014). When assessing potential causality between chronic inflammation and psychiatric outcomes, methods with different sources of

bias can help in triangulation of the evidence (Section 4.3.5).

2.4.5 Summary of research gaps

To summarise, the following research gaps are highlighted:

Exposure to maternal smoking and offspring DNA methylation: It is not known whether the associations between exposure to maternal smoking during pregnancy and differential offspring cord blood DNA methylation endure into the offspring's adolescence and adulthood.

DNA methylation as a mechanism for disease outcomes: It is not known whether differential DNA methylation acts as a mechanism for the observational associations between exposure to maternal smoking during pregnancy and adverse offspring health outcomes.

Multi-omics analysis of ADHD: To date, no integrative analysis across multiple omics types for the association between intrauterine smoke exposure and offspring ADHD have been conducted.

ADHD and obesity: Unclarity remains of the direction of the putative causal association between the traits and the potential role of prenatal environment for this association.

Chronic inflammation and psychiatric outcomes: Potential causality of chronic inflammation on the risk of psychiatric outcomes needs to be more rigorously evaluated.

2.5 Aims and objectives of the thesis

The overall aim of this thesis was to examine the molecular background of ADHD by incorporating information across multiple omics datasets and integrating the evidence from these data sources using different statistical modelling strategies. The objectives were as follows:

1. To examine the association between exposure to maternal smoking during pregnancy and offspring DNA methylation in adolescence and adulthood.
2. To assess the potential mediating role of DNA methylation in the association between exposure to maternal smoking during pregnancy and offspring ADHD as well as other later life adverse health outcomes.

3. To apply advanced integrative prediction modelling and variable selection methods across multiple omics datasets for an investigation of their relationships with ADHD.

4. To investigate the comorbidity and potential causality between ADHD and obesity.

5. To analyse the effect of circulating inflammatory biomarker levels on the risk of psychiatric outcomes.

I address these objectives in five studies, presented in Chapters from 5 to 9 (Table 2.2), in which the corresponding research hypotheses are also stated. The persistence of differential DNA methylation associated with exposure to maternal smoking during pregnancy, and potential mediation for offspring adverse health outcomes is examined in Study I (Chapter 5). In Study II (Chapter 6), the focus is specifically on the mediation on offspring ADHD symptoms. In Study III (Chapter 7), the associations between omics variables and ADHD are examined via different modelling strategies and integrative approaches. Study IV (Chapter 8) examines the comorbidity of ADHD and obesity via genetically informed methods. The potential causality of chronic inflammation on the risk of psychiatric outcomes is assessed in Study V (Chapter 9).

Table 2.2: The studies in this thesis.

Study	Title	Objectives addressed	Chapter
I	Maternal smoking and offspring DNA methylation	1, 2	6
II	Exposure to maternal smoking, offspring DNA methylation and ADHD symptoms	2	7
III	Multi-omics and ADHD	3	8
IV	Causality between ADHD and obesity-related traits	4	9
V	Circulating cytokine levels and psychiatric outcomes	5	10

Chapter 3

Datasets

This chapter summarises the datasets used in this thesis. I give detailed descriptions on Northern Finland Birth Cohorts (NFBC) 1986 and 1966, which were the main datasets used throughout the thesis. NFBC1986 includes phenotypic data on ADHD symptoms at multiple time points by different raters, and the dataset has been widely used in psychiatric research (Miettunen et al., 2019).

In addition, other contributing cohorts and data sources used in this thesis are briefly described. The data sources for each objective are presented in Table 3.1. Detailed data quantification and quality control steps are beyond the scope of this work, but are presented in short here.

Table 3.1: Data sources for each objective.

Study	Datasets	Phenotype	Omics
I	NFBC1986, NFBC1966, ALSPACm, ALSPACc, IOWBC	Offspring DNA methylation at 16 to 48 years; GWAS summary statistics	Epigenomics, genomics
II	NFBC1986, NFBC1966, ALSPACm, ALSPACc, IOWBC	ADHD symptoms at 16 years (NFBC1986); GWAS summary statistics	Epigenomics, genomics
III	NFBC1986	ADHD symptoms at 16 years (NFBC1986); GWAS summary statistics	Epigenomics, genomics, metabolic measures transcriptomics*
IV	NFBC1986	ADHD symptoms at 8 and 16 years (NFBC1986); GWAS summary statistics	Genomics
V	NFBC1966, YFS, FINRISK	GWAS summary statistics	Genomics, protein measures transcriptomics*

NFBC = Northern Finland Birth Cohort; ALSPAC = Avon Longitudinal Study of Parents and Children; IOWBC = Isle of Wight Birth Cohort; YFS = Cardiovascular Risk in Young Finns Study.
* Transcriptomic information incorporated from GTEx database.

3.1 Northern Finland Birth Cohort 1986

Northern Finland Birth Cohort 1986 (NFBC1986) targeted all pregnant women in two northernmost provinces in Finland (Oulu and Lapland) with expected date of delivery between 1st July 1985 and 30th June 1986 (`http://www.oulu.fi/nfbc`) (Northern Finland Cohorts, 2020). The inclusion criterion for the study was that pregnancy had lasted for at least 24 weeks. In total, 9,432 children were live-born to 9,362 mothers (Järvelin et al., 1993). This covered 99% of all deliveries within the region during the target period.

Data collection started during antenatal visits, on average from 12th gestational week onwards. Larger follow-ups have been conducted via postal questionnaires at offspring ages of seven, eight and 16 years, and via a clinical examination for the offspring at 16 years. Data on ADHD symptoms have been collected at 8-year and 16-year follow-ups with different informants. The Ethical Committee of Northern Ostrobothnia Hospital District approved the study, and a written informed consent was provided by parents and adolescents at the 16-year follow-up (Northern Finland Cohorts, 2020). Figure 3.1 shows the flowchart for the NFBC1986 datasets used in this thesis.

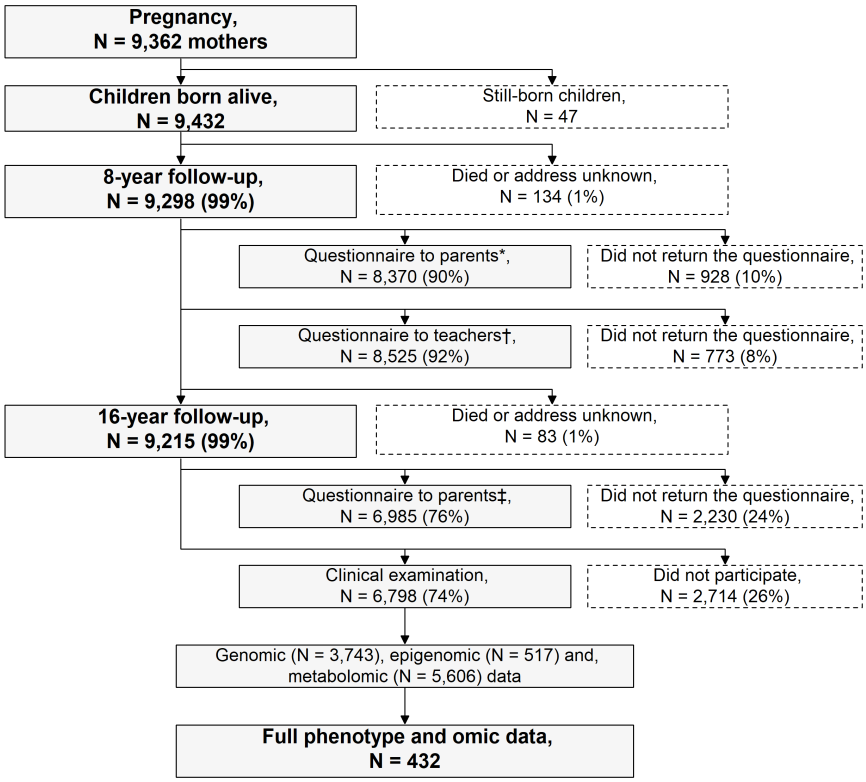

Figure 3.1: Flowchart of the data collections in the Northern Finland Birth Cohort 1986 used in this study. * = Includes data on offspring ADHD symptoms assessed on Rutter A scale. † = Includes data on offspring ADHD symptoms assessed on Rutter B scale. ‡ = Includes data on offspring ADHD symptoms assessed using Strengths and Weaknesses of ADHD symptoms and Normal behaviour (SWAN) scale (N = 6,728).

3.1.1 Perinatal data collection

Background information on mothers – and, to a lesser extent, fathers – was obtained via three separate questionnaires from 24th week of pregnancy onwards until delivery. These questionnaires included questions on parity, pre-pregnancy BMI, parental occupational status and maternal educational level. An indicator of family socioeconomic status (SES) was constructed based on the prestige of paternal occupational status (Kotimaa et al., 2003). If paternal information was missing, maternal occupational status was used.

Maternal smoking habits were enquired using questions on smoking before and during pregnancy, duration of smoking, number of cigarettes smoked per day and change of smoking habits during pregnancy. Exposure to maternal smoking during pregnancy was defined as positive if the mother was still smoking at least one cigarette per day from pregnancy week 8 onwards. Paternal smoking habits were enquired by questions on the smoking status and the amount of smoking of the father during pregnancy.

3.1.2 8-year data collection

In Finland, children start their school in the autumn of the calendar year when they turn seven. A two-part data collection was conducted in NFBC1986 during the children's first school year, first in autumn, and the second in the spring. The second part is referred to here as the 8-year data collection. Parents for children living in Finland and with known address (N = 9,297) received two questionnaires, one for themselves and the other to be forwarded to the children's teachers. Both questionnaires included questions on children's inattention and hyperactivity symptoms. Parents used the Rutter A scale to evaluate the children's behaviour (Rutter, 1967), while teachers used the Rutter B2 scale (Rutter, 1967).

3.1.3 16-year data collection

A 16-year data collection was conducted between April 2001 and June 2002. All individuals of the cohort with known address (N = 9,215) received a postal questionnaire on health and lifestyle and invitation to a clinical examination. In addition, parents received a questionnaire for family background information and to assess the offspring's health, development and behaviour.

Parents used the Strengths and Weaknesses of ADHD symptoms and Normal behaviour (SWAN) scale (Swanson et al., 2012) to rate the offspring's ADHD symptoms. SWAN measures both weaknesses and strengths, and is expected to produce a Gaussian distribution for the inattention and hyperactivity-impulsivity symptoms included in the DSM-5 classification. Parent-assessed ADHD symptoms were available for 6,728 adolescents. The clinical examination included anthropometric measures and blood sample collection, and the data were received for 6,795 (74%) study participants.

3.1.3.1 Omics data

NFBC1986 includes genomic, epigenomic and metabolomic data, all based on the blood sample collection at the 16-year follow-up. For the genomic data, a total of 3,834 samples were genotyped using Illumina HumanOmniExpressExome-8v1.2 platform and Beadstudio calling algorithm. After excluding individuals with low call rate (< 0.95), low mean heterozygosity (< 0.305), related individuals (identical by descent [IBD] pairwise sharing < 0.2), gender mismatch or duplicate samples, genotype data were available for 3,743 adolescents. This comprises of a random sample of 3,371 individuals, and a selected set of 372 individuals exposed to gestational diabetes, gestational hypertensive disorders and preterm birth.

Poor quality SNPs were excluded based on call rate (<0.99) and Hardy-Weinberg equilibrium (HWE, $p < 1 \times 10^{-4}$). After these exclusions, 889,119 SNPs remained in the genotyped dataset. The genetic data were imputed using Haplotype Refererence Consortium (HRC) imputation reference panel (Loh et al., 2016). The imputed dataset was then filtered based on imputation quality (excluded SNPs with imputation $r^2 < 0.5$, MAF < 0.001 and HWE $p < 10^{-12}$), and the final genetic data comprised of 11,009,294 SNPs.

For epigenomic data, DNA methylation was quantified for 546 randomly selected individuals with the most extensive questionnaire and clinical data available and that were part of the genotyped random sample. The quantification was conducted using Illumina HumanMethylation450 array according to manufacturer's instructions. Bisulfite conversion of genomic DNA was performed using the EZ DNA methylation kit according to the manufacturer's instructions (Zymo Research, Orange, CA).

Quality control for DNA methylation data was adapted from the CPACOR (Control Probe Adjustment and reduction of global CORrelation) pipeline (Lehne et al., 2015). Individuals with low call rate (< 0.95) or gender mismatch in DNA methylation data were removed. For the CpGs, Illumina Background Correction was applied to the raw intensity values and a detection threshold was set at p-value $= 10^{-16}$. Quantile normalisation was done separately for six probe-type categories, and these normalised intensity values were used to calculate the methylation beta value at each CpG site, ranging between 0 (no methylation) and 1 (full methylation). CpG sites with call rate < 0.95 were excluded. The final DNA methylation dataset consists of 466,290 CpGs for 517 individuals.

High-throughput nuclear magnetic resonance (NMR) metabolomics was used to quantify 228 metabolic measures by the Nightingale platform (Nightingale Health Ltd, Helsinki, Finland). The list of metabolic measures includes a total of 149 lipid measurements, lipoprotein subclass particle concentrations, abundant fatty acids, amino acids, ketones and glycolysis-related metabolites, and a total of 79 lipid compositions of 14 lipoprotein subclasses or fatty acid ratios (Table A.1). The measurement method is described in Soininen et al. (2015). NMR metabolomics data were available for 5,606 individuals.

3.2 Northern Finland Birth Cohort 1966

Northern Finland Birth Cohort 1966 (NFBC1966) targeted all pregnant women with expected date of delivery in the year 1966 within the study area (Rantakallio, 1969). Information was collected from mothers from 24th gestational week onwards and from clinical maternity data on average 16th gestational week onwards. Overall, the cohort included 12,055 mothers with 12,058 live-born children. This comprises of 96% of all births during 1966 in the study area. A small percentage of births occurred late 1965 or early 1967. Large follow-up data collections for the offspring have been conducted at 14, 31 and 46 years. Ethical approval was received from Ethical Committee of Northern Osthrobothnia Hospital District and the University of Oulu (Northern Finland Cohorts, 2020).

Parental background information and smoking behaviour was enquired with similar questions as in NFBC1986. Both 31-year and 46-year data collections included extensive questionnaires and clinical examinations.

NFBC1966 includes multiple omics data from two time points. For this thesis, I used genetic, metabolomic and proteomic data quantified at 31 years, and DNA methylation data quantified at both 31 and 46 years.

In 1997, at offspring age of 31 years, all cohort participants with known addresses were sent a postal questionnaire on health and lifestyle and those living in Northern Finland or Helsinki area were invited to a clinical examination which included blood sampling. In total, both questionnaire and clinical data were collected for 6,007 participants.

DNA was successfully extracted for 5,753 participants from fasted blood samples (Sovio et al., 2009), and genotyping was conducted using Illumina HumanCNV-370DUO Analysis

BeadChip (Illumina, California, USA). Samples were removed based on low call rate (<0.95), gender mismatch, duplicate samples, high relatedness, outlying heterozygosity, sample contamination and consent withdrawal, with 5,402 individuals in the cleaned dataset. Poor quality SNPs were removed based on low MAF (<0.05) and high missingness (>0.05). Additionally, SNPs with MAF < 0.01 and HWE p-value $< 10^{-4}$ were removed.

The remaining 364,535 SNPs were used for imputation by HRC imputation reference panel. After excluding SNPs with imputation $r^2 < 0.5$, MAF < 0.001 and HWE p-value $< 10^{-12}$), 10,928,335 SNPs remained in the dataset. After two further consent withdrawals, genetic data were available for 5,400 individuals.

DNA methylation at 31 years was quantified for 807 randomly selected subjects of whom both questionnaire and clinical data with cardio-metabolic measures were available at both 31 and 46 years. The quantification and quality control procedures for DNA methylation data were exactly the same as for NFBC1986. Final DNA methylation dataset for NFBC1966 at 31 years includes 459,378 CpGs for 717 individuals.

As for NFBC1986, NMR metabolomics was used to quantify 228 metabolic measures in NFBC1966 at 31 years. These data were available for 5,709 individuals, out of which 692 individuals had also DNA methylation data available.

NFBC1966 contains measurements of 16 proteins, comprising of cytokines, chemokines and growth factors. The proteins were quantified from overnight fasting plasma samples taken at the 31-year data collection, using Bio337 Rad's Bio-Plex 200 system (Bio-Rad Laboratories, California, USA) with Milliplex Human Chemokine/Cytokine and CVD/Cytokine kits (Cat# HCYTOMAG-60K-12 and Cat# SPR349; Millipore, St Charles, Missouri, USA) and Bio-Plex Manager Software V.4.3 as described in Saukkonen et al. (2018). These data were available for 5,199 individuals with genotype data also available.

In 2012, all individuals with known address in Finland were sent postal questionnaires and an invitation for clinical examination. Both questionnaire and clinical data were obtained for 5,539 participants. DNA methylation data at 46 years were extracted for 766 subjects for whom DNA methylation was measured at the 31-year data collection. DNA methylation was quantified by Illumina EPIC array and the quality control steps were exactly the same as for NFBC1986 and NFBC1966-31-year DNA methylation data. This DNA methylation data were available for 832,569 CpGs in 716 individuals.

3.3 Avon Longitudinal Study of Parents and Children

Avon Longitudinal Study of Parents and Children (ALSPAC) started in the early 1990s, when all pregnant women resident in the former county of Avon, UK, with expected dates of delivery between 1st April 1991 and 31st December 1992 were invited to take part in the study. Initially, 14,541 pregnant women enrolled in the study, and there were 14,062 live births in total. Both mothers (ALSPACm) and their children (ALSPACc) have been followed up longitudinally (Boyd et al., 2012; Fraser et al., 2012). Ethical approvals were obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees.

Summarised ALSPAC data were used in Studies I (Chapter 5), II (Chapter 6) and III (Chapter 7), based on the Accessible Resource for Integrated Epigenomic Studies (ARIES) subsample of ALSPAC. This subsample includes 1,018 mothers and their children for whom DNA methylation data has been quantified, selected based on the availability of DNA samples at two time points for the women (antenatal with mean age of 30 years, and offspring adolescent follow-up with mean age of 48 years), and three time points for offspring (neonatal, childhood with mean age 7.5 years, and adolescence with mean age 17.1 years) (Relton et al., 2015). DNA methylation wet-laboratory and pre-processing analyses are described in Relton et al. (2015).

3.4 Isle of Wight Birth Cohort

Isle of Wight Birth Cohort (IOWBC) is a population-based pregnancy-birth cohort recruited on the Isle of Wight, United Kingdom, in 1989 (Arshad et al., 2018). Ethics approvals were obtained from the Isle of Wight Local Research Ethics Committee (now named the National Research Ethics Service, NRES Committee South Central-Southampton B) at recruitment and for the 1, 2, 4, 10, and 18-years follow-up.

Summarised IOWBC data were used in Studies I (Chapter 5) and II (Chapter 6). Maternal smoking status in pregnancy was self-reported and defined as any smoking in pregnancy or no smoking during pregnancy. DNA methylation in peripheral blood samples was analysed from 257 randomly selected study participants at the 18-year follow-up. Quantification of DNA methylation was performed using Illumina HumanMethylation450 array and the

quality control steps were the same as in NFBC cohorts. The cleaned dataset comprised of 461,230 CpGs.

3.5 The Cardiovascular Risk in Young Finns Study

The Cardiovascular Risk in Young Finns study (YFS) is a population-based follow-up study started in 1980 (Raitakari et al., 2008). YFS comprises of randomly chosen individuals from Finnish cities Helsinki, Kuopio, Tampere, Oulu and Turku, and their rural surroundings. Subsequent follow-up visits involving all five centres have been conducted in 1983, 1986, 1989, 2001, 2007, 2011 and 2017.

I used cytokine GWAS summary statistics from YFS in Study V (Chapter 9). Genotyping was conducted for 2,556 study participants based on the blood samples drawn at the 2001 follow-up. Genotyping was performed using a custom-built Illumina 670K array. After sample quality control (missingness > 0.05 or relatedness IBD > 0.20 removed) and SNP quality control (missingness > 0.05, HWE p-value $< 10^{-6}$ excluded), the dataset included 2,443 individuals and 546,674 genetic variants. The imputation for genotype data was performed with IMPUTE2 software by using 1000 Genomes Phase 3 release as reference panel. After imputation, poorly imputed and rare variants (info score < 0.7 and minor allele count < 3) were removed.

Circulating cytokine concentrations were quantified by Biorad's Bio-Plex Pro Human Cytokine 27-plex Assay and 21-plex Assay for 48 cytokines from serum samples drawn at 2007 follow-up visit as previously described (Santalahti et al., 2016). Depending on the cytokine, imputed genotypes and cytokine concentrations were available for 116 to 2,019 samples.

3.6 FINRISK

FINRISK surveys are population-based cross-sectional studies in Finland (Borodulin et al., 2017). The studies started at 1972, and a new sample – a random selection of individuals between 25 and 74 years from five georgraphical areas in Finland – is recruited every five years.

I used cytokine GWAS summary statistics from FINRISK in Study V (Chapter 9). Cytokine

quantification for FINRISK1997 and FINRISK2002 samples was performed analogously as in YFS, with the difference that quantification was done using EDTA plasma in FINRISK1997 and heparin plasma in FINRISK2002 (Santalahti et al., 2017). In FINRISK1997, a custom 20-plex array was used in cytokine quantification, and in FINRISK2002, only participants between 51 and 74 years were selected for the analysis. Imputation for genotype data was performed using 1000 Genomes Phase 3 as reference panel. Poorly imputed and rare variants (info score < 0.7, minor allele count < 3) were excluded. Depending on the cytokine, imputed genotypes and cytokine concentrations were available for 3,440 to 4,613 samples from FINRISK1997 and 843 to 1,705 samples from FINRISK2002.

3.7 External databases

I also used external publicly available databases in my thesis. MR-Base platform (available at `http://www.mrbase.org/`) provides an easy access to GWAS summary statistics for a large range of phenotypes (Hemani, Zheng, et al., 2018). MR-Base was used as a source for GWAS summary statistics in Study I (Chapter 5).

Psychiatric Genomics Consortium (PGC) is a collaborative effort for genomic analyses of psychiatric disorders and provides GWAS summary statistics for a range of phenotypes (P. F. Sullivan et al., 2018). PGC GWAS summary statistics for ADHD were used in Studies II (Chapter 6), IV (Chapter 8) and V (Chapter 9), and summary statistics for ASD, BPD, MDD and schizophrenia were used in Study V (Chapter 9).

In Study IV (Chapter 8), GWAS summary statistics for anthropometric measures were also used, obtained from GIANT (Genetic Investigation of ANthropometric Traits) consortium (Shungin et al., 2015; Yengo, Sidorenko, et al., 2018), and from Neale Lab online repository for biobank-wide high-throughput GWAS summary statistics (the Neale Lab, 2018).

For transcriptomic information, the Genotype-Tissue Expression (GTEx) project provides association statistics for GWAS of gene expression across 53 human tissues (GTEx Consortium, 2017). These summary statistics were used in Studies III (Chapter 7) and V (Chapter 9).

Chapter 4

Statistical methods and analytical approaches

This chapter outlines the methods and analytical approaches used in this thesis. Section 4.1 covers the main statistical methods, with more specific methods described in detail in the corresponding study chapters. In Section 4.2, I outline the basics of omics data analysis. Causal inference, especially based on observational data, is discussed in detail in Section 4.3.

When deciding on which analytical approaches to use, it is crucial to pay attention to the research question at hand, as this determines how a statistical model is developed and applied (Arnold et al., 2020; Shmueli, 2010). In this thesis, I focus on three different modelling strategies: *causal inference*, *prediction modelling* and *variable selection*. Each of these are essentially different tasks and thus require distinct processes for statistical model development and interpretation.

In causal inference, the aim is to conclude whether an exposure causes an outcome, which is fundamental in epidemiology for examining the aetiology of complex traits. In this thesis, causal inference is the most prevalent modelling strategy and is discussed in detail in Section 4.3.

The aim of prediction modelling is to provide predictions of events or conditions for individuals at risk. This is important in medical research and public health to aid decision

making for actionable screening or diagnosis. Prediction modelling can also be used for research purposes by first predicting a phenotype based on multiple variables, and using the predicted values of the prediction model as an input to a subsequent statistical model. A prime example of this are polygenic risk scores (PRS), discussed in Section 4.2.1.1, which are predicted phenotype values based on the effects of multiple genetic variants, and are typically used in association analyses to proxy the polygenic liability to a phenotype (Choi et al., 2020). Prediction models may also be helpful in targeting preventive interventions to individuals at a high risk of having or developing a disease or a condition (Steyerberg, 2019).

In Study II (Chapter 6), I construct a risk score for exposure to maternal smoking during pregnancy based on offspring DNA methylation. This risk score is used in statistical analysis to quantify epigenome-wide differential DNA methylation due to intrauterine smoke exposure. In Study III (Chapter 7), prediction modelling is applied for offspring ADHD symptoms, based on omics variables related to intrauterine smoke exposure. In Study IV (Chapter 8), PRS calculated for BMI and ADHD are used in association analysis.

Variable selection (or feature selection) aims to provide insight to the individual and/or joint relationship of explanatory variables and the outcome, and to identify key influential variables (Heinze et al., 2018; Sauerbrei et al., 2020). These aims are more modest than for prediction modelling or causal inference. Nevertheless, the value of variable selection is evident during the time of high-throughput technologies that enable quantification of a large number of variables across different omics. As the data are usually collected without prior biological knowledge with respect to the phenotype of interest, this fact gives an opportunity to screen omics variables related to the phenotype using a data-driven approach. Particularly, all omics-wide association studies, such as GWAS and EWAS mentioned in Chapter 2 and discussed in more detail in Section 4.2.1.1, can be considered as variable selection studies.

Variable selection is applied in Study I (Chapter 5) for examining the stability of differential DNA methylation of exposure to maternal smoking during pregnancy in adolescents and adults. In Study II (Chapter 6), association analysis is conducted for DNA methylation markers related to intrauterine smoke exposure with ADHD symptoms as the outcome. Variable selection methods leveraging information across multiple omics datasets (Section 4.2.2.2) is applied in Study III (Chapter 7). Finally, in Study V (Chapter 9), GWAS

are conducted for circualting protein levels. Table 4.1 shows the breakdown of different modelling strategies for each study in this thesis.

Table 4.1: Modelling strategies within each objective.

Study	Question	Strategy
I	CpGs associated with exposure to maternal smoking	Variable selection
	Exposure to maternal smoking → DNA methylation → Offspring outcomes	Causal inference
II	CpGs associated with ADHD	Variable selection
	Exposure to maternal smoking predicted by DNA methylation	Prediction modelling
	Exposure to maternal smoking → DNA methylation → ADHD	Causal inference
III	ADHD symptoms predicted by multi-omics	Prediction modelling
	Multi-omics variables associated with ADHD symptoms	Variable selection
IV	BMI↔ADHD	Causal inference
	Polygenic risk scores for ADHD and BMI	Prediction modelling
	Maternal BMI → ADHD	Causal inference
V	Genetic variants associated with circulating cytokine levels	Variable selection
	Cytokines → psychiatric outcomes	Causal inference

Unless otherwise noted, $\mathbf{Y} = (y_1 \dots y_n)^T$ is the outcome variable measured for n observations, $\mathbf{X} = (\mathbf{1}\ \mathbf{x}_1 \cdots \mathbf{x}_p) \in \mathbb{R}^{n\times(p+1)}$ is a data matrix where $\mathbf{1} = (1 \cdots 1)^T \in \mathbb{R}^{n\times 1}$ and $\mathbf{x}_j = (x_{1j} \cdots x_{nj})^T,\ j = 2, \dots, p$ are the observed values for explanatory variables, and $\boldsymbol{\beta} = (\beta_0\ \beta_1\ \cdots\ \beta_p) \in \mathbb{R}^{p+1}$ is a vector of coefficients, and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0\ \hat{\beta}_1\ \cdots\ \hat{\beta}_p)$ their estimated values. It is also assumed that observations $y_i, \dots, y_n$ are independent with an identical distribution.

4.1 Statistical methods

Statistical learning, i.e. application of statistical methods to gain knowledge of data, is in heart of all quantitative research. Statistical learning can be divided to supervised and unsupervised learning. In supervised learning, there exists an outcome variable of interest, which is modelled in relation to explanatory variables. In unsupervised learning there is no specific outcome, and modelling is done to find interrelationships across all variables. I present the basics of regression modelling for supervised learning (Section 4.1.1), and dimension reduction for unsupervised learning (Section 4.1.2).

4.1.1 Supervised learning

Whenever there is a specific outcome variable of interest, supervised learning aims to explore its distribution as a function of explanatory variables. Regression modelling is an ubiquitous

supervised method to model the conditional distribution of an outcome variable, given a set of explanatory variables.

4.1.1.1 Generalised linear models

Generalised linear models (GLMs) encompass a large range of different regression models for supervised learning. Let $\mathbf{Y}$ be observations of a random variable which is assumed a probability distribution that belongs to the exponential family of distributions (McCullagh, 1989). A GLM has a structure of

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

where $\boldsymbol{\mu} = \mathbb{E}(\mathbf{Y}|\mathbf{X})$ is the expected value of outcome $\mathbf{Y}$ given explanatory variables $\mathbf{X}$, g is the *link function* (so that $g^{-1}(\boldsymbol{\eta}) = \boldsymbol{\mu}$) and $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ is the *linear predictor*.

The exponential family covers a large range of probability distributions and they can be used in a range of statistical tasks that deal with causal inference, prediction modelling and variable selection. The general assumptions in GLMs are that the probability distribution, link function and the linear predictor are specified correctly. The 'correct' specification of the linear predictor crucially depends on the research question. The interpretation of a coefficient β_j is the change in the linear predictor $\boldsymbol{\eta}$ due to a unit change in the corresponding explanatory variable $\mathbf{x}_j$ while holding the values of other explanatory variables $\mathbf{x}_{k\neq j}$ constant.

A traditional linear model for a continuous outcome is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I}). \tag{4.1}$$

This can be presented as a GLM, with $\boldsymbol{\mu} = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ and g is the identity function.

A common model for binary outcomes is logistic regression:

$$\mathbb{P}(\mathbf{Y} = 1|\mathbf{X}) = \frac{1}{1 + \exp\{-\mathbf{X}\boldsymbol{\beta}\}}. \tag{4.2}$$

The model (4.2) is a GLM with $\mu_i = \frac{1}{1+\exp\{-\eta_i\}}$ and $g(\mu_i) = \log(\frac{\mu_i}{1-\mu_i}),\ i = 1, \ldots, n$. A com-

mon model for an ordinal outcome with $k+1$ ordered categories $0, 1, \dots, k$ is a proportional odds model:

$$\mathbb{P}(y_i \geq j|\mathbf{X}) = \frac{1}{1+\exp\{-(\beta_{0j} + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p)\}}, \tag{4.3}$$

where $i = 1, \dots, n$, $j = 1, \dots, k$ and the intercept terms β_{0j} are modelled separately for each j. Model (4.3) corresponds to a GLM with $g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$ as above, $\mu_{ij} = \frac{1}{1+\exp\{-\eta_{ij}\}}$ and a different linear predictor for each j:

$$\eta_{ij} = -(\beta_{0j} + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p),\ i = 1, \dots, n.$$

Inverse normal rank transformation In many instances for quantitative outcomes, such as for some biological measurements or behavioural outcomes based on questionnaires, the distribution of the outcome may be highly skewed. In these cases, a transformation to ensure the correct distributional assumptions of the linear model (4.1) is needed. Inverse normal rank transformation (INRT) is a monotonic transformation based on the quantile function for a Gaussian distribution:

$$\text{INRT}(y_i) = \Phi^{-1}\left(\frac{y^{(i)} - 0.5}{n}\right),\ i = 1, \dots, n,$$

where $y^{(i)}$ is the sample rank of ith observation, and Φ^{-1} is the probit function, i.e. the quantile function for a Gaussian distribution. The distribution of the resulting variable is Gaussian, which helps in ensuring that the distribution of the residuals in equation (4.1) is also Gaussian. This transformation is widely used in GWAS (Section 4.2.1.1) to account for skewed distributions of the phenotypes (McCaw et al., 2019).

Modelling non-linearities: splines Linear regression is, by definition, linear in its parameters. Sometimes, a non-linear relationship between any explanatory variable and the outcome is to be expected. For example, there is evidence for a J-shaped association between maternal obesity and offspring ADHD symptoms (Rodriguez, 2010). One way to take the non-linear relationship into account is via spline functions, or splines. They are piecewise polynomial functions, which are connected at prespecified knots $t_1, \dots, t_k$.

Restricted cubic splines are third-order polynomials that are constrained to be linear before the first and after the last knot. They are found to have good properties in estimating the non-linear relationship between two variables (Wood, 2006). Following the formulation by Royston & Parmar (2002), a restricted cubic spline function with k knots $(t_1, \dots, t_k)$ is given by

$$f(x) = \xi_0 + \xi_1 x + \sum_{j=2}^{k-1} \xi_j \big((x - t_j)_+^3 - \lambda_j (x - t_1)_+^3 - (1 - \lambda_j)(x - t_k)_+^3 \big), \tag{4.4}$$

where

$$\lambda_j = \frac{t_k - t_j}{t_k - t_1},\ j = 2, \dots, k-1 \ \text{ and } \ u_+^3 = u^3 \text{ if } u > 0, \text{ and 0 otherwise.}$$

Clearly, the function given in equation (4.4) reduces to linear when $x < t_1$ or $x > t_k$. To allow a non-linear relationship between an explanatory variable $\mathbf{x}_j$ and the outcome, the constructed non-linear terms

$$\mathbf{x}'_j = (\mathbf{x} - \mathbf{t}_j)_+^3 - \lambda_j (\mathbf{x} - \mathbf{t}_1)_+^3 - (1 - \lambda_j)(\mathbf{x} - \mathbf{t}_k)_+^3,\ j = 2, \dots, k-1$$

from equation (4.4) are added as columns to the matrix of explanatory variables $\mathbf{X}$ (Harrell et al., 1988). Typically, a suitable number of knots is between 3 and 7 (Harrell, 2015). The position of the knots is typically pre-specified, e.g. for three knots, Harrell (2015) recommends to place them at quantiles $0.1, 0.5$ and 0.9.

Measures for regression model goodness-of-fit A key way to measure the goodness-of-fit of a regression model is to compare the model fitted values (i.e. model predictions) $\hat{\mathbf{Y}} = (\hat{y}_1, \cdots, \hat{y}_n)^T = \mathbf{X}\hat{\boldsymbol{\beta}}$ to the observed values $\mathbf{Y}$. For a traditional linear model (4.1), a natural metric for model goodness-of-fit is the variance explained:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{(n-1)\mathrm{Var}(\mathbf{Y})}.$$

Larger R^2 implies better fit, and $R^2 < 0$ if the fitted values provide on average less accurate fitted values than a constant mean fitted value $\bar{Y}$ (Harrell, 2015). Another common model goodness-of-fit metric is root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

Smaller values imply better fit. RMSE is on the same scale as the original outcome – for example, if $\text{SD}(\mathbf{Y}) = 1$, then RMSE of 1 would imply that the model fitted values are no better than chance.

In logistic regression (4.2), the comparison is made between the fitted probabilities $\hat{\mathbf{P}} = (\hat{p}_i \cdots \hat{p}_n)^T$ (where $\hat{p}_i = \hat{\mu}_i$) and the outcome values. *Brier score* is a score function that measures the accuracy of fitted probabilities for binary data. It corresponds to the mean squared error of the model fitted values $\hat{p}_i$ for a binary outcome y_i:

$$\text{Brier score} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{p}_i)^2$$

The value for Brier score ranges between 0 and 1, and lower score indicates better fit.

C index is the probability that a randomly selected observation with the outcome has a higher fitted probability than an observation without the outcome. It measures the model's ability to correctly discriminate cases and non-cases. It ranges between 0.5 and 1, and higher values imply better discrimination.

A common way to measure the goodness-of-fit of a logistic regression model is to categorise the fitted probabilities $\hat{p}$ to cases or non-cases based on a cut-off, typically 0.5, and then construct a misclassification table based on these classified values. However, such dichotomisation of fitted probabilities results in loss of information and potentially inadequate model comparison (Harrell, 2015; Wynants et al., 2019).

Model assessment When a regression model is developed for predictive purposes, the specific interest is how well the predictions given by the model perform in predicting observations in external data. Using the same dataset to evaluate the performance tends to produce overly optimistic performance estimates. The aim of model assessment (or model validation) is to measure the predictive performance in an external dataset.

An intuitive way to obtain model performance in an unseen data is data-splitting. The data is split to a model training set and a model test set, the model parameters are estimated in the training set, model performance is evaluated in the test set based on the estimated

model. Metrics presented in Section 4.1.1.1 can be used for model performance evaluation in the test set. In repeated data-splitting, this procedure is repeated multiple times, and this procedure gives more accurate measures for the model predictive ability than a single split (Kuhn & Johnson, 2013).

It should be noted that the model assessment evaluates the performance of the whole model-fitting procedure for unseen observations. The final model should always be fit using the full dataset.

Maximum likelihood estimation Maximum likelihood (ML) estimation is a general technique to estimate parameters for a statistical model (such as regression models) and for conducting statistical inference on these parameters. Assume a probability model $f_{\boldsymbol{\Theta}}$ for each observation $y_i,\ i = 1, \dots, n$, dependent on observed values for explanatory variables $\mathbf{x}_i$ (i.e. ith row of $\mathbf{X}$) and model parameters $\boldsymbol{\Theta} = (\theta_1 \cdots \theta_r) \in \mathbb{R}^r$, measured for n independent observations. The ML estimate is a value $\hat{\boldsymbol{\Theta}}$ that maximises the likelihood function

$$L(\boldsymbol{\Theta}, \mathbf{Y}, \mathbf{X}) = \prod_{i=1}^{n} f_{\boldsymbol{\Theta}}(y_i|\mathbf{x}_i).$$

or, equivalently after a monotonic log-transformation, maximising the log-likelihood function

$$l(\boldsymbol{\Theta}, \mathbf{Y}, \mathbf{X}) = \log(L(\boldsymbol{\Theta}, \mathbf{Y}, \mathbf{X})) = \sum_{i=1}^{n} \log(f_{\boldsymbol{\Theta}}(y_i|\mathbf{x}_i)), \tag{4.5}$$

the latter being numerically easier to deal with. Further details on ML estimation are given in e.g. S. R. Cole et al. (2013).

Bias–variance trade-off In general, increasing the complexity of a regression model leads to better goodness-of-fit measures within the data at hand. However, such models tend to perform poorly in an external dataset. For a traditional linear model (4.1), the expected squared error of the estimated regression model $\hat{f}_{\boldsymbol{\Theta}}$ depending on parameters $\boldsymbol{\Theta}$ can be decomposed as follows (Friedman et al., 2009):

$$\begin{aligned} \mathbb{E}\{(\mathbf{Y} - \hat{f}_{\boldsymbol{\Theta}}(\mathbf{X}))^2\} &= \sigma^2 + (f_{\boldsymbol{\Theta}} - \mathbb{E}\{\hat{f}_{\boldsymbol{\Theta}}\})^2 + \mathrm{Var}(\hat{f}_{\boldsymbol{\Theta}}) \\ &= \text{Irreducible error} + \text{Bias}^2 + \text{Variance}, \end{aligned} \tag{4.6}$$

where $f_{\boldsymbol{\Theta}}$ is the true (but unknown) regression function. Based on (4.6), there are two strategies to reduce the expected squared error: to reduce bias or to reduce variance. It turns out that in some instances, introducing bias decreases variance and consequently the total error. This is called the *bias–variance trade-off*, and also illuminates the reasoning between the different modelling strategies. The same principle applies to other GLMs.

The implications of the bias–variance trade-off are evident when considering the different modelling strategies. For causal inference, the aim is to obtain an unbiased estimate for a specific coefficient β_j. This has implications on the selection of explanatory variables, which will be presented in more detail in Section 4.3.2.

For prediction modelling, the aim is to maximally explain the variance in the outcome. This is analogous to reducing the total error, so the decrease in variance for the cost of adding bias is sometimes desirable. This biased estimation is discussed below. Compared to causal inference, there is less emphasis on the estimated values of coefficients $\hat{\boldsymbol{\beta}}$ for the measured explanatory variables, or predictors. Identifying relevant explanatory variables are still of importance, and it is crucial to consider various biases if observational data is used (Section 4.3.1) (W. R. Robinson et al., 2019).

Variable selection also aims at reducing bias in (4.6). Causality is not considered in a formal manner, however there is typically an implicit causal aim and thus the selection of explanatory variables is similar to causal inference (Section 4.3.2). However, especially in high-dimensional data, variable selection can benefit from biased coefficients if they provide sparsity in the coefficients, as discussed below.

Penalised regression An issue that is encountered with high-dimensional omics datasets is that if the number of variables p is 'too large' in relation to the number of observations n, traditional ML becomes unstable, and, when $p > n$, unidentifiable. However, penalised regression methods, i.e. regression methods based on penalised ML, can deal with this situation. In addition, in penalised regression, the individual coefficients β_j are intentionally biased by including a penalty that shrinks the individual coefficients towards zero, which – as described in Section 4.1.1.1 – may reduce the variance and thus reduce the total expected squared error. Penalised methods are therefore useful primarily for prediction models, and, if the method enables shrinkage of some coefficients to zero, can also be applied for variable selection.

In penalised regression, instead of maximising the log-likelihood function in (4.5), the function to be maximised is

$$l(\boldsymbol{\Theta}, \mathbf{Y}, \mathbf{X}) - P_\lambda(\boldsymbol{\beta}), \tag{4.7}$$

where $\boldsymbol{\beta} \in \boldsymbol{\Theta}$ are the regression coefficients to be penalised, and P is a penalty function depending on parameter $\lambda \geq 0$. The parameter λ is called a *tuning* (or *regularisation*) parameter, which controls the amount of shrinkage from the ML estimates towards zero.

Not all coefficients $\boldsymbol{\beta}$ need to be penalised. In GLMs, typically the intercept term β_0 is not penalised, and in some cases we might not want to shrink coefficients for certain variables at all. In such cases, the corresponding coefficients β_j can be omitted from the input to P_λ.

Typical choices for the penalty term are based on L_1 and L_2 norms. Setting the penalty term as

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \beta_j^2 \tag{4.8}$$

leads to *ridge* regression, based on L_2 norm penalty (Hoerl & Kennard, 1970). A popular alternative to ridge regression is a penalty based on L_1 norm:

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j| \tag{4.9}$$

This corresponds to *lasso* (least absolute shrinkage and selection operator) regression (Tibshirani, 1996), and it has a property of shrinking some coefficients to zero. This is advantageous in situations where such sparsity is desirable, such as variable selection. However, it should be noted that in the case of $p > n$, lasso can only select a maximum of n variables with non-zero coefficients.

In case of multicollinearity, where multiple explanatory variables are highly correlated with each other, lasso tends to be unstable in selecting variables with non-zero coefficients (Friedman et al., 2009). A compromise of L_1 and L_2 penalties is elastic net penalisation (Zou & Hastie, 2005):

$$P_\lambda(\boldsymbol{\beta}) = \alpha\lambda \sum_{j=1}^{p} |\beta_j| + (1-\alpha)\lambda \sum_{j=1}^{p} \beta_j^2 \tag{4.10}$$

This is a combination of ridge and lasso regressions with an additional parameter α that controls the mixing of the penalties (4.8) and (4.9). Ridge and lasso penalties are special cases of (4.10) when $\alpha = 0$ and $\alpha = 1$, respectively.

Selecting the tuning parameter An important part of applying penalised regression models is selecting the tuning parameter λ. It controls the amount of penalty in the model, with larger values increasing the complexity, and $\lambda = 0$ reducing to no penalisation. The aim is to select an optimal value for λ that has good properties for fitting the data at hand and predicting future observations.

For a grid of λ values, the tuning parameter can be estimated using *cross-validation* (CV). CV is a resampling method where a part of the observations is set aside for a *validation set*, and these observations are predicted using a model estimated using the remaining observations. This procedure is repeated so that each observation has been used as a validation set. In 10-fold CV, observations are divided into ten equal subsets at random, and each subset is used as a validation set in turn. The model is estimated using the other nine subsets, and model performance based on regression goodness-of-fit measures is assessed in the validation set. The 10-fold CV has satisfying properties regarding its bias and variance (Friedman et al., 2009). The minimum of (4.7) for a sequence of λ values can be found efficiently using coordinate descent algorithm, as implemented in R package `glmnet` (Friedman et al., 2010).

.

Hypothesis testing based on maximum likelihood Causal inference and variable selection are typically interested in whether there a specific coefficient $\beta_j \in \boldsymbol{\Theta}$ is different from zero. This can be evaluated via hypothesis testing. The methods for hypothesis testing presented in this section apply to GLMs, whereas hypothesis testing and inference based on penalised methods is beyond the scope of this thesis.

Suppose there are $q < r$ parameters of interest, and we wish to test a hypothesis

$$H_0 : \mathbf{Q}\boldsymbol{\Theta} = \mathbf{q}, \tag{4.11}$$

where $\mathbf{Q} \in \mathbb{R}^{q \times r}$ is a matrix of constraints and $\mathbf{q} \in \mathbb{R}^q$ is a vector of null hypothesis parameter values. Different test statistic can be constructed for testing, based on (4.5). Three classical tests are likelihood ratio test, Wald test and score test. It is known that the respective test statistics all have an asymptotic χ^2 distribution with q degrees of freedom. Whereas likelihood ratio test and score test are better for their statistical properties, the test statistic for Wald test is the easiest to compute, and is the only one presented here. The Wald test statistic W is defined as

$$W = (\mathbf{Q}\hat{\mathbf{\Theta}} - \mathbf{q})^T \mathbf{Q} I(\hat{\mathbf{\Theta}}) \mathbf{Q}^T (\mathbf{Q}\hat{\mathbf{\Theta}} - \mathbf{q}),$$

where

$$I(\mathbf{\Theta}) = -\frac{\partial^2 l}{\partial \mathbf{\Theta}^2}$$

is the information matrix and $\hat{\mathbf{\Theta}}$ is the ML estimate. For a single parameter $\hat{\theta}_j$, the square root of W reduces to

$$T = \sqrt{W} = \frac{\hat{\theta}_j}{\mathrm{SE}(\hat{\theta}_j)} \tag{4.12}$$

where

$$\mathrm{SE}(\hat{\theta}_j) = \sqrt{\frac{I^{-1}(\hat{\mathbf{\Theta}})_{jj}}{n}}$$

is the *standard error* of θ_j, and $T \sim N(0,1)$. For an estimated regression coefficient $\hat{\beta}_j$ in a GLM for a continuous outcome as in (4.1), the value of $I^{-1}(\hat{\mathbf{\Theta}})_{jj}$ depends on other estimated parameters, and thus $T \sim t_{n-r}$[1].

What is usually more interesting than a single hypothesis testing is the actual value of $\hat{\theta}_j$ and the uncertainty regarding the estimate. This can be quantified by using a confidence interval (CI) for the estimate. A $100(1-\alpha)\%$ CI for a single parameter $\hat{\theta}_j$ based on the Wald test can be constructed by

[1]Strictly speaking, this is a *t-test*, not a Wald test anymore.

$$\hat{\theta}_j \pm c_{\alpha/2}\text{SE}(\hat{\theta}_j), \tag{4.13}$$

where $c_{\alpha/2}$ is the $\frac{\alpha}{2}$th quantile of the corresponding distribution (i.e. either $N(0,1)$ or t_{n-r}).

Multiple testing Generally for any test statistic and its theoretical distribution, the probability of a test statistic being larger than its observed value under the null hypothesis H_0 is called the *p-value*. If the p-value is lower than a pre-specified threshold α, then this is generally inferred as evidence against H_0. A *false discovery* occurs when p-value $< \alpha$ even though H_0 is true.

A typical choice for α is 0.05. However, when dealing with large dimensions, this α level is very anti-conservative. Adequate controlling for multiple testing is important especially in high dimensions when multiple hypotheses are tested simultaneously. Two main ways to apply multiple testing correction is to either control the family-wise error rate (FWER), i.e. the probability of making at least one false discovery, or to control the false discovery rate (FDR), i.e. the expected number of false discoveries.

For M independent hypotheses with p-values $p_m, m = 1, \dots, M$, FWER is defined as

$$\text{FWER} = \mathbb{P}(1 - (p_m > \alpha \ \forall \ m) | \text{All } H_0 \text{ true}).$$

A common approach for FWER control is a *Bonferroni correction*, where the α is obtained by dividing 0.05 by the number of tests M.

FDR is defined as the proportion of false discoveries of all tests where $p_m < \alpha$:

$$\text{FDR} = \mathbb{E}\left(\frac{\mathbb{P}(p_m < \alpha | H_0 \text{ true})}{R} \middle| R > 0 \right) \mathbb{P}(R > 0),$$

where

$$R = \sum_{m=1}^{M} \mathbb{1}(p_m < \alpha).$$

Benjamini & Hochberg (1995) proposed a widely-used procedure correcting for FDR. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(M)}$ be the ordered set of p-values and $H_0^{(m)}, m = 1, \dots, M$ the corre-

sponding null hypotheses. Let k be the largest m for which $p_{(m)} \leq \frac{m}{M}\alpha$. Then, a discovery is declared (or null hypothesis is rejected) for all $H_0^{(i)}, i = 1, \ldots, k$. Equivalently, new FDR-corrected p-values $p_{\text{FDR}(m)} = p_{(m)}M/m,\ m = 1, \ldots, M$ can be compared to the chosen α level.

In general, at a given α, controlling for FWER provides more stringent control for false discoveries, while FDR correction have greater statistical power (i.e. observing $p < \alpha$ when H_0 is false) at the cost of increased chance of false discoveries.

Standard errors and confidence intervals by bootstrapping *Bootstrapping* (Efron & Tibshirani, 1986) is a general technique that conducts random sampling with replacement from the original data, in order to estimate parameters and their distributions. The sampling with replacement is conducted B times (typically $B \geq$ 1,000). Inference for the statistic of interest is made based on these B samples.

For example, standard error for an estimated parameter $\hat{\theta}_j$ can be constructed via bootstrapping as follows:

1. Take B samples with replacement from original data.
2. Within each $b = 1, \ldots, B$, calculate $\hat{\theta}_j$, denoted by $\hat{\theta}_j^b$
3. The bootstrap standard error for $\hat{\theta}_j$ is

$$\text{SE}(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_j^b - \bar{\theta}_j)^2},$$

where $\bar{\theta}_j = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_j^b$. This estimate for standard error can be plugged into (4.13) to obtain confidence intervals. Alternatively, a bootstrapped $100(1-\alpha)\%$ CI can be obtained directly by obtaining the $\frac{\alpha}{2}$th and $100 - \frac{\alpha}{2}$th quantiles of the empirical cumulative distribution of $\hat{\theta}_j^b$. The bootstrapped CI does not make any distributional assumptions, and thus is a good option especially in cases where there is doubt for the validity of asymptotic approximations. The main limitation of bootstrap is its computational cost.

Meta-analysis In many instances, the results for the association between a given exposure and outcome are available from multiple studies, and combining the results across the studies

increase statistical power to detect true associations. Meta-analysis methods can be used to pool regression coefficients across multiple studies. Let $\hat{\beta}_{jk}$ be the estimated regression coefficient for the association of explanatory variable $\mathbf{x}_{jk}$ and outcome Y in study $k, k = 1, \ldots, K$. An inverse-variance weighted *fixed-effects meta-analysis* estimate $\bar{\beta}_j$ is

$$\bar{\beta}_j = \frac{\sum_{k=1}^{K} w_k \hat{\beta}_{jk}}{\sum_{k=1}^{K} w_k}$$

with weights

$$w_k = \frac{1}{\mathrm{SE}(\hat{\beta}_{jk})^2}.$$

The fixed-effects meta-analysis aims for conditional inference based only on the results of the K studies at hand (Hedges & Vevea, 1998). Random-effects meta-analysis aims for a larger generalisation and allow for heterogeneity across the observed studies (Hedges & Vevea, 1998; Viechtbauer, 2010), and is not considered here.

4.1.2 Unsupervised learning

In unsupervised modelling, there is no specific outcome of interest. Instead, the aim is to capture the structure within the dataset across all variables.

4.1.2.1 Dimension reduction

Dimension reduction refers to the transformation of the original data to a smaller number of new variables constructed from the original data. Typically, the aim is to use a smaller set of $p' < p$ new variables, with p' selected so that there is no substantial loss of information.

Principal component analysis (PCA) is a common method for dimension reduction. In PCA, the original variables are presented by new variables that are linear combinations of the originals and orthogonal of each other. Let $\mathbf{X}$ be a matrix of rank $K \leq \min(n, p)$ and scaled so that the mean $\bar{\mathbf{x}}_j = 0$ and standard deviation $\mathrm{SD}(\mathbf{x}_j) = 1$. A singular value decomposition of $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T,$$

where $\mathbf{U}^T\mathbf{U} = \mathbf{I}_n$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}_p$, and $\mathbf{D}$ a diagonal matrix with diagonal elements $d_1 \geq d_2 \geq \cdots \geq d_K$. Define $\tilde{\mathbf{X}}$ as

$$\tilde{\mathbf{X}} = \mathbf{UD}. \tag{4.14}$$

The columns of $\tilde{\mathbf{X}}$ are the *principal components* (PCs) of $\mathbf{X}$. They are orthogonal and $\mathrm{Var}(\tilde{\mathbf{x}}_1) \geq \cdots \geq \mathrm{Var}(\tilde{\mathbf{x}}_K)$. The proportion of variance explained by the first $p' < K$ components is given by

$$V_{p'} = \frac{\sum_{i=1}^{p'} d_i}{\sum_{i=1}^{K} d_i}.$$

$V_{p'}$ can be used for determining the optimal p'.

4.2 Omics data analysis

In this section, I give an introduction for analysing molecular omics datasets. These datasets are characterised by a large number of measured variables and the correlated structure within datasets. In 4.2.1, analysis of a single omics dataset is considered, with an emphasis on omics-wide association studies.

In order to elicit as much information as possible regarding the underlying biology, it is vital to be able to combine information across multiple omics levels. The principles for such integrative approaches are presented in 4.2.2.

It should be noted that molecular omics data are generally highly sensitive to experimental conditions. Adequate pre-processing and quality control to remove measurement noise due to technical variation is a vital step for any reasonable omics data analysis. Here, the focus is on the analysis of pre-processed data, and in-depth quality control steps are beyond the scope of this work. Quality control procedures applied for the datasets used in this thesis were given in Chapter 3, and are discussed here in less detail.

4.2.1 Analysing single omics dataset

For supervised modelling of a single high-dimensional omics dataset with a large number of variables p (and disregarding any other potential explanatory variables for a moment), there are two distinct approaches: either univariate approach by analysing each $\mathbf{x}_j, j = 1, \ldots, p$ separately, or a multivariable approach by joint modelling of $\mathbf{X}$ with respect to the outcome $\mathbf{Y}$. When applying regression models, the approach taken crucially depends on the modelling strategy: the interpretation of the regression coefficients in the multivariable approach is conditional on the inclusion of other variables. Whereas this conditioning is less of a concern for prediction modelling, it might not be desirable for causal inference (Section 4.3.2). Variable selection can be conducted by either univariate approaches, or by multivariable approaches that also provide sparsity (Section 4.1.1.1). The basics of genome-wide and epigenome-wide association studies using the univariate approach are presented in Section 4.2.1.1.

Unsupervised methods can be applied to omics data to capture the joint variability in the datasets. For genomic data, PCA (Section 4.1.2.1) is a common method to measure population structure, that is, the differences in genetic variants' allele frequencies between subpopulations within a population. Some level of population structure exists in virtually any population, and there is known strong population structure in the NFBC cohorts (Cardon & Palmer, 2003; Sabatti et al., 2009). PCA is applied to the genomic data, and the resulting PCs can be used in supervised learning to capture the phenotypic variation due to population structure, discussed below and in Section 4.3.1.2.

For epigenomic data, unsupervised methods can be used to capture the technical variation due to varying experimental conditions. DNA methylation quantification arrays include technical control probes that measure the experimental conditions (Bibikova et al., 2011). In the quality control pipeline proposed by Lehne et al. (2015) and as applied to NFBC datasets (Chapter 3), a PCA is conducted for the control probes, and the resulting PCs are used in association analysis to account for the technical variation.

The unsupervised learning methods on omics datasets demonstrated here are auxiliary for supervised modelling, with the intention of capturing the variance not related to the research question at hand. In the next sections, I will provide the main approaches for supervised learning in molecular omics datasets.

4.2.1.1 Genome-wide association studies

As briefly mentioned in Section 2.1, genome-wide association studies (GWAS) have been a popular way for examining complex trait genetics. The aim of GWAS is to detect genetic variants that are associated with a given phenotype, and consequently gain more knowledge of its biology. Specifically, GWAS are most powerful in detecting common genetic variants, i.e. variants with MAF > 0.01 (or > 0.05) (McCarthy et al., 2008). Methods for analysing rare variants are beyond the scope of this thesis.

By far the most common way to conduct GWAS is to assume an additive genetic model: A typical biallelic SNP at a specific position in the human genome has two different options, say, A and B, for the occurring allele. Therefore, together in both pairs of the chromosome, the possible values for this specific SNP are AA, AB or BB. One of the alleles is defined as the reference allele and the other as the effect allele, and the numeric value for the SNP is the number of effect alleles, i.e. either 0, 1 or 2.

The traditional GWAS applies a GLM (Section 4.1.1.1) on the phenotype of interest, separately for each genetic variant $G_j,\ j = 1, \dots, J$:

$$g(\mathbb{E}(\mathbf{Y}|\mathbf{X})) = \mathbf{X}\boldsymbol{\beta}, \tag{4.15}$$

where $\mathbf{x}_1 = (G_{1j} \dots G_{nj})^T$ are the values for the genetic variant j, and the coefficient of interest is the effect of $\mathbf{x}_1$ on $\mathbf{Y}$, estimated as $\hat{\beta}_1$.

Although GWAS are literally merely association (or variable selection) studies, typically there is an implicit causal target when selecting other explanatory variables in the model (Section 4.3.2). Typical variables included are sex, age and genetic PCs (Section 4.3.1.2). The GWAS summary statistics are increasingly made publicly available (Buniello et al., 2018).

Using imputed genetic variants Up to hundreds of thousands of SNPs can be measured by widely used genotyping arrays. However, the number of SNPs analysed in GWAS can be considerably increased up to tens of millions by using imputation methods to infer the values of non-genotyped genetic variants, based on a set of reference haplotypes (group of alleles inherited together) of reference populations (Marchini & Howie, 2010). This is helpful

for combining GWAS results from multiple studies (Section 4.1.1.1), with possibly varying genotyping platforms.

The imputation procedure is based on the information of the measured genetic variants **G** and the reference haplotype **H**. Genetic imputation methods exploit the LD information of genetic variants in reference populations, where a set of individuals are genotyped by high-density SNP arrays (Loh et al., 2016; Marchini et al., 2007). The genotype imputation methodology is beyond the scope of this thesis.

The output of genotype imputation is a probabilistic prediction of the value of the genotype:

$$p_{ijk} = \mathbb{P}(G_{ij} = k|\mathbf{G}, \mathbf{H}), k \in 0, 1, 2, \sum_k p_{ijk} = 1,$$

where G_{ij} is the genotype for individual i at SNP j. Factors that affect the final imputation accuracy of each SNP are the SNP allele frequency (rarer variants are more difficult to impute), and the reference panel used (large sample size and an increased number of haplotypes in the reference panel improve imputation accuracy) (Das et al., 2016).

Different measures for imputation quality has been proposed (Marchini & Howie, 2010), the values for which range from 0 to 1, with 0 indicating complete uncertainty and 1 meaning no uncertainty in the imputed values for a given SNP. The measure applied for NFBC datasets (Chapter 3) is the estimated squared correlation between the imputed and true unobserved genotypes:

$$r^2 = \frac{\sum_{i=1}^{2n}(D_i - \hat{q}^2)}{\hat{q}(1-\hat{q})2n},$$

where D_i is the imputed effect allele probability at haplotype i (Das et al., 2016), and $\hat{q}$ is the estimated effect allele frequency for the variant. It is advised to filter out poorly imputed SNPs (e.g. $r^2 < 0.5$) from GWAS results.

To take the imputation uncertainty into account in GWAS, a popular option is to use allele dosages

$$d_{ij} = p_{ij1} + 2p_{ij2},$$

which are the expected allele counts for the effect allele in the imputed SNP. These can have

any value between 0 and 2, and $\mathbf{x}_1 = (d_{1j} \dots d_{nj})$ for the association analysis in equation (4.15).

Another way to incorporate the imputation uncertainty into the regression model is to incorporate each possible genotype to the likelihood function, weighted by their imputation probabilities (Marchini et al., 2007):

$$L(\mathbf{\Theta}, \mathbf{Y}, \mathbf{G}, \mathbf{X}', \mathbf{H}) = \prod_{i=1}^{n} \sum_{k} f_{\mathbf{\Theta}}(y_i | \mathbf{x}_i', G_{ij} = k) \mathbb{P}(G_{ij} = k | \mathbf{G}, \mathbf{H}),$$

where $f_{\mathbf{\Theta}}$ is the assumed probability model – such as a GLM – for the phenotype with parameters $\mathbf{\Theta}$, G_{ij} is the genotype for SNP j in individual i, $\mathbf{X}'$ is the matrix of other explanatory variables, and $\mathbf{G}$ and $\mathbf{H}$ are the genotype data and haplotype reference data, respectively, used for imputation.

Both of these methods have similar performance in GWAS when the effect sizes are small, which is typically the case for GWAS (Guan & Stephens, 2008). The dosage method is implemented in Plink software (C. C. Chang et al., 2015; Purcell et al., 2007), while both dosage and the probability method are implemented in SNPTEST software (Marchini et al., 2007).

Multiple testing correction for GWAS Due to the high number of variables, an adequate control of multiple testing is important. In GWAS, a conventional choice for controlling the FWER for multiple hypothesis testing is $\alpha = 5 \times 10^{-8}$, corresponding to a Bonferroni correction for million independent tests (Dudbridge & Gusnanto, 2008; International HapMap Consortium, 2007).

Reverse regression for correlated phenotypes The traditional approach in GWAS is to consider each phenotype separately, even though many phenotypes are correlated with each other. For example, ADHD symptoms can be split into the core symptoms subtypes, inattention and hyperactivity, which are strongly correlated (Sokolova et al., 2016). Joint modelling of such correlated outcomes may increase statistical power (Teixeira-Pinto et al., 2009).

Mägi et al. (2017) introduced SCOPA (Software for COrrelated Phenotype Analysis) software which applies a *reverse regression* framework for GWAS. In reverse regression, the

vector of observed values for each genetic variant $\mathbf{G}_j, j = 1, \dots, J$ is treated as the outcome in turn, and the h correlated phenotypes $\mathbf{Y} = (\boldsymbol{y}_1 \dots \boldsymbol{y}_h) \in \mathbb{R}^{n \times h}$, $\mathbf{y}_m = (y_{1m} \dots y_{nm})^T$, $m = 1, \dots, h$ are treated jointly as explanatory variables using a traditional linear model:

$$G_{ij} = \beta_{0j} + \sum_{m=1}^{h} \beta_{mj} y_{im} + \varepsilon_{ij},\ \varepsilon_i \sim N(0, \sigma_j^2)\ i = 1, \dots, n. \tag{4.16}$$

Both genotyped and imputed SNPs can be used, and the reverse regression can incorporate both categorical and quantitative phenotypes simultaneously in the same model. Other explanatory variables can be taken into account by regressing the outcomes on the explanatory variables, and using the remaining residuals for association analysis. The evaluation for an association for each SNP $\mathbf{G}_j$ can be done via testing $\mathrm{H}_0 : (\beta_{1j}\ \dots\ \beta_{hj}) = 0$. Reverse regression GWAS is applied in Study III (Chapter 7).

Identifying quantitative trait loci As well as for complex traits, GWAS can be conducted for another omics data type, i.e. variables on an omics level, such as metabolome or proteome, are treated as a phenotype one by one. Such omics-wide GWAS provide information on the genetic makeup of the omics levels themselves. Genetic variants associated with other omics are called molecular quantitative trait loci (QTL). For example, genetic variants associated with differential DNA methylation, gene expression, or circulating protein levels are termed methylation QTL (mQTL), expression QTL (eQTL) or protein QTL (pQTL), respectively. QTLs are traditionally classified into *cis*-QTLs, which are variants associated with the expression of nearby genes (e.g. within 500 kB), and *trans*-QTLs, which are associated with the expression of distal genes.

The *cis* and *trans*-QTLs show different features, discussed in detail by e.g. Ye et al. (2020). QTLs are increasingly catalogued in publicly available databases, such as GTEx for eQTL (GTEx Consortium, 2017), and ARIES for mQTL (Gaunt et al., 2016), both presented earlier in Chapter 3.

Combining results from multiple studies As outlined in Section 2.1, the individual effect sizes of common variants on complex traits tend to be small and the need for collaborative efforts to increase sample sizes and consequently improve statistical power has been highlighted (Risch & Merikangas, 1996; W. Y. S. Wang et al., 2005). Meta-analysis meth-

ods can be used to pool association analysis results across multiple studes. As of date, the largest sample sizes for GWAS meta-analysis based on large international consortia reach up to over 1 million individuals (J. J. Lee et al., 2018).

Polygenic risk scores Despite the large sample sizes in GWAS, the observed contributions of individual genetic variants on the variance or the risk liability of complex traits is small. Maher (2008) introduced the concept of *missing heritability*, i.e. the proportion of phenotypic variance attributable to additive genetic factors based on GWAS results is much lower than the one calculated from pedigree studies. This has increased the interest in aggregating GWAS results across the full genome.

Polygenic risk scores (PRS) combine the effects of multiple common DNA variants on a phenotype, and they can be used in epidemiological studies as a measure for the genetic liability to a trait (Dudbridge, 2013). A PRS for individual i is a weighted sum of M genotypes of a set of SNPs:

$$\text{PRS}_i = \sum_{j=1}^{M} w_j G_{ij}, \tag{4.17}$$

where $i = 1, \ldots, n$, w_j are the weights for each genotype j and G_{ij} is the genotype of variant j for individual i. The number of SNPs M can be defined by a p-value cut-off, and it has been shown that a lenient threshold improves the predictive accuracy of PRS for complex traits (Choi et al., 2020; Purcell et al., 2009).

The strength of PRS is heavily dependent on the base GWAS summary statistics sample size (Dudbridge, 2013). Therefore, the weights are commonly based on the regression coefficients of the genetic variants from GWAS summary statistics. It is also important to take the correlation of genetic variants, known as *linkage disequilibrium* (LD) (Slatkin, 2008), into account when selecting the weights. The traditional way is to clump the variants so that a variant with a lowest p-value within a specific genetic locus is detected, and all other variants in LD with the lead variant are pruned out (Euesden et al., 2014; Privé et al., 2018). Alternatively, all variants can be taken into account and the weights can be adjusted based on the underlying LD structure of a reference population, using e.g. either penalised regression or Bayesian methods, as implemented in lassosum (Mak et al., 2017) or LDpred (Vilhjálmsson et al., 2015) softwares, respectively.

4.2.1.2 Epigenome-wide association studies

Partially due to the missing heritability, interest in non-genetic variation in the aetiology of complex traits has increased. Akin to GWAS, other omics-wide association analysis approaches are an efficient way to examine the molecular variation with respect to the phenotype of interest.

The statistical principles of GWAS mostly apply to other association studies across the different omics. Here, I cover EWAS, which assess the epigenetic variation (mainly DNA methylation at CpG sites, Section 2.2.1) across the full genome.

Methylation at a CpG site is commonly measured using methylation Beta value:

$$\text{Beta} = \frac{M}{M + U + c},$$

where M is the methylated and U is the unmethylated intensity at the CpG site, and c an offset term to ensure numerical stability for low intensity values, with $c = 100$ is commonly used (Bibikova et al., 2011). The Beta value ranges between 0 and 1, and it can be transformed to M-value by $M = \log_2(\text{Beta})$ to improve numerical stability in EWAS (Du et al., 2010). A widely used Illumina HumanMethylation450 array quantifies DNA methylation at over 450,000 CpG sites (Bibikova et al., 2011).

Similarly as for GWAS, regression models such as GLMs (Section 4.1.1.1) can be used for EWAS. The methylation Beta value (or M-value) can be used as either the outcome or as an explanatory variable in EWAS regression models. For multiple testing correction, a widely-used threshold is $\alpha = 10^{-7}$ (Lehne et al., 2015), corresponding to a Bonferroni-corrected FWER threshold of approximately 450,000 independent tests. For multiple correlated outcomes, reverse regression methods applied in GWAS naturally extend to EWAS by replacing the genetic variant G_j by the methylation Beta value (or M-value) as the outcome in equation (4.16).

The main differences between GWAS and EWAS are the tissue specificity and cell-type heterogeneity of epigenetic markers and the dynamic nature of the epigenomic profile. DNA methylation varies across tissues, and typically samples from easily accessible tissues, such as blood, are available in epidemiological cohorts. Whether a possible differential DNA methylation related to a phenotype manifests also in the relevant tissue, such as brain

tissue for psychiatric outcomes, remains a challenge.

The cell-type heterogeneity of DNA methylation implies that as all tissues are composed of multiple cell types, possible differential DNA methylation with respect to the phenotype may actually reflect differential cell-type composition. Houseman et al. (2012) described a method for estimating relative proportions of different cell type components in whole blood. This method uses DNA methylation at 500 CpG sites as surrogate measures for cell proportions. The estimated cell-type proportions in whole blood are obtained from the reference probes using a non-linear random effects model (Houseman et al., 2012; Jaffe & Irizarry, 2014). This method is shown to explain much of the observed variability in blood DNA methylation (Jaffe & Irizarry, 2014). These estimated relative proportions are commonly added as explanatory variables in EWAS when DNA methylation from whole blood is used.

Other common explanatory variables used in EWAS are age, sex, and potentially other relevant confounders (Section 4.3.1.2). Finally, as the epigenome is prone to environmental perturbations, inferring the direction of association of epigenetic markers and a phenotype can be challenging. Causal inference in epigenetic epidemiology is discussed further in Section 4.3.1.2.

4.2.2 Omics data integration

In general, biological processes involve regulation and signalling across different omics. Thus, it is crucial to integrate different omics data types to unravel the molecular patterns and processes of complex traits in more detail.

Data integration is a general term for combining data from different omics layers in a meaningful way to improve the understanding of the underlying biology of interest (Ritchie et al., 2015). The motivation is that combining information across multiple omics datasets may give additional insight to the biological processes and putative mechanisms. Moreover, data integration can compensate for missing or unreliable information of a single data type (Ritchie et al., 2015). Convergence of results based on multiple omics datasets strengthens the evidence for a hypothesised mechanism.

In the surplus of omics data, integrative methods have attracted plenty of interest in recent years (Hasin et al., 2017; Yan et al., 2017). Moreover, even though especially genomic data

is highly sensitive, summarised association statistics by large consortia are increasingly publicly available, which also has accelerated the interest in integrative data analysis approaches (Richardson et al., 2016).

4.2.2.1 Data integration approaches

Different approaches for data integration are *early*, *intermediate* and *late integration* (Figure 4.1) (Zitnik et al., 2019). In early integration, all datasets are combined to a single large dataset to develop a statistical model. Any analysis method that is applicable for single omics data method can be applied in early integration. A challenge in this approach is how to normalise the variables to a common scale, as typically the variable distributions between omics datasets are highly heterogeneous. Early stage data integration is applied in Study III (Chapter 7), where prediction models on ADHD symptoms were built using genomic, epigenomic and metabolomic data.

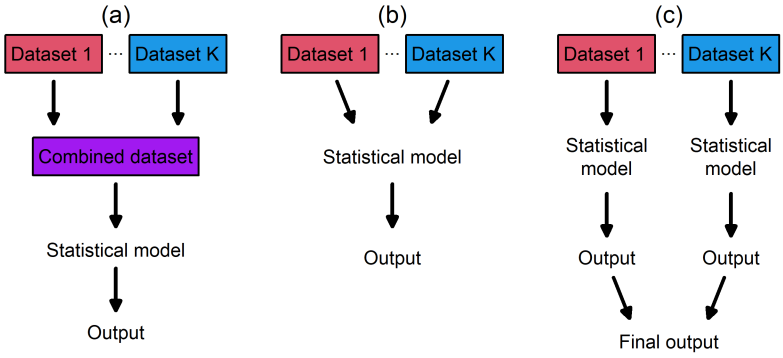

Figure 4.1: Different data integration approaches across K datasets. (a) Early integration, where all datasets are combined to a single dataset, and statistical modelling is applied to the full dataset. (b) Intermediate integration: statistical modelling is applied to model the relationships across the datasets. (c) Late integration: statistical models are built in each dataset independently, and the results are combined to obtain a final output.

In intermediate integration, a statistical model is applied to the different omics simultaneously to map the relationships across omics datasets. An intermediate data integration method based on canonical correlation analysis is presented in Section 4.2.2.2, and applied in Study III (Chapter 7). In addition, omics-wide GWAS for QTL discovery described in Section 4.2.1.1 can be considered as an intermediate data integration method.

In late integration, a first-level model is built for each omics data independently. The results from these models are then taken forward and integrated across omics. Integrating QTL information across multiple omics data is an example of late integration, and it may reveal insights into biological processes. If there is an association signal for the same QTL (or a genetic locus) in multiple omics datasets, this strengthens the hypothesis that this variant is related to the corresponding biological pathway. In Study V (Chapter 9), a late-stage integrative approach is taken for Mendelian Randomisation (Section 4.3.3).

4.2.2.2 Sparse canonical correlation analysis

Canonical correlation analysis (CCA) was introduced by Hotelling (1936) and the traditional CCA can be used to quantify relationships across two datasets. CCA can therefore be considered as an intermediate data integration method. Let $\mathbf{X}_1 \in \mathbb{R}^{n \times p_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{n \times p_2}$ be two data matrices, with mean 0 and standard deviation 1 for all columns in both $\mathbf{X}_1$ and $\mathbf{X}_2$. The aim of CCA is to find pairs of *canonical vectors* $\mathbf{w}_1^{(1)} \in \mathbb{R}^{p_1}$ and $\mathbf{w}_2^{(1)} \in \mathbb{R}^{p_2}$ that maximise the correlation between the two data matrices.

The objective function to minimise is

$$\min_{\mathbf{w}_1^{(1)}, \mathbf{w}_2^{(2)}} -\text{cor}(\mathbf{X}_1 \mathbf{w}_1^{(1)}, \mathbf{X}_2 \mathbf{w}_2^{(1)})$$

subject to constraint $\text{Var}(\mathbf{X}_1 \mathbf{w}_1^{(1)}) = \text{Var}(\mathbf{X}_2 \mathbf{w}_2^{(1)}) = 1$. Further canonical pairs $\mathbf{w}_1^{(r)}, \mathbf{w}_2^{(r)}, r = 2, \ldots, \min(p_1, p_2)$ can be computed with additional constraints that the new canonical vector pairs are orthogonal of all preceding pairs. This index is dropped from further notation for simplicity.

Variable selection can be induced in CCA by adding penalty functions $P_{\tau_{\mathbf{w}_1}}$ and $P_{\tau_{\mathbf{w}_2}}$, that depend on the regularisation parameters τ_1 and τ_2, respectively:

$$\min_{\mathbf{w}_1,\mathbf{w}_2} -\text{cor}(\mathbf{X}_1\mathbf{w}_1, \mathbf{X}_2\mathbf{w}_2) + P_{\tau_{\mathbf{w}_1}}(\mathbf{w}_1) + P_{\tau_{\mathbf{w}_2}}(\mathbf{w}_2) \tag{4.18}$$

Such sparse solution for CCA can be considered as a network model for two omics datasets (Shi et al., 2019). A common selection for penalty function P is the lasso penalty based on L_1:

$$P_{\tau_{\mathbf{w}_j}}(\mathbf{w}_j) = \tau_j||\mathbf{w}_j|| = \tau_j \sum_{i=1}^{p_j} |w_{ij}|$$

for $j = 1, 2$. Witten et al. (2009) described how to obtain a sparse solution for CCA via penalised matrix decomposition (PMD):

$$\begin{gathered}\min_{\mathbf{w}_1,\mathbf{w}_2} -\text{cor}(\mathbf{X}_1\mathbf{w}_1, \mathbf{X}_2\mathbf{w}_2) + \tau_1||\mathbf{w}_1|| + \tau_1||\mathbf{w}_2|| \\ \text{subject to } \text{Var}(\mathbf{w}_j) = 1,\ j = 1, 2.\end{gathered} \tag{4.19}$$

This assumes that $\mathbf{X}_j^T\mathbf{X}_j = \mathbf{I}_{p_j}, j = 1, 2$, that is, all columns of $X_j, j = 1, 2$ are independent of each other. With multi-omics data, different variables may be highly correlated with each other, and therefore this is not a realistic assumption. Suo et al. (2017) proposed a relaxed version of PMD:

$$\begin{gathered}\min_{\mathbf{w}_1,\mathbf{w}_2} -\text{cor}(\mathbf{X}_1\mathbf{w}_1, \mathbf{X}_2\mathbf{w}_2) + \tau_1||\mathbf{w}_1|| + \tau_2||\mathbf{w}_2|| \\ \text{subject to } \text{Var}(\mathbf{X}_j\mathbf{w}_j) \leq 1,\ j = 1, 2.\end{gathered} \tag{4.20}$$

This can be generalised to M datasets:

$$\begin{gathered}\min_{\mathbf{w}_m} -\sum_{q<r} \text{cor}(\mathbf{X}_q\mathbf{w}_q, \mathbf{X}_r\mathbf{w}_r) + \sum_{m=1}^{M} \tau_m||\mathbf{w}_m|| \\ \text{subject to } \text{Var}(\mathbf{X}_m\mathbf{w}_m) \leq 1,\ m = 1, \ldots, M,\end{gathered} \tag{4.21}$$

with lists of canonical vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m$ minimising the objective function.

The problem in (4.21) is multi-convex, i.e. it is convex in $\mathbf{w}_r$ when $\mathbf{w}_q\ \forall\ r \neq q$ are fixed. The solution can be obtained via linearised alternating direction method of multipliers described in Parikh & Boyd (2014) and recently implemented for multiple datasets by Rodosthenous

et al. (2020). The regularisation parameters τ_m can be selected via CV as described in Section 4.1.1.1.

The method presented above can be considered as an intermediate stage data integration method across multiple omics datasets with simultaneous variable selection, applied in Study III (Chapter 7).

4.2.3 Summary

Overall, all the methods presented in Section 4.2 aim provide enhanced knowledge on complex trait biology. A much more stringent consideration is required if the aim is to infer causal pathways between the molecular variables and the outcome. In the next section, I introduce the key concepts of causal inference and how to apply causal inference in genetic and other molecular epidemiology using observational data.

4.3 Causal inference

In this section, the interest is in the potential causal effect of an *exposure* variable (e.g. exposure to maternal smoking in pregnancy) on an *outcome* (e.g. offspring DNA methylation at a CpG site). Causality is defined within the *counterfactual* (or *potential outcomes*) framework as described by Hernán & Robins (2019). The counterfactual is a value of the exposure that a study participant does not have. The outcome that would have been obtained under the counterfactual exposure is called the potential outcome. As the counterfactual situation is, by its definition, not observed, this poses a fundamental problem to causal inference on an individual level (Hernán, 2004). However, the average causal effects (hereafter causal effects) can be consistently estimated on a population level, provided that the assumption of *exchangeability* holds, that is, the potential outcome is independent of the actual value of the exposure (Hernán & Robins, 2019).

Identifying the causal effect of interest is informative to the potential interventions from public health perspective (Glass et al., 2013). A presence of causality suggests a possibility for an intervention for a modifiable exposure, whereas robust accumulation of absence of evidence for a causal effect would imply that such intervention would not be efficient in preventing an adverse outcome.

Causal inference is discussed from the perspective from observational studies using generalised linear models (Sections 4.3.1 and 4.3.2), instrumental variable methods as implemented in Mendelian Randomisation (Section 4.3.3), and causal mediation analysis (Section 4.3.4).

4.3.1 Causal inference in observational studies

The exchangeability assumption mentioned above is achieved if the exposure can be randomised, as is done in randomised controlled trials (RCTs) (Hernán & Robins, 2019). However, as discussed in Hernán & Robins (2019) and Rothman (2008), in many instances RCTs are not feasible, due to financial or ethical reasons. RCTs are also generally labour-intensive, and require high patient costs. These financial factors limit the number of study participants, and consequently, statistical power to detect small or moderate effects. Regarding the ethics, it is unethical to randomise study participants into harmful exposures, such as smoking during pregnancy. Thus, it is frequent that causal inference has to be based on observational studies, where there is no randomisation (or other intervention) by the investigators.

In observational studies and other non-randomised settings, eliciting the causal effect of interest from all other uncontrolled systematic variation requires additional structural knowledge and assumptions about the sources of systematic error. This structural knowledge is typically incomplete (Hernán et al., 2019). Thus, the validity of causal inference based on observational studies depends on the correspondence between the assumptions and reality (Rothman, 2008). Causal graphs, such as directed acyclic graphs (DAGs) may be helpful in identifying variables that should be taken into account (Greenland et al., 1999; Pearl, 1995).

The key limitations to causal inference from observational studies include confounding, selection bias, reverse causation and measurement error (Figure 4.2) (Davey Smith & Ebrahim, 2002, 2003; Hernán & Robins, 2019). Both confounding and selection bias are forms of lack of exchangeability between the study participants with different values of exposure variable. Confounding occurs when a third variable, a *confounder*, influences both exposure and the outcome (Porta, 2014).

Selection bias relates to various biases that arise from the procedure by which study partic-

ipants are selected into the analysis (Hernán et al., 2004). Selection bias is a form of a more general *collider bias*, when conditioning on a common outcome causes a spurious correlation between two variables (Hernán et al., 2004; Munafò et al., 2017). In reverse causation, the outcome affects the exposure. This can be an issue especially in cross-sectional observational studies, where the temporality of the exposure and outcome is not always clear. Measurement error in the exposure, confounders or the outcome can also cause bias to the causal estimates of interest (Figure 4.2).

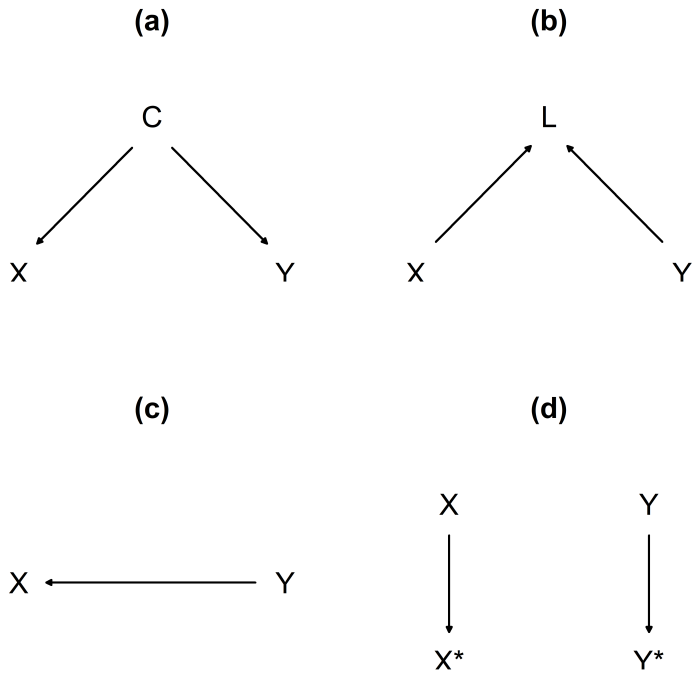

Figure 4.2: Limitations to causal inference from observational studies. The interest is in the causal effect of X on Y. (a) A confounder C causes a spurious association between X and Y. (b) A collider L is a common outcome of X and Y, and conditioning on this would cause a spurious association between X and Y. Selection bias would occur if L is a selection criterion to the study. (c) Reverse causation, i.e. Y causes X. (d) Measurement error: inference is based on variables X^* and Y^* which are measured with error, causing bias to the causal estimate.

4.3.1.1 Causal inference in birth cohorts

In birth cohorts, or pregnancy-birth cohorts, women are recruited during pregnancy and their offspring are then followed up. The longitudinal and inter-generational design provides an ideal opportunity to examine the effects of pregnancy or early life exposures on

later life outcomes (Richmond, Al-Amin, et al., 2014). The temporality of the pregnancy-time exposure and offspring outcome is clear, and is thus devoid of reverse causality. The prospective data collection minimises the possibility of recall bias when measuring exposures during pregnancy.

Many birth cohorts measure data from both parents. If available, paternal (or partner) exposures give a natural *negative control exposure* to examine the potential causal intrauterine effects. The intention of using negative controls is to mimic the hypothesised causal mechanism with the same sources of bias and similar confounding structure, but where the hypothesised mechanism is implausible (Davey Smith, 2008; Davey Smith et al., 2012; Lipsitch et al., 2010). The associations from negative control analysis are compared to the results from the actual exposure-outcome analysis. If the magnitude of association is stronger for the actual exposure-outcome analysis, this is consistent with a causal explanation. In turn, the similarity of the results between exposure-outcome analysis and negative control analysis would imply that the hypothesised association is due to unmeasured confounding.

As an example, exposure to paternal smoking during pregnancy can be used as a negative control exposure to evaluate the causal intrauterine effect of exposure to maternal smoking during pregnancy on offspring health outcomes. Maternal and paternal smoking are likely to have similar familial confounding factors and to be subject to similar biases.

A limitation for causal inference in birth cohorts is that due to the nature of birth cohorts, the long time between exposure during pregnancy and offspring's later life outcomes increases the possibility of confounding (Richmond, Al-Amin, et al., 2014). Prospective follow-up is also time-consuming. Differential loss to follow-up due to health conditions is likely in long follow-up studies. This may lead to bias in causal estimates of interest (Hernán et al., 2004).

The main limitation of using paternal smoking as a negative control is that second-hand smoking may be its causal mechanism on offspring health outcomes. However, A. E. Taylor, Davey Smith, et al. (2014) examined the cotinine levels – a biomarker for smoking – among non-smoking pregnant women whose partner was smoking, and concluded that the causal pathway via second-hand smoking is unlikely. In addition, paternal smoking as a negative control might be biased due to *assortative mating* (the phenomenon of partners selecting each other based on phenotypic similarities), or differential confounding structure or measurement error of the exposure in the two groups (Keyes et al., 2014; Madley-Dowd et al., 2020; Sanderson et al., 2017).

4.3.1.2 Causal inference in genetic and other molecular epidemiology

The emergence of multi-omics has enabled valuable data for establishing causality based on observational studies. Detailed analysis of molecular mechanisms for complex traits, once considered as a 'black box' in aetiological research, is being thoroughly investigated in modern molecular epidemiology (Fedak et al., 2015).

In particular, genetically informed methods have distinct advantages for causal inference based on observational studies. Family-based designs can be used to adjust for genetic confounding, based on pedigree information on genetic relatedness between individuals (Pingault et al., 2018). Here, I discuss causal inference in genetic epidemiology using non-pedigree data.

The genetic variants are fixed at conception, and therefore not subject to reverse causality (Davey Smith & Ebrahim, 2003), as the phenotype cannot affect the genotype. Moreover, the genetic variants are randomly assorted at conception, conditional on the parents' genotypes. This fact has benefits for Mendelian Randomisation (Section 4.3.3) as a tool for causal inference.

However, even in the absence of reverse causation, there are several other ways how an association between a genetic variant and a phenotype can arise (Morris et al., 2020). *Population stratification* refers to the confounding of a variant-phenotype association by population structure (Price et al., 2006), and can cause a correlation between a genotype and a phenotype. A classic example of population stratification is a hypothetical genetic variant that, when analysed among Asian and European populations, is associated with chopstick use (Hamer, 2000). However, this association reflects different allele frequencies of the variant among the different populations rather than a causal effect of the variant on the preferred cutlery. Population stratification is typically taken into account in genetic analyses by restricting the sample into a homogeneous population, and adjusting for the genetic similarity by e.g. genetic principal components (Section 4.2.1) (Lawson et al., 2020).

Pleiotropy refers to the situation where a genetic variant is associated with more than one phenotype (Paaby & Rockman, 2013). This can be either vertical pleiotropy, where the phenotypes are on the same causal pathway, or horizontal pleiotropy, where the variant affects the phenotypes through different pathways (Figure 4.3) (Hemani, Bowden, & Davey Smith, 2018). Moreover, correlation between a genotype and a phenotype can also

arise due to *linkage disequilibiurm* (non-random association between two different variants, LD) (Slatkin, 2008), where the genetic variant examined is correlated (i.e. in LD) with the causal variant, but is not itself causal to the phenotype.

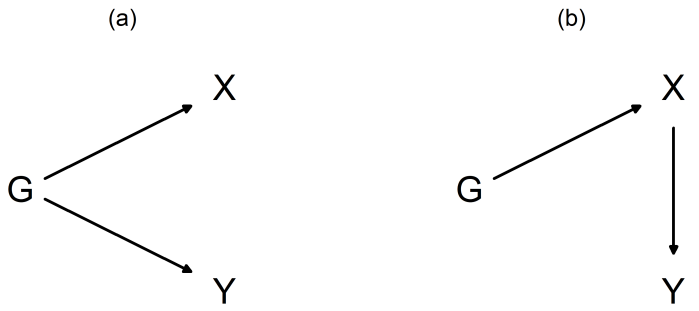

Figure 4.3: Pleiotropy. (a) horizontal pleiotropy, where a genetic variant G causes both X and Y via different pathways; (b) vertical pleiotropy, where a genetic variant G causes X, which causes Y in turn.

Even the genetic variants from the parents that are not transmitted to the offspring can affect the child. In this scenario, the parental genotype influences the parental phenotype, which in turn affects the offspring phenotype. Kong et al. (2018) gives an example of parental non-transmitted alleles being associated with offspring educational attainment, possibly due to parental genetics leading to an education-friendly home environment. This effect is called the genetic nurture (Kong et al., 2018), or *dynastic effects* (Morris et al., 2020). Assortative mating can also cause a correlation between a genotype and a phenotype (M. R. Robinson et al., 2017; Yengo, Robinson, et al., 2018).

In contrast to genetic epidemiology, non-genetic molecular epidemiology is more limited by issues of confounding and reverse causation that are typical in 'traditional' epidemiology. In

contrast to genetic variants, other omics data types can be either a cause or a consequence of the phenotype of interest. For example, DNA methylation is shown to be altered by smoking status on an epigenome-wide scale (Joehanes et al., 2016), and the results by Wahl et al. (2017) suggest that adiposity is driving differential DNA methylation, rather than the other way round.

For causal inference in epigenetic epidemiology, further challenges are tissue specificity and cell-type heterogeneity (Section 4.2.1.2). DNA methylation data is typically available from easily accessible tissues, such as blood. However, they are not always the most relevant tissues for the outcome examined, such as for psychiatric outcomes, where brain would be the ideal tissue to measure DNA methylation. The results by Walton et al. (2015) suggest that some, but not all, blood DNA methylation markers can be used to proxy DNA methylation in brain tissue.

In addition to variability across tissues, variations in the proportions of cell types within the tissue can also affect associations between DNA methylation markers and the phenotype, either as a confounder or a mediator (Section 4.3.4). All tissues are composed of multiple cell types, and possible differential DNA methylation with respect to the phenotype may actually reflect differential cell-type composition if the cell type heterogeneity is not taken into account.

4.3.2 Generalised linear models for causal inference

Let $\mathbf{Y}$, $\mathbf{X}$ and $\boldsymbol{\beta}$ be as defined in the beginning of this chapter. The GLM (Section 4.1.1.1)

$$g(\mathbb{E}(\mathbf{Y}|\mathbf{X})) = \mathbf{X}\boldsymbol{\beta}$$

can be used in causal inference framework to examine the effect of exposure on the outcome. The measured values for the exposure are added as jth column to $\mathbf{X}$. The parameter of interest is the coefficient β_j. In linear regression, the estimated value for the coefficient $\hat{\beta}_j$ is the amount of change in observed $\mathbf{Y}$ per unit change in $\mathbf{x}_j$. In logistic and proportional odds regression, β_j is the change in log-odds in observed $\mathbf{Y}$ per unit change in $\mathbf{x}_j$. Evidence for causality can be evaluated by examining whether the observed parameter estimate $\hat{\beta}_j$ differs from zero, for example by using methods based on maximum likelihood, such as Wald

test. To improve statistical power, results for the same causal question can be pooled from estimates across multiple studes via meta-analysis methods.

A confounder can be taken into account by adding it as an additional explanatory variable $\mathbf{x}_{k,k\neq j}$ in the model. Thus, when multiple explanatory variables are included, the coefficient β_j is conditional on holding the values of other explanatory variables $\mathbf{x}_{k\neq j}$ constant. This adjustment is usually done in association studies (Section 4.2.1.1) as well, although causality is not inferred in a formal manner.

It is of equal importance to be aware of the variables that should not be adjusted for. Adjusting for a variable that is influenced by (rather than influencing) both exposure and outcome leads to collider bias and consequently to a spurious association between the exposure and the outcome (Munafò et al., 2017). Especially in GWAS, there is a risk of introducing collider bias if heritable phenotypes are included as explanatory variables (Day et al., 2016). For example, ADHD is associated with lower educational attainment in later life (Erskine et al., 2016), and adding educational attainment as an explanatory variable in ADHD GWAS would cause a spurious association between variants that influence educational attainment and the risk of ADHD (Figure 4.2 (b)).

The regression coefficient β_j can be interpreted as a causal estimate, albeit with strong assumptions (Gelman, 2007). All confounders should be taken into account with proper functional forms, and there should be no unmeasured confounding present. This is never certain when using observational data. Moreover, statistical methods themselves cannot distinguish between exposure and outcome, nor between confounders, colliders, and mediators (Section 4.3.4) (VanderWeele, 2019).

The interpretation of the causal effect is always specific to the selected exposure. Inferring multiple causal effects from the same regression model can lead to so-called 'Table 2 fallacy', as coined by Westreich & Greenland (2013). For example, if the causal effect of interest is the impact of differential DNA methylation at a given CpG site on ADHD, exposure to maternal smoking should be considered as a confounder. However, the causal effect of exposure to maternal smoking on offspring ADHD cannot be inferred from the same model, as now DNA methylation is not a confounder, but a mediator (Section 4.3.4).

4.3.3 Mendelian Randomisation (MR)

Instrumental variable methods can be used to consistently estimate causal effects in the absence of exchangeability assumption (Hernán & Robins, 2019). Consider a DAG in Figure 4.4, with the interest in the effect of exposure variable X on outcome Y. In an observational setting, this association is typically affected by confounders C. However, the causal effect of X on Y can be examined if there exists a variable V which complies to the following instrumental variable assumtions (Greenland, 2000):

i. V is robustly associated with X
ii. V is independent of the confounders C
iii. V affects Y only via X

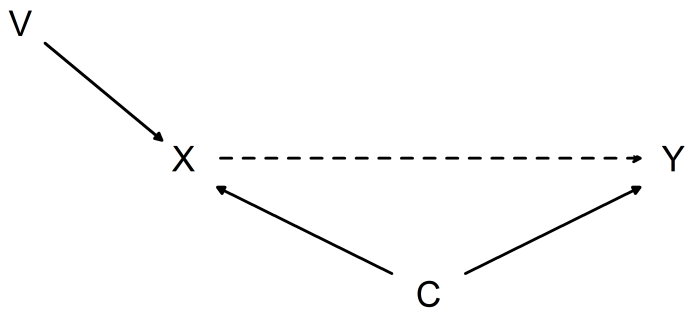

Figure 4.4: A DAG for instrumental variable methods. The aim is to examine the causal effect of exposure X on outcome Y (dashed line). In an observational setting, this association is typically biased by confounders C. However, an instrumental variable V can be used for causal analysis. It is required that V is robustly associated with X, V is independent of confounders C (i.e. no arrow between them), and that V affects Y only via X (i.e. no other path between V and Y).

Mendelian Randomisation (MR) uses genetic variants to assess causality between exposure and outcome (Davey Smith & Ebrahim, 2003; Katan, 1986), and can be thought as a method that uses the genetic variants as instrumental variables (Thomas & Conti, 2004).

The name of the method derives from the Mendelian principles in genetics and the random assortment of alleles at conception, which results in a randomisation of genetically different levels of exposure, independent of confounding factors. This is analogous to an RCT, where study participants are randomly assigned to different therapeutic interventions (Lawlor et al., 2008). Therefore, MR corresponds to an intention-to-treat analysis in RCTs, that is, study participants are analysed based on the randomisation group, independent of the treatment compliance. This ensures that no confounding is reintroduced after potential non-compliance (Davey Smith & Hemani, 2014).

The instrumental variable assumptions imply that the confounders C need not be measured at all. This is advantageous as typically all observational studies suffer from unmeasured confounding. Also, it should be noted that the aim in MR is not to identify functional variants for the outcome, but to exploit the information of variant-exposure association for examining the causality between the exposure and the outcome. Estimating the causal effect size in MR requires an additional assumption of homogeneity or monotonicity (Burgess & Thompson, 2015; Hernán & Robins, 2006). Even in the case when neither of these assumptions hold, MR still provides a valid test for the null hypothesis of no causality (Burgess et al., 2013; Didelez & Sheehan, 2007).

Another advantage of MR is the avoidance of reverse causation (Davey Smith & Ebrahim, 2003), as the genetic variants used as instruments for X are selected based on the association with the exposure only. Furthermore, if there are genetic variants that can be used as instrumental variables available for the outcome, one can conduct *bidirectional* MR, i.e. also examine the causality from Y to X by treating Y as the exposure and X as the outcome. The main disadvantages are related to violations of instrumental variable assumptions, and they are discussed in detail in Section 4.3.3.3.

4.3.3.1 Strategies for selecting instrumental variables for MR

In an ideal situation, the instrument for the exposure would be a variant within a single gene – a *cis*-variant – which has a known link with the exposure and whose biological function is

well-understood. For example, if the exposure of interest is differential DNA methylation at a CpG site within a given gene locus, a natural choice for an instrumental variable would be a genetic variant within the same gene locus that is also a methylation QTL, i.e. a variant that is associated with the corresponding DNA methylation levels. Using such a *cis*-variant is likely to be reliable for molecular exposures, such as DNA methylation (Burgess et al., 2020). The variant used in MR need not be the causal variant itself, as any variant in LD with the causal variant will suffice (Davey Smith & Hemani, 2014).

However, such monogenic instruments are not always available, especially for complex traits. Even if there exists a suitable *cis*-variant, the genetic effect may be miniscule, leading to a weak instrument (Section 4.3.3.3) and to a requirement of very large sample sizes (Davey Smith & Hemani, 2014). In addition, the polygenicity of complex traits implies widespread pleiotropy (Visscher & Yang, 2016), and horizontal pleiotropy violates the instrumental variable assumption (iii) as depicted in Figure 4.5 (Lawlor et al., 2008).

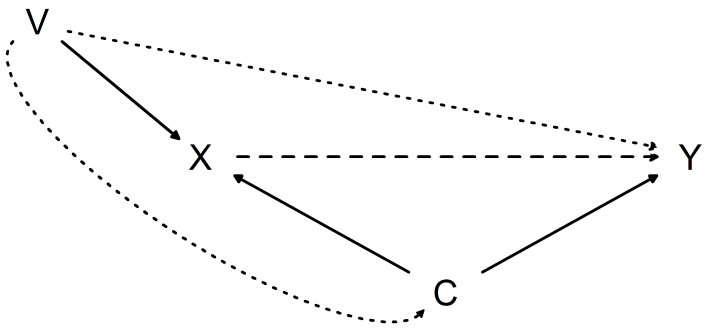

Figure 4.5: Horizontal pleiotropy in MR. The aim is to examine the causal effect of exposure X on outcome Y (dashed line). Horizontal pleiotropy implies an alternative path from genetic instrumental variable V on outcome Y (dotted lines) independently of X.

Despite the polygenicity, MR using multiple genetic variants typically has higher power compared to single-variant MR. The proprtion of variance explained by the instrumental variable is increased by including multiple independent variants, which in turn improves precision (Davey Smith & Hemani, 2014). A typical way to select genetic variants as instrumental variables in MR is based on a p-value threshold, typically at genome-wide significance of $p < 5 \times 10^{-8}$, in the GWAS of the exposure variable. More relaxed p-value thresholds can also be applied, and polygenic risk scores (Section 4.2.1.1) can also be used as instrumental variables, however in these cases there might be more issues regarding horizontal pleiotropy (Section 4.3.3.3) (Pingault et al., 2018). Independence of the variants can be ensured by clumping the variants, i.e. detecting a leading variant with the lowest p-value within a specific locus (such as 10,000 kb window from the lead variant), and then excluding other variants that are correlated (e.g. $r^2 > 0.1$) with the lead variant (Privé et al., 2018).

The downside of using multiple genetic variants as instrumental variables is that the risk of including invalid instruments increases, which in turn may lead to either under- or overestimating the causal estimates of interest (Burgess & Thompson, 2013). However, numerous methods have been developed to allow for some invalid instruments in MR using multiple genetic variants (see Section 4.3.3.4).

4.3.3.2 Data sources and effect size estimation

MR can be carried out using either one-sample MR or two-sample MR. In one-sample MR, the variant-exposure and variant-outcome association estimates are measured in the same individuals. One-sample MR ensures the absence of problems in comparing variant-phenotype associations in two separate populations, including population characteristics or ethnicity (Burgess et al., 2016). In two-sample MR, the variant-exposure and variant-outcome associations are estimated in two separate datasets (Lawlor et al., 2008). This enables comparisons between an exposure and outcome that need not be measured in the same individuals. This can be particularly useful if either the exposure or the outcome are difficult to measure (Davies et al., 2018), and larger sample sizes can also be reached. The two-sample MR also protects against the bias due to winner's curse in overestimating the variant-exposure effect sizes (Lawlor, 2016).

MR can be conducted using either individual-level data or summarised data. Individual-level data allows more flexibility for sensitivity analyses, such as subgroup analyses or non-linear

effects (Burgess et al., 2020).

Individual-level data are not always available, for example due to data sharing policies. Using summarised data enables the use of variant-phenotype effect estimates from large-scale GWAS, which can substantially increase power in MR (Burgess et al., 2020). These GWAS summary statistics are increasingly being made publicly available (Burgess et al., 2015). Here, only the two-sample MR setting using summarised data is considered.

Let $\hat{\beta}_{X|g}$ and $\hat{\beta}_{Y|g}$ be the effect size estimates of genetic variant g on exposure X and outcome Y, respectively, from large GWAS for each variable. In the case of only one genetic variant g used as an instrumental variable, the MR effect size can be estimated by the Wald ratio estimate

$$\hat{\beta}^{MR}_{\text{Wald}} = \frac{\hat{\beta}_{Y|g}}{\hat{\beta}_{X|g}} \tag{4.22}$$

with a standard error

$$\text{SE}(\hat{\beta}^{MR}_{\text{Wald}}) = \frac{\text{SE}(\hat{\beta}_{Y|g})}{\hat{\beta}_{X|g}}$$

based on the approximation by the delta method. Even though this can be estimated from individual data, another option is to use increasingly publicly available summary statistics from large GWAS to find both $\hat{\beta}_{X|g}$ and $\hat{\beta}_{Y|g}$ and their respective standard errors, as proposed by Burgess et al (2013).

In the case of MR using multiple genetic variants $g = 1, \dots, G$ and aligning the effects so that $\hat{\beta}_{X|g} > 0 \ \forall \ g$, the inverse-variance weighted (IVW) estimate is defined as

$$\hat{\beta}_{IVW} = \frac{\sum_g \hat{\beta}_{X|g}\hat{\beta}_{Y|g}\text{SE}(\hat{\beta}_{Y|g})^{-2}}{\sum_g \hat{\beta}^2_{X|g}\text{SE}(\hat{\beta}_{Y|g})^{-2}} \tag{4.23}$$

with a standard error

$$\text{SE}(\hat{\beta}_{IVW}) = \sqrt{\frac{1}{\sum_g \hat{\beta}^2_{X|g}\text{SE}(\hat{\beta}_{Y|g})^{-2}}}$$

approximated by the delta method (Burgess et al., 2013). This is equivalent to regressing

the SNP-outcome effect size estimate $\beta_{Y|g}$ on SNP-exposure effect size estimate $\hat{\beta}_{X|g}$ without an intercept and using the standard error of $\hat{\beta}_{Y|g}$ as weights:

$$\mathbf{Y}' = \mathbf{X}'\hat{\beta}_{IVW} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(0, (\sigma^2/w_g)\mathbf{I}), \tag{4.24}$$

where $\mathbf{Y}' = (\hat{\beta}_{Y|1} \cdots \hat{\beta}_{Y|G})^T$, $\mathbf{X}' = (\hat{\beta}_{X|1} \cdots \hat{\beta}_{X|G})^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1 \cdots \varepsilon_G)^T$ and $w_g = \mathrm{SE}(\hat{\beta}_{Y|g}), g = 1, \ldots, G$.

4.3.3.3 Key limitations of MR

The IVW method provides an efficient estimate when all of the instrumental variable assumptions (i), (ii) and (iii) hold. Violations to these assumptions may lead to biased MR estimates (Burgess & Thompson, 2013).

Assumption (i) is straightforward to verify, as the variants used are selected based on their association with the exposure. However, if the instrumental variable used explains only a small proportion of the exposure, this can lead to a *weak instrument bias* (Burgess & Thompson, 2011). In one-sample MR, weak instrument bias is towards the confounded observational estimate of exposure-outcome relationship (Burgess & Thompson, 2011). In two-sample MR, the same bias is towards the null (Lawlor, 2016). In addition, using a weak instrument also leads to limited power to detect an association.

One way to measure the strength of the instrument is via F statistic:

$$F = \frac{R^2(n-1-k)}{(1-R^2)k},$$

where k is the number of instrumental variables, n is the sample size for which the instruments are based on, and

$$R^2 = \frac{2\beta_{X|g}^2 \mathrm{MAF}_g(1-\mathrm{MAF}_g)}{\mathrm{Var}(Y)}$$

is the variance explained by the instrumental variable. $F > 10$ has been suggested as sufficient (Lawlor et al., 2008), however, as shown by Burgess & Thompson (2011), this might not completely eliminate weak instrument bias, and the authors advise to conduct sensitivity analyses with different methods where possible.

Instrumental variable assumptions (ii) and (iii) cannot be proven explicitly. Horizontal pleiotropy (Figure 4.3 (a)) violates instrumental variable assumption (iii), that is, there is an association between the genetic instrumental variable and the outcome independently of the exposure. For MR methods using multiple genetic variants, robust methods have been developed to allow some degrees of violation to this assumption (Section 4.3.3.4).

Possible other sources of bias in MR analysis include dynastic effects (Sanderson et al., 2019), assortative mating (Hartwig et al., 2018), and insufficient adjustment to population stratification in the source GWAS (Section 4.3.1.2) (Lawson et al., 2020). If the variant-exposure estimates $\hat{\beta}_{X|g}$ are derived from summary statistics that are adjusted for other heritable traits, there is a risk of introducing collider bias in MR (Day et al., 2016; Gkatzionis & Burgess, 2018).

4.3.3.4 Sensitivity analyses to assess violations to instrumental variable assumptions

There is a range of sensitivity analyses available for MR (Slob & Burgess, 2020). Here, I present three commonly used methods, with different types of consistencies: MR-Egger, weighted median, and MR-PRESSO (Pleiotropy RESidual Sum and Outlier).

MR-Egger method proposed by Bowden et al. (2015) regresses variant-outcome effect size estimates $\hat{\beta}_{Y|g}$ on variant-exposure effect size estimate $\hat{\beta}_{X|g}$, weighted by $\text{SE}(\hat{\beta}_{Y|g})$, as in IVW method. Unlike in the IVW method, the intercept is also estimated in the regression:

$$Y_g^{'} = \hat{\beta}_{0E} + \hat{\beta}_{\text{Egger}} X_g^{'} + \varepsilon_g, \varepsilon_g \sim N(0, \sigma^2/w_g)$$

for $g = 1, \ldots, G$. The estimated intercept $\hat{\beta}_{0E}$ can be interpreted as the estimate for the average pleiotropic effect across the variants. Testing whether $\hat{\beta}_{0E}$ equals zero serves as a test for horizontal pleiotropy. MR-Egger gives consistent MR estimates if the InSIDE (Instrument Strength Independent of Direct Effect) assumption holds (Bowden et al., 2015). Burgess & Thompson (2017) suggest that MR-Egger is considered only as a sensitivity analysis after there is evidence for a causal effect from IVW analysis.

The weighted median method proposed by Bowden et al. (2016) does not have a straightforward representation as a regression. Let $\hat{\boldsymbol{\beta}}_{(g)}^{MR}$ be an ordered sequence of ratio estimators

$\hat{\beta}_{Y|g}/\hat{\beta}_{X|g}$, and $\mathbf{w}_{(g)}$ be the corresponding weights, with $w_g = \hat{\beta}^2_{X|g}\text{SE}(\hat{\beta}_{Y|g})^{-2}/W$, where $W = \sum_g \hat{\beta}^2_{X|g}\text{SE}(\hat{\beta}_{Y|g})^{-2}$ so that $\sum w_g = 1$. The weighted median estimator is the median of a distribution that has $\hat{\beta}_{(g)}$ as its $100(1 - \frac{w_{(g)}}{2})$th percentile. Standard errors and p-values for the weighted median estimate can be calculated via bootstrapping (Section 4.1.1.1) (Bowden et al., 2016). Weighted median method provides consistent estimates if at least half of the weight for the analysis comes from valid instrumental variables (Bowden et al., 2016).

MR-PRESSO proposed by Verbanck et al. (2018) is based on the regression representation of IVW estimate (equation (4.24)). The residual sum of squares (RSS) are calculated for each variant g by:

$$\text{RSS} = \sum_{g=1}^{G} \text{RSS}_g = \sum_{g=1}^{G} (\hat{\beta}_{Y|g} - \hat{\beta}^{-j}_{IVW}\hat{\beta}_{X|g})$$

where $\hat{\beta}^{-j}_{IVW}$ is the IVW estimate when excluding genetic variant g. An expected distribution of RSS is then based on simulated RSS (SimRSS) of K random samples, under the null hypothesis of no pleiotropy:

$$\text{SimRSS}^k = \sum_{g=1}^{G} \text{SimRSS}^k_g = \sum_{g=1}^{G} (\tilde{\beta}_{Y|g} - \hat{\beta}^{-j}_{IVW}\tilde{\beta}_{X|g}),$$

where

$$\tilde{\beta}_{X|g} \sim N(\hat{\beta}_{X|g},\ \text{SE}(\hat{\beta}_{X|g})^2)$$

and

$$\tilde{\beta}_{Y|g} \sim N(\hat{\beta}^{-j}_{IVW}\hat{\beta}_{X|g}, \text{SE}(\hat{\beta}_{Y|g})^2)$$

for $k = 1, \ldots, K$.

The observed RSS is then compared to the distribution of the SimRSS, and an empirical p-value is computed as

$$p_E = \frac{\sum_k \mathbb{1}\left(\text{RSS} > \text{SimRSS}^k\right)}{K}$$

For MR-PRESSO Outlier test, the p-values are computed variant-wise:

$$p_{Eg} = \frac{\sum_k \mathbb{1}\left(\text{RSS}_g > \text{SimRSS}_g^k\right)}{K}$$

The variants which show evidence against the null are removed, and the MR-PRESSO estimate is the IVW estimate without the pleiotropic variants. This method assumes both of the assumptions of weighted median and MR-Egger methods (Verbanck et al., 2018).

Another way to examine the violations to instrumental variable assumptions is to utilise *negative control outcomes* (Hemani, Bowden, & Davey Smith, 2018; Lipsitch et al., 2010). The negative control outcome should be an outcome for which there should be no plausible association with the exposure. Any association would indicate invalidity of the used genetic variants as instrumental variables (Hemani, Bowden, & Davey Smith, 2018).

In the presence of multiple genetic variants as instrumental variables, Burgess et al. (2020) advises to use IVW estimate as the main MR analysis, as it is the most powerful MR method when all variants are valid instrumental variables. For those analyses that show association between genetically predicted exposure and the outcome, it is suggested to conduct further sensitivity analyses with different MR methods (that make different assumptions with respect to underlying horizontal pleiotropy) to assess the robustness of the results.

4.3.4 Causal pathways and mediation analysis

The aim of mediation analysis is to examine the potential mechanisms of a causal association between the exposure X and the outcome Y. The effect of the exposure is divided into two separate effects – an indirect effect which affects the outcome via a *mediator* M, and a direct effect that is separate from the mediated effect. The mediation model also implies that the mediator is causally affected by the exposure.

Figure 4.6 depicts the hypothesised mediation. Mediation can be complete or partial. Complete mediation implies that M is the only mechanism of the effect of X on Y and that the direct effect of X on Y is null. In partial mediation – which is more likely in practice – X

may affect Y via other paths as well.

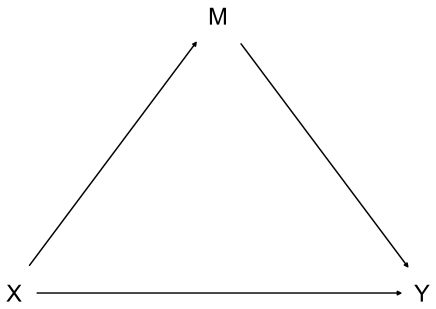

Figure 4.6: A DAG on mediation. A mediator M is on the causal pathway from X to Y. In the case of full mediation, there would be no direct arrow from X to Y.

As an example, exposure to maternal smoking is associated with offspring ADHD (Huang et al., 2018). There is also evidence that maternal smoking has an impact on offspring DNA methylation (Joubert et al., 2016). Thus, DNA methylation could be a mechanistic pathway from exposure to maternal smoking during pregnancy to offspring ADHD, i.e. mediate the association between exposure to maternal smoking and offpsring ADHD. In addition, MR can also be interpreted in terms of mediation, where it is assumed that the exposure completely mediates the effect of genetic instrumental variable on the outcome.

The mediator of interest is determined by the scientific theory under investigation. In regression modelling, a mediator is mathematically identical to adding a confounder in the model. Thus, it is important to motivate the hypothesised data generating process carefully, especially when using observational data.

There are different strategies to estimate mediation effects, reviewed by Preacher (2015). The approach taken here is model-based causal mediation analysis (Pearl, 2001; Robins & Greenland, 1992).

Imai, Keele, & Yamamoto (2010) described a sequential ignorability assumption for estimating average causal mediated effects. This assumption implies that, when confounders C for all of exposure-outcome, exposure-mediator and mediator-outcome associations are taken into account (Forastiere et al., 2018; VanderWeele, 2016), the average causal medi-

ated effect can be estimated (Imai, Keele, & Yamamoto, 2010). Whereas non-parametric inference is also possible, here I present only the estimation for parametric inference using a quasi-Bayesian Monte Carlo approximation (Imai, Keele, & Tingley, 2010).

Let $\mathbf{X}$, $\mathbf{C}$, $\mathbf{M}$ and $\mathbf{Y}$ be the observed values for the exposure, confounders, the mediator and the outcome, respectively. First, a regression model is fitted for the mediator

$$\mathbf{M} = \mathbf{X}\beta^M + \mathbf{C}\boldsymbol{\gamma}^M \tag{4.25}$$

and the outcome

$$\mathbf{Y} = \mathbf{X}\beta^Y + \mathbf{M}\phi^Y + \mathbf{C}\boldsymbol{\gamma}^Y \tag{4.26}$$

Let $\boldsymbol{\theta}_M$ and $\boldsymbol{\theta}_Y$ present the model parameters for equations (4.25) and (4.26), respectively. In the second step, parameters for both models are simulated J times from their sampling distributions, with $\boldsymbol{\theta}_M^{(j)}$ and $\boldsymbol{\theta}_Y^{(j)}$ denoting the jth simulation. Finally, for each $j = 1, \ldots, J$,

1. Potential values are simulated for the mediator
2. Potential outcomes are are simulated, conditional on the simulated mediator values
3. Causal mediation effects are calculated based on the simulated values for the mediator and the outcome.

The mediation effect of interest is $\hat{\beta}^M\hat{\phi}^Y$. Summary statistics for the mediation effects can be computed from the simulated distribution. Further details are given in Imai, Keele, & Tingley (2010). The strong assumptions of no unmeasured confounding and no measurement error applies also for causal mediation analysis.

4.3.5 Triangulation

Other methods for causal inference from observational studies not presented here include cross-context comparisons (Gage et al., 2016), family-based designs (Pingault et al., 2018), and inverse probability weighting (Hernán & Robins, 2019). In any case, causal inference should never be based on a single piece of evidence from a specific analysis.

A famous list of principles for evaluating causality in epidemiology was presented by Austin Bradford Hill (1965). According to Hill, the nine aspects of association are strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment and analogy, commonly known as 'Bradford Hill criteria' (Fedak et al., 2015). These aspects can be used for evaluating potential causality of the exposure–outcome relationship of interest. However, as acknowledged by Hill, the list should not be used as checklist to be fulfilled. Instead, these aspects are merely ways to examine whether causality is the most reasonable inference. It should be also noted that none of the aspects are sufficient alone to infer causality.

Lawlor et al. (2017) proposed *triangulation* to combine evidence from different analytical strategies. Triangulation is the practice of integrating evidence from multiple approaches – with different underlying assumptions and unrelated key sources of bias – to answer a research question. Triangulation requires at least two – but ideally more – modelling approaches that try to answer the same research question and that have different sources of bias. These biases should be explicitly acknowledged when comparing the results, and, if feasible, the expected direction of all key sources of bias should be made clear.

Results from various methods with different sources of bias are compared with respect to the research question. Triangulation would provide qualitative evidence for causality if the results from the different methods converge to a causal explanation. If the results from different methods do not agree, understanding the key sources of bias of the approaches may help in identifying further research questions (Lawlor et al., 2017).

Chapter 5

Maternal smoking and offspring DNA methylation[1]

Exposure to maternal smoking during pregnancy is associated with many adverse birth and offspring later life outcomes (Hofhuis et al., 2003; Obel et al., 2008). Cigarette smoking is a known modifier of individual's DNA methylation profiles (Joehanes et al., 2016), and Joubert et al. (2016) reported associations between exposure to maternal smoking and differential cord blood DNA methylation at 6,073 CpG sites. Whether this differential DNA methylation changes are stable throughout the lifecourse of the offspring is not clearly known. It is also hypothesised that DNA methylation acts as a mechanism for the associations between intratuterine smoke exposure and offspring health outcomes (Knopik et al., 2012).

In this chapter, I examined 1) whether the CpG sites associated with exposure to maternal smoking during pregnancy persist into adolescence and adulthood, and 2) whether differential DNA methylation mediates the effect of exposure to maternal smoking on later life disease outcomes.

5.1 Methods

Table 5.1 shows the analytical steps in this study and data availability for each step. Data were available from five European cohorts, with DNA methylation in the offspring measured

[1]The material in this chapter has been published in Wiklund et al. (2019).

between 16 and 31 years in the main analysis, and up to 48 years in the longitudinal analysis. I conducted all analyses for NFBC1966 and NFBC1986 cohorts, and pooled all results using inverse-variance weighted fixed-effects meta-analysis.

Table 5.1: Data availability.

Study (mean age in years)	Meta-analysis $N = 2{,}821$	Never-smokers only $N = 1{,}325$	Negative control analysis $N = 1{,}774$	Longitudinal analysis* $N = 1{,}220$	Mendelian Randomisation	Mediation analysis $N = 636$
NFBC1986 (16)	Y	Y	Y			
NFBC1966 (31)	Y	Y	Y	Y†		Y
ALSPACm (30)	Y	Y	Y	Y‡	Y	
ALSPACc (17)	Y	Y	Y		Y	
IOWBC (18)	Y	Y				
MR-Base					Y	

Y indicates that the data were used in the corresponding analysis.
* Longitudinal analysis for the persistence of the effects.
† Mean age 46 years.
‡ Mean age 48 years.

5.1.1 Association analysis for exposure to maternal smoking during pregnancy and offspring DNA methylation

Association analysis was conducted on the exposure maternal smoking during pregnancy and 6,073 CpG sites identified by Joubert et al. (2016). Offspring DNA methylation was the outcome in linear regression model, and maternal smoking was the exposure of interest. Additional explanatory variables used were offspring sex, BMI, smoking status, SES (at the time of DNA methylation measurement), first four genetic PCs, technical covariates related to DNA methylation, and white blood cell subpopulation estimates. A threshold for CpGs to be picked for subsequent senstivity analysis was set at p-value $< 1 \times 10^{-7}$, which corresponds approximately to an epigenome-wide Bonferroni-correction for 450,000 independent tests (Lehne et al., 2015). The conservative multiple testing correction was used for a stringent control of false positives, and the CpGs with p-value $< 1 \times 10^{-7}$ were deemed worthy of further examination.

5.1.2 Associations within never-smokers

As cigarette smoking is known to affect DNA methylation levels (Joehanes et al., 2016), there is a possibility that the detected associations are due to participants' own smoking status. In order to examine this possibility, the same analysis were conducted excluding all

participants who reported smoking regularly.

5.1.3 Paternal smoking as a negative control

The same analysis was conducted using paternal smoking status as a negative control exposure. Assuming a biological effect of intrauterine smoke exposure on offspring DNA methylation, one would expect the effects of paternal smoking on offspring DNA methylation be negligible, whilst still having a similar sources of confounding (Lipsitch et al., 2010). If the effects are of similar magnitude, it is likely that the detected effects of exposure to maternal smoking during pregnancy and offspring DNA methylation are due to unmeasured confounding (Section 4.3.1.1).

5.1.4 Longitudinal analysis

To examine whether the differential DNA methylation related to maternal smoking persists into middle age, longitudinal analysis were conducted using NFBC1966 data at 31 and 46 years, and ALSPACm data at 30 and 48 years. DNA methylation at each time point was regressed on technical and white blood cell covariates. The corresponding residuals were used as the outcome in a generalised least squares model, with exposure to maternal smoking as the exposure of interest. Offspring sex, smoking, BMI and family social class at each time point, the time point of measurement and its interaction with the exposure were added as covariates in the model. The model residuals were allowed to be correlated within each individual and be heteroskedastic between time points:

$$y_{it}^{res} = \beta_0 + \beta_1 x_i + \beta_2 T_t + \beta_3 x_i T_t + \beta \mathbf{z}_{it} + \varepsilon_{it},$$

where $i = 1, \ldots, n$, y_{it}^{res} are the residuals from regressing the DNA methylation values on technical and white blood cell covariates, x_i is the indicator variable for exposure to maternal smoking, $T_t, t = 1, 2$ is the indicator for time point of measurement ($T_1 = 31$ years, $T_2 = 46$ years), $\mathbf{z}_{it}$ is the matrix for the other covariates (sex, BMI, smoking status, SES and first four genetic PCs) and

$$\begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right).$$

The effect estimates for maternal smoking at time points 31 and 46 years are β_1 and $\beta_1 + \beta_3$, respectively. Thus, the equality of the effect sizes at each individual time point can be tested by whether $\beta_3 = 0$. Generalised least squares model can be thought of as a linear mixed model without random effects (Pinheiro & Bates, 2000).

5.1.5 MR for the effect of DNA methylation on disease outcomes

I conducted MR (Section 4.3.3) to examine putative causal effects of maternal smoking related DNA methylation on 106 offspring disease outcomes for which GWAS summary statistics were available in MR-Base. Based on the association analysis described above, 69 CpG sites were associated with exposure to maternal smoking, for which methylation quantitative trait loci (mQTL) were sought from publicly available ARIES database (Gaunt et al., 2016; Relton et al., 2015). I examined the mQTLs with p-value $< 1 \times 10^{-7}$ at any of four different life stages found in the database – birth, childhood, adolescence and middle age. After clumping SNPs (1 Mb window, $r^2 < 0.001$) and pruning the CpG sites to one per locus (1 Mb window), mQTLs were found for 15 CpG sites. The consistency of the SNP-CpG associations across the four time points was examined, and the time-varying associations were excluded from further analyses. The SNP-CpG effect size estimates were extracted from ARIES middle age data point for the MR analysis. The SNP-outcome effect size estimates were extracted from MR-Base platform (Hemani, Zheng, et al., 2018). MR estimates were calculated using Wald ratio or, in case of multiple mQTL available, IVW method. To correct for multiple testing, FDR correction was applied with $\alpha = 0.05$.

5.1.6 Mediation analysis

To further examine the putative causal associations detected in MR, I conducted mediation analysis (Section 4.3.4) between maternal smoking during pregnancy and disease outcomes, with DNA methylation as a mediator, in NFBC1966. Model-based causal mediation as described by Imai, Keele, & Yamamoto (2010) was performed by first estimating the effect of maternal smoking on DNA methylation at a CpG site, and then estimating the effect

of the CpG site on the outcome, adjusted for maternal smoking. The additional covariates in both models were offspring sex, smoking status and techinal covariates related to DNA methylation data. Estimates for the total effect, average direct effect, and average causal mediation effect were calculated using quasi-Bayesian Monte Carlo method based on normal approximation with 2,000 simulations and robust standard errors (Tingley et al., 2014). The estimated proportion mediated was calculated as described in Imai, Keele, & Yamamoto (2010).

5.2 Results

The study participant characteristics for NFBC1966 and NFBC1986 are shown in Tables A.2 and A.3, respectively. After meta-analysing the results from all five participating cohorts, there were 69 CpG sites in 36 genomic regions with p-value $< 1 \times 10^{-7}$ for association with exposure to maternal smoking (Table 5.2). All of these CpG sites showed directionally concordant effects with the ones reported earlier for newborn cord blood DNA methylation (Joubert et al., 2016).

When conducting the same association analysis within never-smokers only, the effect size estimates were highly similar, both in magnitude and direction (Figure 5.1). In the negative control analyses, using paternal smoking as the negative control exposure, the directions of the effect size estimates were similar, however the estimates for the negative control exposure were notably smaller (Figure 5.2). The results from longitudinal analyses showed no evidence for change in direction or magnitude of associations between the two time points (Figure 5.3).

Table 5.2: Lead CpG sites within 1 Mb window with $p < 1 \times 10^{-7}$ for the association of exposure to maternal smoking during pregnancy and offspring peripheral blood DNA methylation.

CpG	Chr	Position	Gene	Effect size (SE)	p-value
cg19089201	7	45002287	*MYO1G*	0.035 (0.003)	1.20e-31
cg05549655	15	75019143	*CYP1A1*	0.010 (0.001)	1.00e-28
cg25949550	7	145814306	*CNTNAP2*	-0.007 (0.001)	5.40e-27
cg18493761	11	125386885		0.037 (0.004)	7.30e-21
cg14179389	1	92947961	*GFI1*	-0.028 (0.003)	5.00e-20
cg00253658	16	54210496		0.037 (0.004)	3.40e-19
cg11813497	10	14372879	*FRMD4A*	0.026 (0.003)	9.30e-19
cg05575921	5	373378	*AHRR*	-0.019 (0.002)	3.50e-18
cg11207515	7	146904205	*CNTNAP2*	-0.023 (0.003)	1.00e-13
cg14157435	2	206628692	*NRP2*	-0.046 (0.006)	1.10e-13
cg01952185	5	134813213		0.019 (0.003)	3.30e-12
cg05204104	2	235403141	*ARL4C*	0.017 (0.003)	4.80e-12
cg05697249	11	111789693	*C11orf52*	0.015 (0.002)	1.30e-11
cg11025974	2	152830521	*CACNB4*	0.016 (0.002)	9.10e-11
cg25189904	1	68299493	*GNG12*	-0.020 (0.003)	1.30e-10
cg06758350	21	36259460	*RUNX1*	0.030 (0.005)	1.70e-10
cg00794911	6	166260532		-0.012 (0.002)	5.20e-10
cg14563637	9	98931801		0.016 (0.003)	1.20e-09
cg12984635	19	44032076	*ETHE1*	0.015 (0.002)	2.20e-09
cg17199018	8	28206278	*ZNF395*	-0.017 (0.003)	3.90e-09
cg00174179	3	49450293	*RHOA;TCTA*	-0.006 (0.001)	7.00e-09
cg15578140	7	147718109	*MIR548F3;CNTNAP2*	0.011 (0.002)	7.50e-09
cg14540913	9	132458514	*PRRX2*	0.013 (0.002)	7.60e-09
cg21253335	5	87835928		0.017 (0.003)	1.30e-08
cg25879142	7	4671391		0.018 (0.003)	1.80e-08
cg20117519	7	8429907		0.022 (0.004)	2.40e-08
cg05783384	2	218843735		0.021 (0.004)	2.60e-08
cg13822849	9	137999757	*OLFM1*	0.007 (0.001)	2.90e-08
cg16449012	4	17781880	*FAM184B*	0.014 (0.002)	3.10e-08
cg05634495	6	122364658		0.016 (0.003)	3.10e-08
cg13834112	15	90361639		0.013 (0.002)	3.60e-08
cg06635952	2	70025869	*ANXA4*	0.011 (0.002)	5.50e-08
cg04598670	7	68697651		-0.019 (0.003)	8.30e-08
cg04749740	2	65935124		0.015 (0.003)	9.20e-08
cg15325070	1	2792704		0.014 (0.003)	9.20e-08
cg04358214	16	67143304	*C16orf70*	0.022 (0.004)	9.60e-08

Chr = chromosome; SE = standard error.

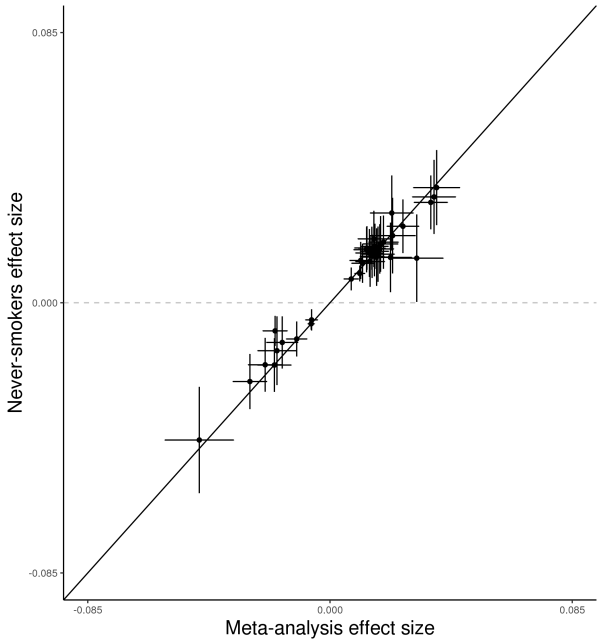

Figure 5.1: Comparison of meta-analysis effect size estimates and their 95% confidence intervals in all participants (x-axis) and never-smokers (y-axis) for the 36 top CpG sites. All effect size estimates are adjusted for study-specific covariates as necessary and meta-analysed using inverse-variance weighted fixed-effects model.

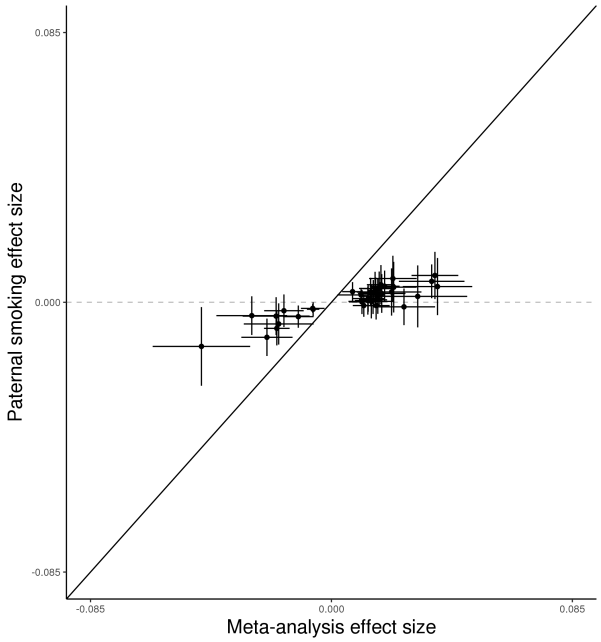

Figure 5.2: Comparison of meta-analysis effect size estimates and their 95% confidence intervals for exposure to maternal smoking (x-axis) and exposure to paternal smoking (y-axis) for the 36 top CpG sites. All effect size estimates are adjusted for study-specific covariates as necessary and meta-analysed using inverse-variance weighted fixed-effects model.

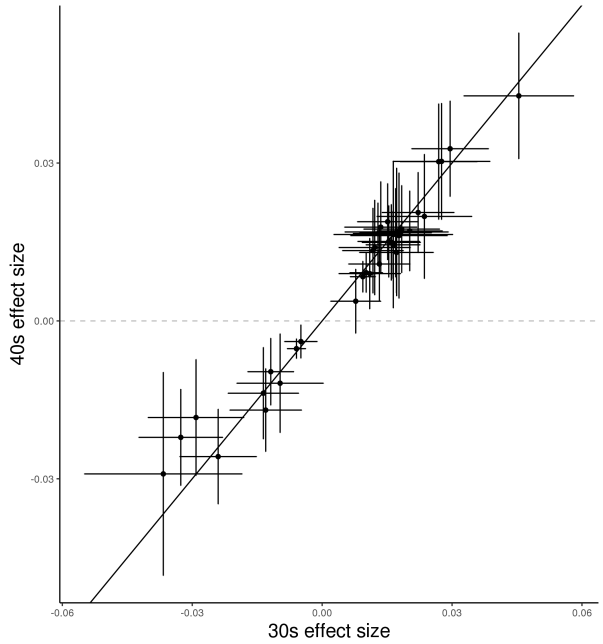

Figure 5.3: Comparison of meta-analysis effect size estimates and their 95% confidence intervals at age 30-31 years (x-axis) and age 46-48 years (y-axis) for the 36 top CpG sites. All effect size estimates are adjusted for study-specific covariates as necessary and meta-analysed using inverse-variance weighted fixed-effects model.

There were mQTLs for 15 out of 36 CpG sites associated with exposure to maternal smoking. When examining the stability of the SNP–CpG associations throughout different time points within ARIES dataset, the associations were consistent between time points, apart from rs4306016–cg01825213 association, which was excluded from the final MR analysis (Figure 5.4).

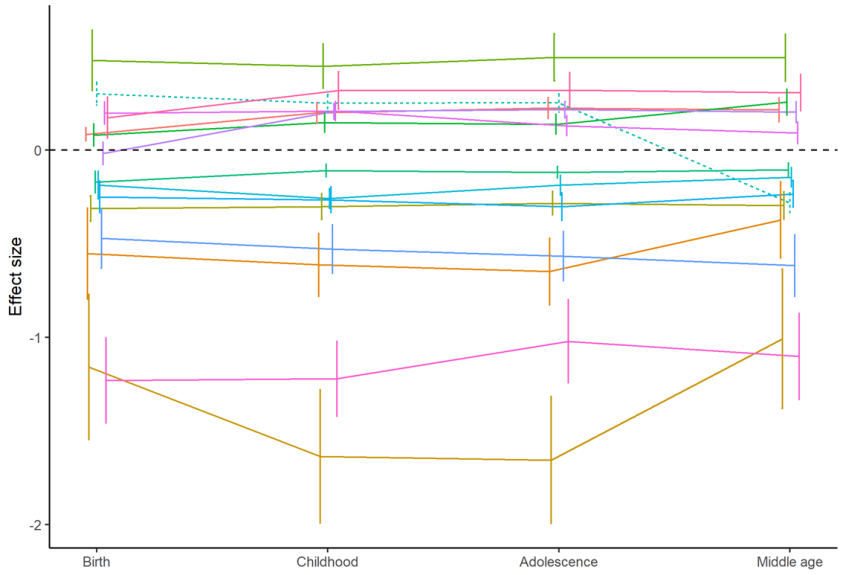

Figure 5.4: Effect sizes and their 95% confidence intervals of each available SNP-CpG association across different time point in the ARIES data. Horizontal lines represent the same SNP-CpG association at each time point. Dotted green line indicates SNP-CpG association that was not consistent across all time points.

MR analyses for the effect of 14 CpGs (mQTL available with longitudinally consistent effect sizes) on 106 disease outcomes (available in MR-Base) showed evidence for associations between three genetically predicted CpG sites (cg15578140 in microRNA 548f-3 (*MIR548F3*), cg09935388 in Growth Factor Independent Protein 1 (*GFI1*) and cg04598670 (unknown gene)) and the risk of inflammatory bowel disease, and genetically predicted cg25189904 (in Guanine Nucleotide Binding Protein, Gamma 12, *GNG12*) and the risk of schizophrenia ($p_{FDR} < 0.05$, Table 5.3).

Table 5.3: MR results with $p_{\text{FDR}} < 0.05$ for the effect of CpGs associated with intrauterine smoke exposure on 106 disease outcomes from MR-Base platform. The sign of the effect sizes are for DNA methylation-outcome associations; the direction of the association between exposure to maternal smoking and offspring disease outcome depends also on the exposure to maternal smoking–DNA methylation associations, see Table 5.2.

Exposure	Gene	Disease	Beta	SE	p-value	p_{FDR}
cg15578140	*MIR548F3*	Inflammatory bowel disease	-0.104	0.018	3.73e-09	2.54e-07
cg09935388	*GFI1*	Inflammatory bowel disease	-0.152	0.034	7.27e-06	2.18e-04
cg04598670		Inflammatory bowel disease	-0.410	0.091	7.27e-06	5.24e-04
cg09935388	*GFI1*	Crohn's disease	-0.162	0.040	4.74e-05	7.12e-04
cg04598670		Crohn's disease	-0.439	0.108	4.74e-05	1.71e-03
cg09935388	*GFI1*	Ulcerative colitis	-0.160	0.042	1.47e-04	1.47e-03
cg04598670		Ulcerative colitis	-0.433	0.114	1.47e-04	3.52e-03
cg25189904	*GNG12*	Schizophrenia	-0.222	0.053	3.37e-05	1.82e-03

Beta = effect size estimate; SE = standard error; FDR = false discovery rate.

In the mediation analyses results, there was evidence for cg25189904 mediating the association between exposure to maternal smoking and Bipolar II Scale (p-value = 0.024) and Hypomanic Personality Scale (p-value = 0.018). The estimated mediated proportions were 30% and 28%, respectively (Figure 5.5).

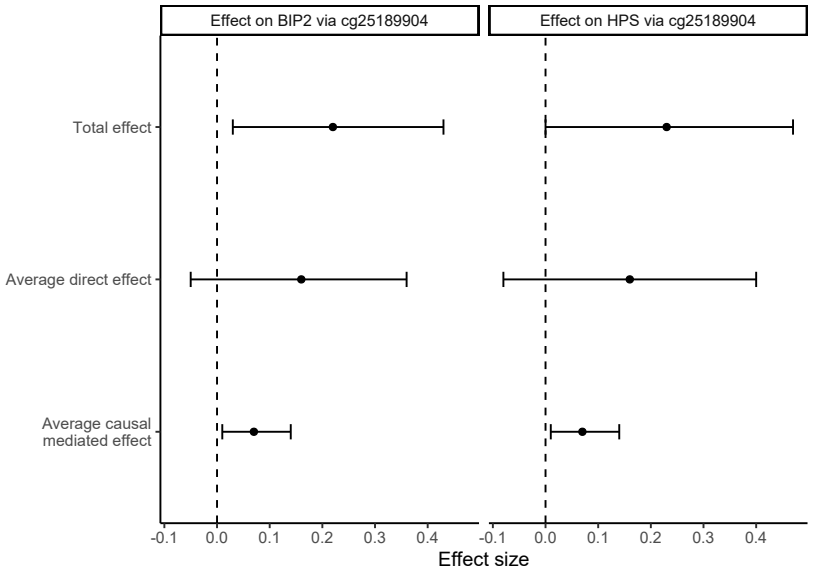

Figure 5.5: Effect size estimates and their 95 % confidence intervals for mediation analysis examining the indirect effect of exposure to maternal smoking during pregnancy on Bipolar II Scale (BIP2, left panel) and Hypomanic personality scale (HPS, right panel) through differential methylation of cg25189904 in *GNG12*.

5.3 Discussion

I examined the association between exposure to maternal smoking during pregnancy and offspring blood DNA methylation in adolescents and adults. The results combined across five cohorts showed evidence for differential DNA methylation at 69 CpGs in 36 genomic regions. These CpGs showed robust associations across sensitivity analysis within never-smokers, and no evidence for association with paternal smoking or longitudinal changes. MR and mediation analyses suggested that differential DNA methylation may play a role in the observed association between maternal smoking during pregnancy and an increased risk of psychiatric morbidity in the exposed offspring.

Earlier evidence has showed that exposure to maternal smoking during pregnancy is associated with differential offspring blood DNA methylation in newborns, children and adoles-

cents (Joubert et al., 2016; Küpers et al., 2015; K. W. Lee et al., 2015; Richmond, Simpkin, et al., 2014). The findings presented here provide evidence that these associations persist up to several decades following the exposure (Richmond et al., 2018; Tehranifar et al., 2018).

Parental smoking is known to associate with offspring's smoking behaviour also via genetic predisposition (A. E. Taylor, Howe, et al., 2014). It is therefore possible that differential DNA methylation act as a mediator between parental smoking and offspring smoking. I conducted sensitivity analyses for offspring that reported no smoking during their lives. The results were similar across all CpG sites, which highlights differential DNA methylation independently of offspring's own smoking status.

Paternal smoking was used as a negative control to evaluate the extent of potential unmeasured confounding due to familial effects, such as post-natal passive smoke exposure or genetic predisposition. The effect sizes for paternal smoking were notably smaller, indicating that it is unlikely that the detected associations were attributable to unmeasured familial confounding.

Longitudinal analyses using blood DNA methylation data from two time points showed no evidence for change in the associations. These results imply that the detected changes in blood DNA methylation due to intrauterine smoke exposure may be irreversible (Richmond et al., 2018).

I performed MR to assess whether the detected differential DNA methylation is implicated with the risk of later life disease outcomes. There was evidence for potential impact of three CpGs (cg15578140, cg09935388, cg04598670) on the risk of inflammatory bowel disease and one CpG (cg25189904) on the risk of schizophrenia. Further mediation analysis provided evidence for differential methylation in cg25189904 mediating the effect of maternal smoking on offspring Bipolar II Scale and Hypomanic Personality Scale. These results are in line with earlier evidence showing association between maternal smoking during pregnancy and an increased risk of offspring psychiatric morbidity (Ekblad et al., 2010; Lahti et al., 2009; Niemelä et al., 2016; Talati et al., 2013).

There is some experimental evidence that suggests *GNG12* acting as a regulator of inflammatory signalling in microglia cells, which are the resident macrophages of the central nervous system (Larson et al., 2010). Furthermore, inflammation is suggested to play a role in the aetiology of schizophrenia (Müller et al., 2015). Further studies are needed to investigate

the potential relevance of these findings.

This study has some limitations. The main weakness is that the DNA methylation was measured in whole blood, whereas brain tissue would be the most relevant for psychiatric outcomes. However, differential DNA methylation during prenatal development may affect multiple tissues, and therefore methylation measured in blood may be informative for psychiatric conditions (Aberg et al., 2013). Maternal smoking was determined from self-reported questionnaires, which may have induced under-reporting or recall bias. In the ALSPACm cohort, adult offspring reported their mothers' smoking, and this could also be subject to recall bias.

As a conclusion, these results are consistent with a direct and persistent biological effect of in utero exposure to cigarette smoke on offspring DNA methylation. DNA methylation may represent a biological mechanism through which maternal smoking is associated with an increased risk of psychiatric morbidity in the exposed offspring.

Chapter 6

Exposure to maternal smoking, offspring DNA methylation and ADHD symptoms

As discussed in Section 2.4, the existence of a causal mechanism between exposure to maternal smoking during pregnancy and offspring ADHD is questioned. Particularly, evidence from negative control exposure analysis with paternal smoking indicate that the association is likely due to unmeasured confounding (Thapar & Rice, 2020). Paternal smoking as a negative control exposure in this context serves to examine potential causal intrauterine effects due to maternal smoking (Lipsitch et al., 2010). In the presence of intrauterine causal effect of maternal smoking, no association with paternal smoking would be expected. Any detected association would imply unmeasured genetic and/or family-level environmental confounding between exposure to maternal smoking and ADHD (Davey Smith et al., 2012).

Smoking is a strong modifier of an individual's epigenetic profile (Joehanes et al., 2016). DNA methylation is the most studied epigenetic mechanism, and intrauterine exposure to smoking is known to be associated with differential DNA methylation at 6,073 cytosine-phosphate-guanine (CpG) sites ($p_{\text{FDR}} < 0.05$) in the offspring cord blood (Joubert et al., 2016). Further evidence suggests that these associations are independent of paternal smoking and offspring's own smoking, and that the associations persist into adulthood (Chapter

5) (Richmond et al., 2018; Sikdar et al., 2019; Wiklund et al., 2019).

Differential DNA methylation may contribute to the variation in gene expression between individuals, and consequently to the risk of complex traits (Relton & Davey Smith, 2010). DNA methylation may play a role in the aetiology of ADHD and other mental health outcomes (Barker et al., 2018; Guintivano & Kaminsky, 2016; Hamza et al., 2019), with an increasing number of large-scale human studies examining the associations between epigenome-wide DNA methylation and ADHD being conducted in recent years (Mooney et al., 2020; van Dongen et al., 2019; Walton et al., 2017).

The aim in this chapter was to examine whether DNA methylation is a mediator in the association between exposure to maternal smoking during pregnancy and offspring ADHD. This hypothesis implies a pathway from maternal smoking to offspring ADHD symptoms via DNA methylation, and that the DNA methylation related to intrauterine smoke exposure is independent of familial unmeasured confounders that are common with maternal and paternal smoking (Figure 6.1).

I first conducted association analysis between CpG sites and ADHD symptoms using cross-sectional data at 16 years in NFBC1986. In addition, I developed two alternative offspring DNA methylation risk scores for exposure to maternal smoking. The risk scores and top CpG sites from the association analysis were used as mediators in model-based causal mediation analysis. Finally, I conducted MR analysis to evaluate the causality of differential methylation at CpG site cg05575921 within aryl-hydrocarbon receptor repressor (*AHRR*) gene on the risk of ADHD.

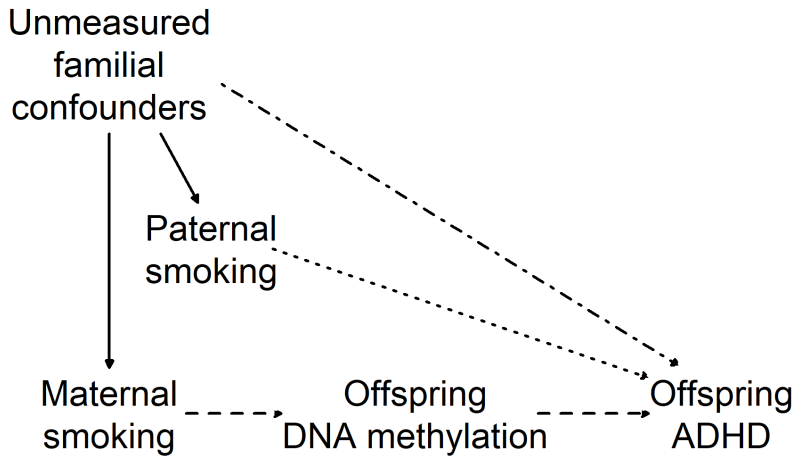

Figure 6.1: A DAG for the analysis of the effect of exposure to maternal smoking during pregnancy on the risk of ADHD in the offspring, mediated by differential DNA methylation. Negative control exposure studies using paternal smoking suggest that the association between intrauterine smoke exposure and offspring ADHD is due to unmeasured familial confounding (dot-dash line), which cause an association between paternal smoking and offspring ADHD (dotted line). An alternative pathway is via differential DNA methylation that is specific to exposure to maternal smoking (dashed line). Other possible confounders affect all nodes in the graph, and these are omitted for clarity.

6.1 Methods

6.1.1 Dataset and outcomes

NFBC1986 dataset was used for individual participant analysis, detailed in Section 3.1. Briefly, DNA methylation was measured from blood samples given by the participants at the clinical examination during the 16-year follow-up. Inattention, hyperactivity, and total ADHD symptoms evaluated by offspring's parents using SWAN questionnaire (Swanson et al., 2012) at the 16-year follow-up were used as outcomes. Complete DNA methylation and ADHD symptom data were available for 432 individuals.

6.1.2 Association analysis

The methylation beta values at each CpG site were used as the exposure and ADHD, inattention, and hyperactivity symptoms as outcomes in separate regression models. To adjust for potential confounding and technical variation, maternal smoking, maternal age, family SES, offspring sex, five first genetic PCs, first 30 PCs of technical probes (Lehne et al., 2015) and white blood cell subpopulation estimates (Houseman et al., 2012) were included as additional explanatory variables in the regression models.

Association analyses were conducted for all 473,864 CpG sites in 450K array which passed the quality control. However, I focus only on the results of 6,073 CpG sites associated with exposure to maternal smoking during pregnancy ($p_{\text{FDR}} < 0.05$) as reported in Joubert et al. (2016). The results for other CpG sites were used only to compare the test statistic distribution within the subset of the selected CpGs. To evaluate the impact of offspring's own smoking status, the analyses were repeated among those who reported to never having smoked regularly.

I hypothesise that in the presence of a true effect of maternal smoking related differential DNA methylation profile on ADHD symptoms, there would be an excess of small effects across the CpGs associated with exposure to maternal smoking in pregnancy. To measure the excess of the small effects, I calculated the inflation factor of the test statistics, as proposed by Devlin & Roeder (1999). The inflation factor compares the observed median of the χ^2 test statistics to the expected value under the null hypothesis of no associations. An inflation factor larger than 1.1 is suggested to show evidence for inflation. Reasons for inflation might be due to either unmeasured confounding, or a real signal that associates with the full DNA methylation profile. The latter is suggested to be likely in EWAS (van Iterson et al., 2017). The inflation factor was calculated separately within all participants for the 6,073 CpG sites and within all participants for other CpG sites. Confidence intervals for the inflation factors were constructed by 1,000 bootstrap samples of size 6,073 within both sets of CpG sites.

CpG sites that showed association with ADHD at a Bonferroni-corrected p-value threshold ($p < 0.05/6073 \approx 8 \times 10^{-6}$) and that were independent of other top hits ($r^2 < 0.1$) were taken forward for futher analysis.

6.1.3 DNA methylation risk scores for exposure to maternal smoking during pregnancy

In an effort to capture the differential offspring DNA methylation that is related to exposure to maternal smoking, I developed DNA methylation risk scores for exposure to maternal smoking, based on offspring blood DNA methylation data. I built prediction models of exposure to maternal smoking as the outcome using logistic regression, with two alternative sets of CpGs as predictors (Figure 6.2). The first set included all 6,073 CpG sites previously associated with exposure to maternal smoking (Joubert et al., 2016). For the second set, I aimed to select only CpG that were specific to exposure to maternal smoking in adolescents and adults. Summary statistics were obtained from ALSPAC and IOWBC studies as in Chapter 5. As the prediction model would be built using NFBC1966 and validated in NFBC1986, the summary statistics from these two datasets were excluded, to avoid doubly using these data and thus biasing the results. The results of 6,073 CpG sites from ALSPAC and IOWBC cohorts were combined using fixed-effects meta-anaylsis. CpG sites with either $p > 0.05$ or a non-concordant direction of effect size compared to the ones in Joubert et al. (2016) were excluded. Furthermore, those CpGs that were either not associated with maternal smoking ($p > 0.05$) within non-smokers, or were associated with paternal smoking ($p < 0.05$) in NFBC1966 dataset were excluded. After these exclusions, the second set included 177 CpG sites.

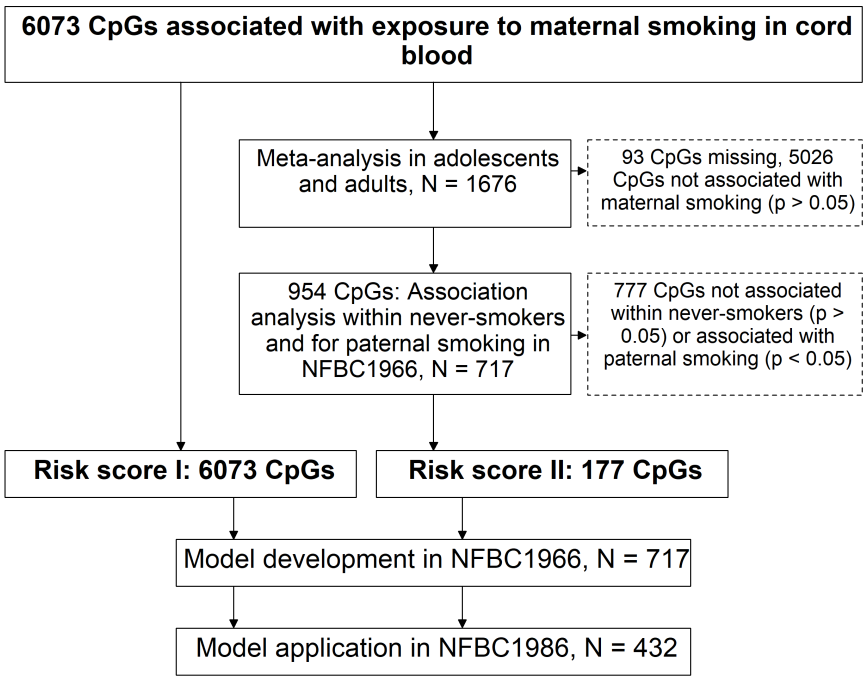

Figure 6.2: A flowchart of the selection of CpGs for DNA methylation risk score development.

The prediction models were built using NFBC1966 dataset with DNA methylation measured at 31 years ($N = 717$), separately for both sets of CpG sites. The methylation beta values at each CpG site were first regressed on 30 first PCs of technical covariates (Lehne et al., 2015) and estimated white blood cell counts (Houseman et al., 2012) to adjust for technical variation and cell type heterogeneity. The resulting residuals were used in the subsequent prediction models.

For prediction model assessment, I compared the model building procedure of five different penalised regression models (Section 4.1.1.1): ridge ($\alpha = 0$), elastic net with $\alpha \in (0.2, 0.5, 0.8)$ and lasso ($\alpha = 1$).

The out-of-sample model performance was estimated by 1,000 repetitions of data splitting (80% training set, 20% test set). The tuning parameter λ was estimated based on 10-fold CV in the training set, and the predictive performance of the model-fitting procedure was evaluated in the test set. I evaluated the out-of-sample model performance by Brier score and C index.

The model with the highest mean C index in the test set was selected as the final model, and the tuning parameter of the final model was estimated by 10-fold CV in the full dataset.

The final model was then used to create a predicted probability of exposure to maternal smoking in NFBC1986 dataset ($N = 516$) based on blood DNA methylation measured at 16 years. The methylation beta values at each CpG site were regressed on technical covariates and white blood cell counts as above, and the predicted probabilities of exposure to maternal smoking were calculated using the penalised regression model with its tuning parameters α and λ estimated in NFBC1966. I compared the two risk scores by Pearson correlation, and evaluated their association with exposure to maternal smoking during pregnancy by C index.

To validate that the DNA methylation scores are specific to exposure to maternal smoking, I compared the associations of exposure to maternal and paternal smoking in a mutually adjusted regression model, with additional adjustments for family SES, maternal age and offspring sex. Any association with paternal smoking would imply that the associations between the risk score and maternal smoking are due to confounding factors.

To evaluate whether offspring's own smoking status distort these associations, I also conducted the same analyses within offspring never-smokers only. A difference in associations between all participants and never-smokers would suggest that offspring's own smoking status drives the exposure–risk score association.

6.1.4 Mediation analysis

I performed model-based causal mediation (Imai, Keele, & Yamamoto, 2010) to evaluate the mediating effect of maternal smoking related DNA methylation on offspring ADHD symptoms. The different mediators tested were: 1) DNAm risk score based on 6,073 CpG sites; 2) DNAm risk score based on 177 CpG sites (Figure 6.2); 3) independent top hits from association analysis. First, the mediator in question was regressed on the exposure

and the outcome was regressed on both exposure and mediator. Maternal age, family SES and offspring sex were added as additional explanatory variables in both exposure–mediator and exposure/mediator–outcome regression models to adjust for confounding. Maternal age was fit as a continuous variable using restricted cubic splines with three knots (Harrell, 2015). The average causal mediated effect was estimated based on 2,000 simulations with robust standard errors, using `mediate` R package (Tingley et al., 2014).

6.1.5 MR analysis

To further examine potential causality for CpGs that were associated with ADHD, I conducted summary statistics based two-sample MR for the effect of CpG site cg05575921 in *AHRR* gene on ADHD. To obtain instruments for cg05575921, I examined an ARIES database on methylation QTL (Gaunt et al., 2016). In addition, I conducted a GWAS for cg05575921 in both NFBC1986 ($N = 508$) and NFBC1966 ($N = 717$) datasets. In each cohort, the CpG site methylation beta values were regressed on technical covariates, blood cell subpopulation estimates and ten first genetic PCs, to adjust for techincal variation, cell type heterogeneity and population stratification. The inverse normal rank transformed residuals were used for association analysis. The results from the two cohorts were meta-analysed using the fixed-effects method. Genome-wide summary statistics for ADHD were obtained from Demontis et al. (2019). For instruments on cg05575921, I selected independent genetic variants ($r^2 < 0.1$ within a 10 Mb window) associated with the methylation value at $p < 5 \times 10^{-8}$ within 200kb of the *AHRR* gene locus and that were available in the ADHD GWAS summary statistics. The strength of the instrument was measured by F statistic. An instrument with $F > 10$ is considered sufficiently powered not to suffer from weak instrument bias (Davies et al., 2018). The MR estimate was calculated using the Wald ratio.

6.2 Results

6.2.1 Association analysis

Figure 6.3 shows the distributions for the total, inattention and hyperactivity ADHD symptoms measured by SWAN scale and the descriptive statistics for other demographic variables

used are in Table A.4. Out of the 6,073 CpG sites associated with exposure to maternal smoking during pregnancy, two CpG sites – cg26703534 and cg05575921 – were associated with ADHD symptoms at $p < 8 \times 10^{-6}$. Both of these CpG sites are within *AHRR* gene, and were correlated with each other (Pearson correlation $r = 0.42$). When restricting the analyses to offspring never-smokers, the effect size estimates were of similar magnitude for cg05575921 (-0.058 in all participants, -0.051 in never-smokers), and somewhat decreased for cg26703534 (-0.084 in all participants, -0.057 in never-smokers), both with wide confidence intervals (Table 6.1). The directions of the associations are negative, which is in line with the association of exposure to maternal smoking and decreased levels of DNA methylation in *AHRR* gene (Joubert et al., 2016). When looking at the inattention and hyperactivity symptoms separately, the top associations were slightly stronger for inattention symptoms (Tables 6.2 and 6.3).

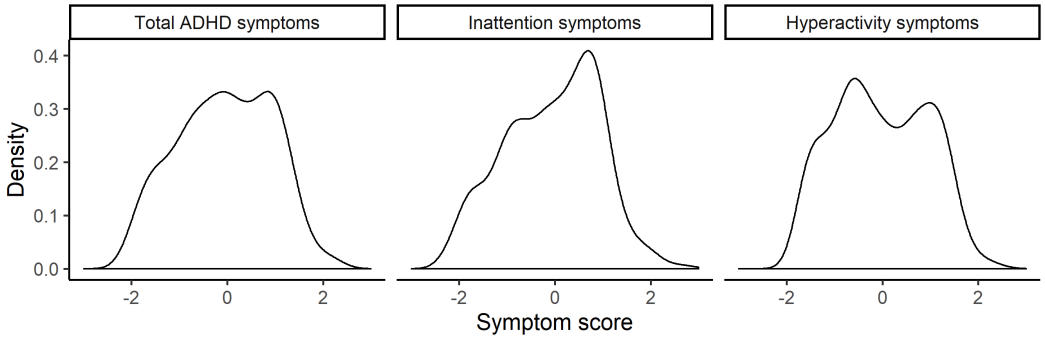

Figure 6.3: Distributions for the total, inattention and hyperactivity symptom scores as measured by SWAN rating scale for those with DNA methylation data available in NFBC1986 ($N = 432$).

QQ-plots showed inflation of the test statistic for the 6,073 CpG sites (inflation factor = 1.29, 95% CI 1.22; 1.37), compared to the rest 467,201 CpG sites (inflation factor = 1.02, 95% CI 0.96; 1.08, Figure 6.4). Similar patterns were observed for inattention and hyperactivity symptoms (Figures 6.5 and 6.6).

Table 6.1: Top results ($p < 8 \times 10^{-6}$) for the associations between CpG sites related to exposure to maternal smoking during pregnancy and offpsring ADHD symptoms.

CpG	Chr	Pos	Gene		*N*	B (95% CI)	*p*-value
cg05575921	5	373378	*AHRR*	All individuals	432	-0.058 (-0.083; -0.033)	6.3e-06
				Never-smokers	321	-0.051 (-0.094; -0.008)	
cg26703534	5	377358	*AHRR*	All individuals	432	-0.084 (-0.119; -0.049)	4.4e-06
				Never-smokers	321	-0.057 (-0.104; -0.011)	

Chr = chromosome; Pos = position; B = effect size estimate, CI = confidence interval.

Table 6.2: Top results ($p < 8 \times 10^{-6}$ for total ADHD symptoms) for the associations between CpG sites related to exposure to maternal smoking during pregnancy and offpsring inattention symptoms.

CpG	Chr	Pos	Gene		*N*	B (95% CI)	*p*-value
cg05575921	5	373378	*AHRR*	All individuals	432	-0.060 (-0.084; -0.035)	3.8e-06
				Never-smokers	321	-0.051 (-0.094; -0.009)	
cg26703534	5	377358	*AHRR*	All individuals	432	-0.087 (-0.122; -0.052)	1.8e-06
				Never-smokers	321	-0.061 (-0.107; -0.014)	

Chr = chromosome; Pos = position; B = effect size estimate, CI = confidence interval.

Table 6.3: Top results ($p < 8 \times 10^{-6}$ for total ADHD symptoms) for the associations between CpG sites related to exposure to maternal smoking during pregnancy and offpsring hyperactivity symptoms.

CpG	Chr	Pos	Gene		*N*	B (95% CI)	*p*-value
cg05575921	5	373378	*AHRR*	All individuals	432	-0.049 (-0.074; -0.024)	0.00015
				Never-smokers	321	-0.043 (-0.088; 0.001)	
cg26703534	5	377358	*AHRR*	All individuals	432	-0.069 (-0.105; -0.034)	0.00016
				Never-smokers	321	-0.046 (-0.094; 0.001)	

Chr = chromosome; Pos = position; B = effect size estimate, CI = confidence interval.

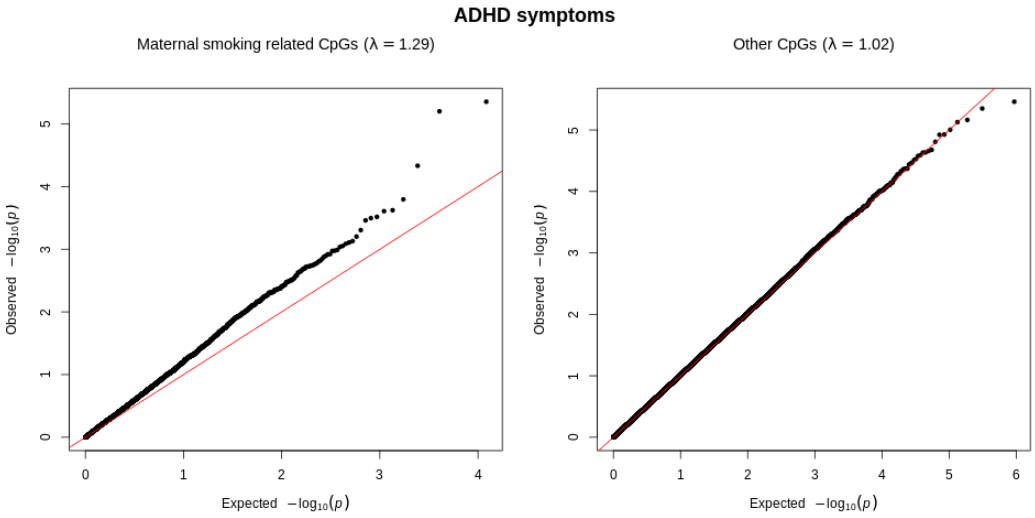

Figure 6.4: QQ-plots for the association analyses of blood DNA methylation and ADHD symptoms for CpGs associated with exposure to maternal smoking (left, 6,073 CpGs) and other CpGs (right, 467,201 CpGs).

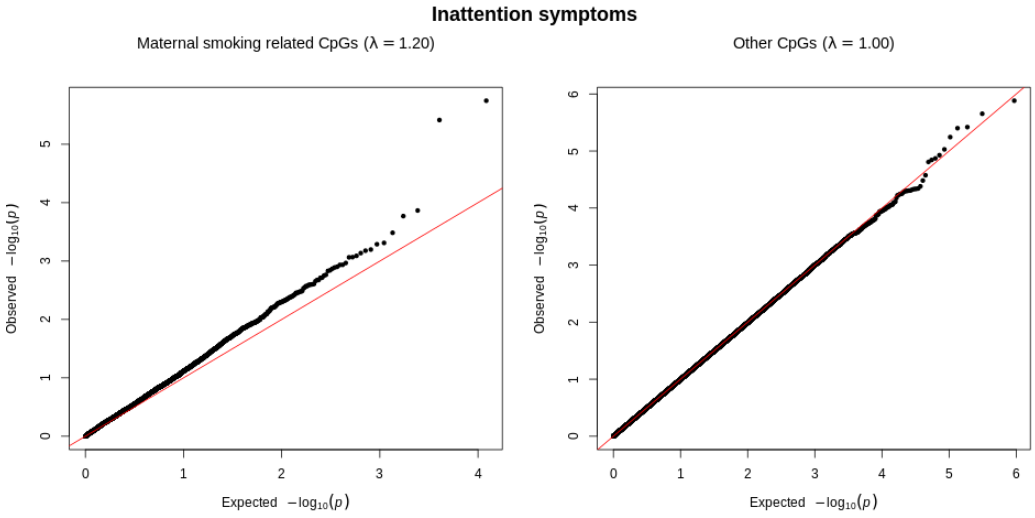

Figure 6.5: QQ-plots for the association analyses of blood DNA methylation and inattention symptoms for CpGs associated with exposure to maternal smoking (left, 6,073 CpGs) and other CpGs (right, 467,201 CpGs).

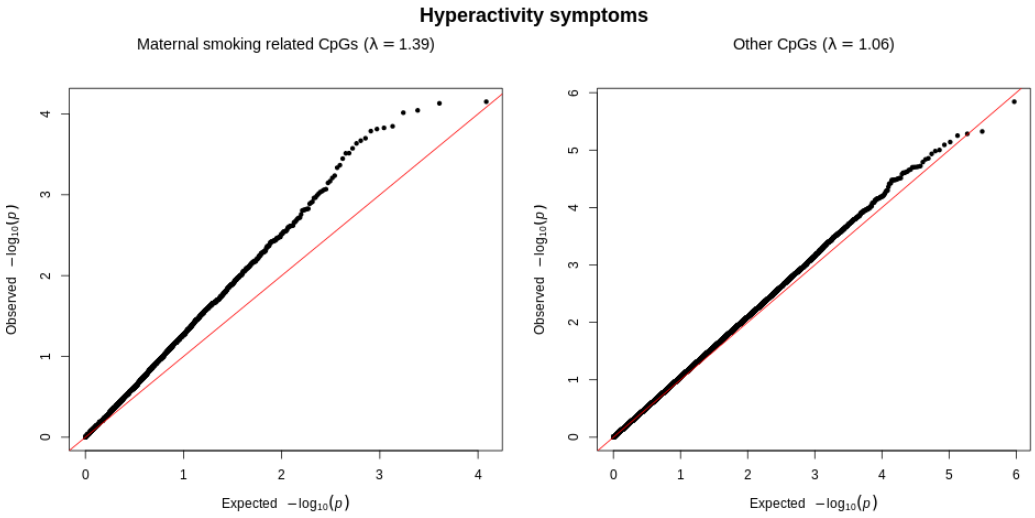

Figure 6.6: QQ-plots for the association analyses of blood DNA methylation and hyperactivity symptoms for CpGs associated with exposure to maternal smoking (left, 6,073 CpGs) and other CpGs (right, 467,201 CpGs).

6.2.2 DNA methylation based prediction of exposure to maternal smoking

Model performance was similar across the different penalised regression models in the test dataset based on 1,000 repetitions of data splitting (Table 6.4). The lowest mean Brier score and highest mean C index was detected for $\alpha = 0.2$ for the model with 6,073 CpG sites and for $\alpha = 0$ for the model with 177 CpG sites. These values were used for the final tuning parameter λ estimation using the full NFBC1966 dataset, where the final $\lambda = 0.084$ for the first risk score (6,073 CpGs), and $\lambda = 0.100$ for the second (177 CpGs). The resulting model was then used in NFBC1986 dataset to generate the predicted probabilites for exposure to maternal smoking from offspring blood DNA methylation.

Table 6.4: Mean (standard deviation) for Brier score and C index in the test set based on 1,000 repeated data splits for both risk scores using different shrinkage parameters in penalised regression models.

	Risk score I: 6,073 CpGs		Risk score II: 177 CpGs	
α	Brier score	C index	Brier score	C index
0	0.105 (0.019)	0.753 (0.055)	0.087 (0.016)	0.856 (0.044)
0.2	0.099 (0.017)	0.801 (0.051)	0.087 (0.016)	0.851 (0.045)
0.5	0.099 (0.017)	0.797 (0.051)	0.088 (0.016)	0.841 (0.045)
0.8	0.099 (0.017)	0.795 (0.051)	0.089 (0.016)	0.836 (0.046)
1	0.099 (0.017)	0.794 (0.052)	0.090 (0.016)	0.833 (0.046)

The DNA methylation risk scores correlated with each other (Pearson correlation $r = 0.87$) and were predictive of the targeted exposure (C index values 0.84 and 0.85 for the risk scores with 6,073 and 177 CpG sites, respectively). The risk scores were then tested for association for exposure to both maternal and paternal smoking with mutual adjustment. There was clear evidence for association for exposure to maternal smoking, but not paternal smoking, for both risk scores. Similar results were obtained within offspring never-smokers (Figure 6.7).

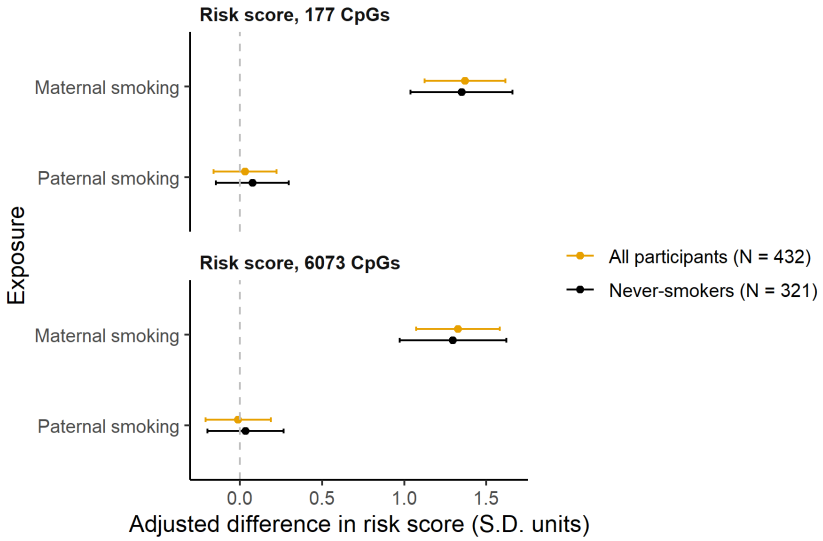

Figure 6.7: The differences (point estimates and their 95% confidence intervals, adjusted for family SES, maternal age and offspring sex) in offspring DNA methylation risk scores between exposed and non-exposed, separately within all participants and offspring never-smokers. S.D. = standard deviation.

6.2.3 Mediation analysis

I then assessed the evidence for potential mediation of the effect of intrauterine smoke exposure on offspring ADHD symptoms, via blood DNA methylation. There was evidence for mediation by CpG site cg05575921 in *AHRR* gene, with an estimated average causal mediated effect (in standard deviation changes in ADHD symptoms per exposure to maternal smoking) of 0.12 (95% CI [0.03; 0.22], $p = 0.004$, Figure 6.8), with a point estimate for the proportion of the mediated effect = 38%. Similar effect sizes were detected for the DNA methylation risk scores (risk score with 6,073 CpGs: 0.12, 95% CI [-0.05; 0.23]; risk score with 177 CpGs: 0.12, 95% CI [-0.02; 0.26]), as well as for separate inattention and hyperactivity symptoms.

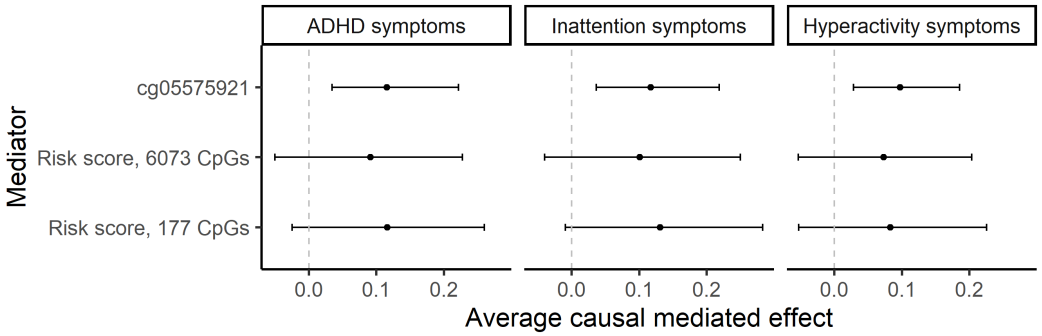

Figure 6.8: Point estimates and their 95% confidence intervals for the average causal mediated effects of exposure to maternal smoking during pregnancy on offspring ADHD symptoms, mediated via offspring blood DNA methylation.

6.2.4 MR results

For MR, there were no genetic variants available in ARIES database to serve as instruments for cg05575921, that could be found from the ADHD GWAS summary statistics. Based on the GWAS meta-analysis results from the NFBC cohorts, the I used variant rs2672766 within *AHRR* gene as an instrumental variable. The F-statistic for the instrument was 37, indicating a sufficiently strong instrument not to suffer from weak instrument bias. The results from MR analysis did not give evidence for association between genetically determined cg05575921 levels and the risk of ADHD (effect size estimate in SD units = 0.09, 95% CI [-0.03; 0.21], $p = 0.19$).

6.3 Discussion

The aim in this chapter was to examine the possible pathway between intrauterine smoke exposure and offspring ADHD symptoms, via differential offspring DNA methylation. I conducted association analysis between DNA methylation and ADHD symptoms, concentrating on the CpG sites with evidence for association with exposure to maternal smoking during pregnancy from previous literature (Joubert et al., 2016). Specifically, two CpG sites within *AHRR* gene showed the strongest associations. There was evidence of inflation in the association results, implying more associations with small effect sizes that would be expected by chance. Even though residual confounding cannot completely be ruled out, the inflation was not replicated among other CpG sites, implying that this is specific to the

maternal smoking related CpGs.

According to earlier literature, exposure to maternal smoking seem to affect offspring DNA methylation on numerous CpG sites (Joubert et al., 2016; Wiklund et al., 2019). Here, I aimed to capture the full profile of this effect by constructing a DNA methylation risk score on exposure to maternal smoking. Two alternative sets of CpGs were used – the other using 6,073 CpGs from Joubert et al. (2016), and the other with a selected set of 177 CpGs not associated with exposure to paternal smoking or driven by offspring smoking status. Both risk scores were predictive of the targeted exposure, and not associated with paternal smoking or driven by offspring's own smoking status.

The mediation analysis showed positive effects for the estimated average causal mediated effect. There was evidence for differential DNA methylation at cg05575921 mediating the effect of exposure to maternal smoking on offspring ADHD symptoms. The effect sizes were similar for both risk scores, however the condfidence intervals overlapped with zero. Finally, MR analysis provided no evidence for causality of cg05575921 on ADHD.

These results do not provide strong evidence for DNA methylation mediating the observational association between exposure to maternal smoking during pregnancy and offspring ADHD symptoms. However, concordant positive point estimates suggest that this hypothesis should be researched further in different study populations.

I developed a DNA methylation risk score that was predictive of intrauterine smoke exposure. As the DNA methylation profile seem to be altered by exposure to maternal smoking on a large scale with small effects, accumulating small individual effects into a risk score is likely to be a productive approach, as shown here. There have been earlier attempts in building a similar risk score (Reese et al., 2017; Richmond et al., 2018). The risk scores created in this work have two distinctive properties. First, the risk scores were independent of offspring's own smoking status. Second, the risk scores were specific to exposure to maternal smoking, and were not associated with paternal smoking. This provides further evidence for the plausibility and specificity of the created risk scores. The risk score with 177 CpGs, selected not to be driven by offspring's own smoking and not associated with paternal smoking in the training set, was marginally more predictive of the exposure, however the results for mediation analysis were similar.

Paternal smoking is often used as a negative control exposure when examining the effect

of intrauterine exposure to offspring health outcomes (Davey Smith et al., 2012). In the absence of an intrauterine effect, one would expect similar effect sizes with maternal and paternal smoking. Negative control exposure analysis can be an effective way to distinguish potentially causal effects from confounded effects. Although unmeasured confounding cannot be definitely ruled out, the use of DNA methylation specific to maternal smoking partially circumvents this problem and gives an opportunity to examine the potential causal pathway from intrauterine smoke exposure to offspring ADHD, via offspring DNA methylation (Figure 6.1).

There is known genetic correlation between smoking and ADHD (Demontis et al., 2019). The detected associations may be driven by genetic confounding, which causes both differential offspring DNA methylation and ADHD symptoms. Unravelling the true causal associations is further complicated by the evidence of liability to ADHD increasing the likelihood of smoking (Treur et al., 2019). Differential DNA methylation at cg05575921 is well known to be affected by individual's own smoking status (Bojesen et al., 2017), so distinguishing the effects of own smoking and maternal smoking is difficult. For sensitivity, I tested the association analysis within offspring never-smokers, however it should be noted that conditioning on the smoking status might induce a spurious association due to potential collider bias (Munafò et al., 2017).

Another potential source of complexity in the relationship between smoking and ADHD is the self-medication hypothesis (Potter et al., 2006; Wilens et al., 2007): it is suggested that nicotine acts as an indirect dopamine agonist to alleviate ADHD symptoms (De Biasi & Dani, 2011; van Amsterdam et al., 2018). Therefore, those expecting mothers who have a higher susceptibility to ADHD symptoms may smoke as a form of self-medication. However, the current evidence for the self-medicating properties of smoking on ameliorating ADHD or other neuropsychiatric symptoms is not robust (Sousa et al., 2011; Vermeulen et al., 2019), and it is hypothesised that the perceived relief in ADHD symptoms may simply be the relief of tobacco withdrawal symptoms (G. M. J. Taylor & Munafò, 2019).

Some key limitations should be taken into consideration when interpreting these results. Sample size is a key limiting factor, and replication of the analysis in other datasets is required. Another issue is the tissue specificity of DNA methylation. The current analysis used blood DNA methylation, however the most likely relevant tissue for neuropsychiatric outcomes, such as ADHD, is brain. Nevertheless, early exposures – such as intrauterine

smoke exposure – might manifest themselves in multiple tissues (Aberg et al., 2013). In addition, there is evidence for correlation between methylation QTL estimates between blood and brain tissues (Qi et al., 2018). Finally, there is no maternal genotype information available in this dataset to adjust for potential genetic confounding.

In summary, I built two prediction models on exposure to maternal smoking based on offspring DNA methylation that were predictive of the exposure, independent of paternal smoke exposure and not driven by offspring's own smoking. I found no sufficient evidence for DNA methylation mediating the association between exposure to maternal smoking during pregnancy and ADHD. Similar approach in other datasets and triangulation of other evidence is needed to disentangle the role of maternal smoking-induced differential DNA methylation in the aetiology in ADHD.

Chapter 7

Multi-omics variable selection and prediction of ADHD symptoms

Recent GWAS (Demontis et al., 2019) and EWAS (Mooney et al., 2020; Rovira et al., 2020; van Dongen et al., 2019) on ADHD have given new insights to the aetiology of ADHD. However, these studies on their own are not able to detect complex interplay across omics datasets that mirror different levels of biological regulation. Datasets with multiple omics measured in the same individuals offer an intriguing platform to apply integrative analytical methods (Section 4.2.2). The results from such integrative analyses, even with modest sample sizes, may give insightful results that can be prioritised in further analyses (Hasin et al., 2017). In addition, joint modelling of ADHD symptom subtypes, inattention and hyperactivity, as a multivariate outcome may increase power in omics-wide association analysis (Mägi et al., 2017; Teixeira-Pinto et al., 2009).

The aim in this chapter was to further examine the molecular constitution of ADHD. I conducted omics-wide association analyses for ADHD symptoms by joint modelling of ADHD symptom subtypes, inattention and hyperactivity, as a multivariate outcome.

In the two previous chapters, differential DNA methylation related to maternal smoking during pregnancy and its potential mediating effect on offspring ADHD and other adverse

health outcomes was evaluated in detail. In this chapter, I add information from both genomic data and from metabolic measures quantified by NMR spectroscopy (Soininen et al., 2015). The integration of these molecular datasets may reveal more of the molecular composition of ADHD, especially related to the putative impact due to intrauterine smoke exposure.

Furthermore, using two QTL databases, GTEx for expression QTL (eQTL) (GTEx Consortium, 2017) and ARIES for methylation QTL (mQTL) (Gaunt et al., 2016; Relton et al., 2015), multi-omics variables were selected based on their relationships with differential DNA methylation related to intrauterine smoke exposure. To evaluate whether ADHD symptoms could be reliably predicted from these maternal smoking related multi-omics variables, regression models were applied for prediction of ADHD symptoms. Additionally, I performed multivariable selection via phenotype-specific network modelling based on canonical correlation analysis.

7.1 Methods

Data from NFBC1986 collected at 16 years were used. For association analysis, the two symptom types of ADHD, inattention and hyperactivity/impulsivity as measured by SWAN scale (Swanson et al., 2012), were treated as a bivariate continuous outcome.

7.1.1 Omics-wide association analyses

I conducted a GWAS, an EWAS and an association analysis across NMR-quantified metabolic measures for the multivariate outcome of inattention and hyperactivity symptoms. Table 7.1 gives information on these association analyses. The explanatory variables used in association analysis were sex and first 20 genetic PCs for GWAS; maternal smoking and age, family SES, offspring sex, five first genetic PCs, technical covariates of CPACOR pipeline (Lehne et al., 2015) and white blood cell subpopulation estimates for EWAS; and sex and family SES for metabolic measures. The phenotypes were first separately regressed on these explanatory variables, and the remaining residuals were used in a reverse regression framework (Section 4.2.1.1) in the omics-wide association studies.

As hormonal contraception is known to alter individual's metabolic profile (Q. Wang et al.,

2016), those study participants who reported current (regular or irregular) use of oral contraceptives were excluded from the analysis of metabolic measures. Additional adjustment for BMI was used only for sensitivity analysis, as the directionality between ADHD and BMI is unclear (Cortese & Tessari, 2017).

Table 7.1: Adjustments and sample sizes for omics-wide association analysis.

Omics data	Adjustments	Variables	*N*
GWAS	Sex, 20 genetic PCs	11,009,294	3,185
EWAS	Maternal smoking and age, family SES, offspring sex, five genetic PCs, technical covariates, white blood cell subpopulation estimates	466,290	432
NMR metabolic measures	Sex, family SES*	228	4,713†

N: sample size; GWAS: genome-wide association study; EWAS: epigenome-wide association study; NMR: nuclear magnetic resonance; PC: principal component; SES: socioeconomic status.
* Adjustment for BMI as a sensitivity analysis.
† Oral contraceptive users excluded.

To apply the reverse regression framework here, the residualised phenotypes $\mathbf{Y} = \left(\mathbf{Y}^{\text{resid.}}_{\text{Inattention}} \; \mathbf{Y}^{\text{resid.}}_{\text{Hyperactivity}}\right)^T$ are treated as explanatory variables, and the omics variables $\mathbf{x}_j, j = 1, \ldots, p_m, m = 1, 2, 3$ as the outcome:

$$\mathbf{x}_j = \mathbf{Y}\boldsymbol{\beta}.$$

The evaluation for an association can be done via testing $\mathrm{H}_0 : (\beta_1 \; \beta_2) = (0 \; 0)$. For GWAS and EWAS, I used existing SCOPA and MethylSCOPA softwares, respectively (Mägi et al., 2017; Draisma et al., 2019).

To correct for multiple testing, statistical significance was set based on the conventional thresholds of $p < 5 \times 10^{-8}$ for GWAS (Dudbridge & Gusnanto, 2008) and $p < 10^{-7}$ for EWAS (Lehne et al., 2015). The metabolic measures data are highly correlated, and therefore, for multiple testing correction in the analysis of metabolic measures, PCA was first applied to the data to approximate the number of independent tests. PCA showed that the first 28 PCs explained 95% of the variation. Based on this, the multiple testing threshold was set at $0.05/28 \approx 0.0018$.

7.1.2 Multi-omics variables related to CpGs associated with maternal smoking

To further examine the potential impact of specifically maternal smoking-related omics, prediction modelling techniques were applied by using a pre-selected set of multi-omics variables. The selection was done using information from ARIES database, GTEx database, and NFBC1966. The starting point was the list of 6,073 CpGs that have shown robust associations in cord blood with exposure to maternal smoking during pregnancy (Joubert et al., 2016). Of this list, 954 GpGs measured in later life associated at $p < 0.05$ with exposure to maternal smoking with a concordant sign of the effect size estimate as in Joubert et al. (2016) (Chapter 6).

For these CpG sites, mQTLs were sought from ARIES database (Gaunt et al., 2016). This database contains information on SNPs that are associated with differential blood DNA methylation ($p < 1 \times 10^{-7}$) at five different time points: pregnancy, birth, childhood, adolescence, and middle-age. The adolescence time point (15 years) was used here as this is the most relevant to the timing of the outcome measurement in the phenotypic data (16 years).

The SNPs (and the corresponding CpG sites) that also associated with gene expression (eQTL) at $p < 1 \times 10^{-4}$ across tissues in GTEx database (GTEx Consortium, 2017) were retained. As these genetic variants are associated with both DNA methylation and gene expression, it is assumed that the differential DNA methylation has functional relevance to the expression (Figure 7.1). As the DNA methylation was measured in blood and the most likely relevant tissue in ADHD is brain, the meta-analysed (rather than tissue-specific) association p-value was used.

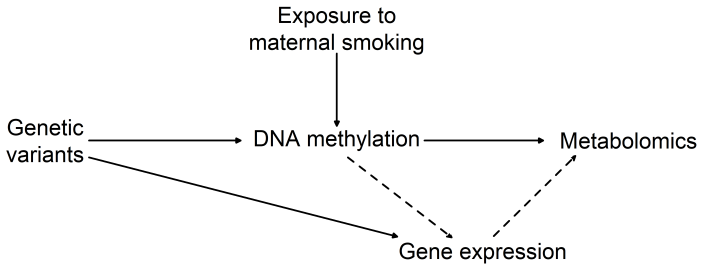

Figure 7.1: Schematic illustration of the associations between omics. Solid lines represent associations for omics variable inclusion criteria. Dashed lines represent hypothesised associations in the present study.

After removing SNPs with MAF < 0.05 and clumping for $r^2 < 0.1$ with a window of 250kb, there were 534 SNPs that were both mQTL for 356 CpG sites and eQTL for gene expression.

To obtain metabolic measures associated with DNA methylation, I conducted association analysis for the 356 CpG sites with 228 metabolic measures measured by NMR metabolomics platform (Soininen et al., 2015) in NFBC1966 dataset ($N = 692$). Metabolite values were first regressed on sex, and an inverse normal rank transformation was applied to the residuals. These residuals were then regressed on DNA methylation beta value at each CpG site. The first 30 PCs of technical control probes and white blood cell subpopulation estimates were included as additional explanatory variables to account for technical variation and cell type heterogeneity.

To apply an appropriate multiple testing correction, I first conducted PCA as a dimension reduction to the highly correlated NMR metabolomics data. PCA showed that 28 PCs explained 95% of the variation. Based on this, the multiple testing threshold was set at $0.05/28 \approx 0.0018$. There were 132 metabolic measures that associated with any of the 356 CpGs at p-value < 0.0018, and these variables were taken forward for the subsequent modelling.

Table 7.2 summarises the criteria for omics variables for the prediction and network modelling. The outcome variable used was the total ADHD symptoms at 16 years evaluated by

offspring's parents using SWAN questionnaire (Swanson et al., 2012). All omics variables were inverse-normal rank transformed to normalise them to a common scale. Full data for the outcome and omics variables were available for 432 participants.

Table 7.2: Criteria for omics variables selected for prediction of ADHD symptoms and network modelling.

Omics data	Selection criteria	Variables
Genomics	mQTL at $p < 10^{-7*}$ eQTL at $p < 10^{-4}$	534
Epigenomics	Associated with maternal smoking in adolescents and adults mQTL available	356
Metabolic measures	Associated with DNA methylation at $p < 0.0018$†	132

mQTL = methylation quantitative trait loci; eQTL = expression quantitative trait loci
* Availability in ARIES database.
† Bonferroni correction for principal components in metabolomic data explaining 95% of the total variance.

7.1.2.1 Prediction models

Regression models were used for prediction modelling, and the minimal model included only sex as a predictor. The additional predictors considered were the 534 SNPs, 356 CpG sites and 132 metabolites. I compared the performance of the following high-dimensional regression methods:

1. Elastic net regression (i.e. penalised regression with penalty term as in equation (4.10)), with α values 0 (Ridge), $0.2, 0.5, 0.8$ and 1 (Lasso).

2. Relaxed sparse multiple canonical correlation analysis (rsmCCA) regression: sparse CCA (Section 4.2.2.2) was conducted between genetic, epigenetic and metabolic measures datasets, and the first 1 to 5 canonical variate lists (Section 4.2.2.2) were included as predictors.

3. Principal component regression (PCR): PCA is first conducted for the explanatory variables, and the first one, five or ten PCs are used as predictors in the regression. The PCA was conducted for genetic, epigenetic and metabolic measures datasets, either separately (sepPCR) or for a combined dataset (ComPCR).

Sex was added to all of these models as an unpenalised predictor. To evaluate the predictive

performance of the models, I applied repeated data splitting: the dataset was split into training and test sets with probabilities 0.8 and 0.2, respectively. The training set was used to estimate model parameters. The tuning parameters for elastic net or rsmCCA were estimated using 10-fold CV. Model performance was calculated in the test set, and the measures used were root mean squared error (RMSE) and variance explained (R^2) (Section 4.1.1.1). The data splitting procedure was repeated 100 times.

7.1.2.2 Phenotype-specific networks

To conduct variable selection across multiple omics datasets, I applied an adapted version of SmCCNet (Sparse multiple Canonical Correlation Network analysis) method proposed by Shi et al. (2019) to discover multi-omics networks specific to the phenotype. The SmCCNet method is based on sparse CCA with lasso penalty. In the adaptation applied here, the originally proposed sparse multiple CCA method (Witten et al., 2009) is replaced by a relaxed sparse multiple CCA with lasso penalty, originally proposed by Suo et al. (2017) and recently implemented by Rodosthenous et al. (2020).

As detailed in Section 4.2.2.2, for normalised datasets $\mathbf{X}_m \in \mathbb{R}^{n\times p_m}, m = 1, \dots, M$, the relaxed sparse multiple CCA with lasso penalty corresponds to finding *canonical weights*, i.e. vectors $\mathbf{w}_1, \dots, \mathbf{w}_M$ that minimise equation (4.21), repeated here for clarity:

$$\min_{\mathbf{w}_m} -\sum_{q<r} \text{cor}(\mathbf{X}_q\mathbf{w}_q, \mathbf{X}_r\mathbf{w}_r) + \sum_{m=1}^{M} \tau_m ||\mathbf{w}_m||$$
$$\text{subject to } \text{Var}(\mathbf{X}_m\mathbf{w}_m) \leq 1,\ m = 1, \dots, M.$$

The phenotype $\mathbf{Y} \in \mathbb{R}^{n\times 1}$ is taken into account by treating it as an additional unpenalised data type. Therefore, in the current data, $M = 4$, and the number of variables p_m in each dataset $m = 1, 2, 3$ were $p_1 = 534$, $p_2 = 356$, $p_3 = 132$ and $p_4 = 1$ (with $\tau_4 = 0$).

Selection of the regularisation parameters The penalty parameters $\tau_m, m = 1, 2, 3$ were selected based on 5-fold CV. A subsampling scheme was added to improve the robustness of the networks (Y. R. Wang et al., 2015). Within each fold and for each candidate penalty vector, pseudo canonical weights were estimated based on 1,000 repetitions of subsampling variables with subsampling proportion of 0.8 for each omics data type. The relaxed sparse multiple CCA was applied for each subsample, and the pseudo canonical weights were

defined as the mean of the absolute canonical weights of each subsample. The prediction error was calculated in the test set using the pseudo canonical weights. The mean of the errors within each fold were calculated for each candidate penalty vector.

Robust canonical weights and scaled similarity matrix Using the penalties corresponding to the smallest prediction error as per above, I applied the sparse CCA method on 1,000 subsamples, with subsampling proportion = 0.8 for each omics data type. As above, the subsampling was done to increase the robustness of the network construction (Y. R. Wang et al., 2015). A relationship matrix A is calculated for each subsample as $A_{ij} = |u_i \times u_j|, \forall i, j \in \{1, \ldots, p_1 + p_2 + p_3\}$. The mean of the matrices is taken over all subsamples, the diagonal is set to zero, and the matrix is then normalised to have maximum relatedness of 1 – this matrix is denoted as $\bar{A}$, which is the scaled similarity matrix.

Hierarchical tree cut for phenotype-specific network In order to focus only on the strongest network associations, a hierarchical tree is constructed based on the dissimilarities $1 - \bar{A}$, and the tree is cut using a liberal threshold (≈ 1). The remaining clusters are trimmed by removing singletons (as these are not networks) and edges with low weights (here $d < 0.5$), to focus on stronger network connections. The remaining clusters can be taken as multi-omics networks related to the phenotype.

I further visualised the relationships among the final variables in the phenotype-specific multi-omics networks by a Pearson correlation heatmap of the variables. For the SNPs and CpGs selected for the network, I also searched Kyoto Encyclopedia of Genes and Genomes (KEGG) database for enriched genes.

7.2 Results

7.2.1 Omics-wide association analyses

Inattention and hyperactivity symptom scores were strongly correlated with each other (Pearson correlation coefficient = 0.74, Figure 7.2). For GWAS and EWAS, there were no SNPs or CpGs with $p < 5 \times 10^{-8}$ or $p < 1 \times 10^{-7}$, respectively (Figure 7.3).

In the association analysis for metabolic measures there were nine measures with $p < 0.0018$

(Table 7.3). These included albumin, proportion of monounsaturated fatty acids out of total fatty acids, different concentrations of triglycerides in medium and small low-density lipoproteins (LDL), proportions of phospholipids and triglycerides in different particle size groups, mean diameter for LDL particles, and creatinine. Out of the top hits, seven remained at $p < 0.0018$ after additional adjustment for BMI. The associations were positive for albumin, creatinine and mean LDL particle diameter, and negative for other top hits.

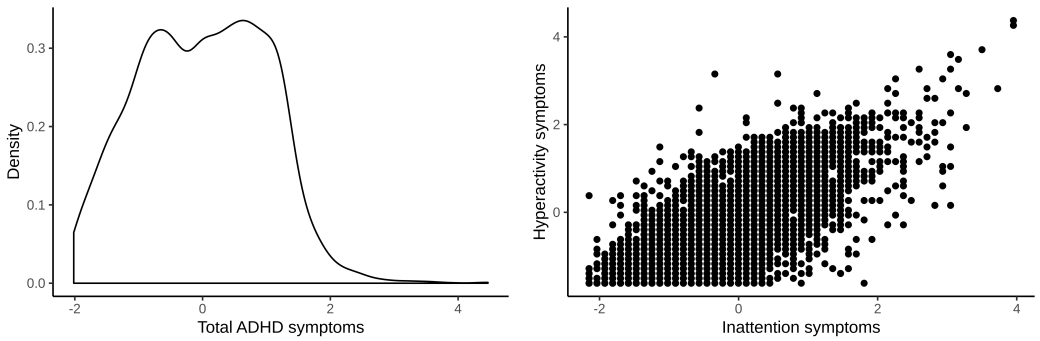

Figure 7.2: Phenotype distributions in NFBC1986 with NMR metabolic measures available, $N = 4{,}713$. Left panel: density plot for total ADHD symptoms; right panel: scatterplot of inattention and hyperactivity symptom scores.

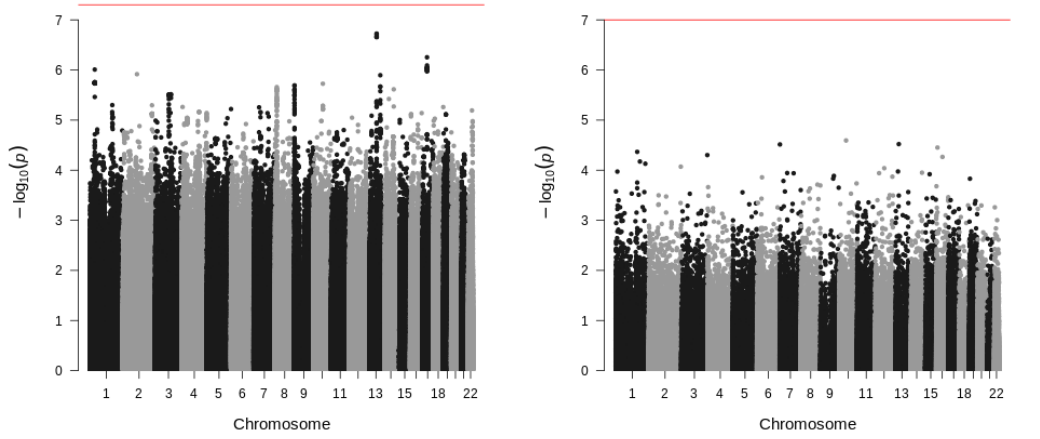

Figure 7.3: Manhattan plots for omics-wide association studies. Left panel: GWAS; right panel: EWAS. The red lines present the significance thresholds ($p = 5 \times 10^{-8}$ for GWAS, $p = 10^{-7}$ for EWAS).

Table 7.3: NMR-quantified metabolic measures with $p < 0.0018$ in association analysis with ADHD symptoms, both without and with adjustment for BMI.

Metabolic measure	Direction	p-value	$p_{\text{adj. BMI}}$
Albumin	−	2.7e-05	2.1e-05
Ratio of monounsaturated fatty acids to total fatty acids	+	3.1e-05	2.6e-04
Triglycerides in small LDL	+	1.8e-04	5.3e-04
Mean diameter for LDL particles	−	2.2e-04	3.1e-04
Triglycerides to total lipids ratio in small LDL	+	2.4e-04	3.4e-04
Creatinine	−	5.2e-04	4.4e-04
Phospholipids to total lipds ratio in large HDL	+	6.0e-04	7.3e-03
Phospholipids to total lipds ratio in small VLDL	+	9.3e-04	4.0e-04
Triglycerides in medium LDL	+	1.1e-03	1.8e-03

LDL: low-density lipoprotein; HDL: high-density lipoprotein; VLDL: very low-density lipoprotein.

7.2.2 Prediction models

The different methods to predict ADHD symptoms were compared by RMSE and R^2 in the test set, with 100 repetitions (Figures 7.4 and 7.5). Similar performance was detected across models and outcomes.

The minimal model with sex as the only predictor showed weak predictive performance, with mean $R^2 = 0.02$ in test set. None of the methods showed superiority to a minimal model with sex as the only predictor. The best average predictive performance was detected for PCR with separate PCA for each omics dataset and using five PCs.

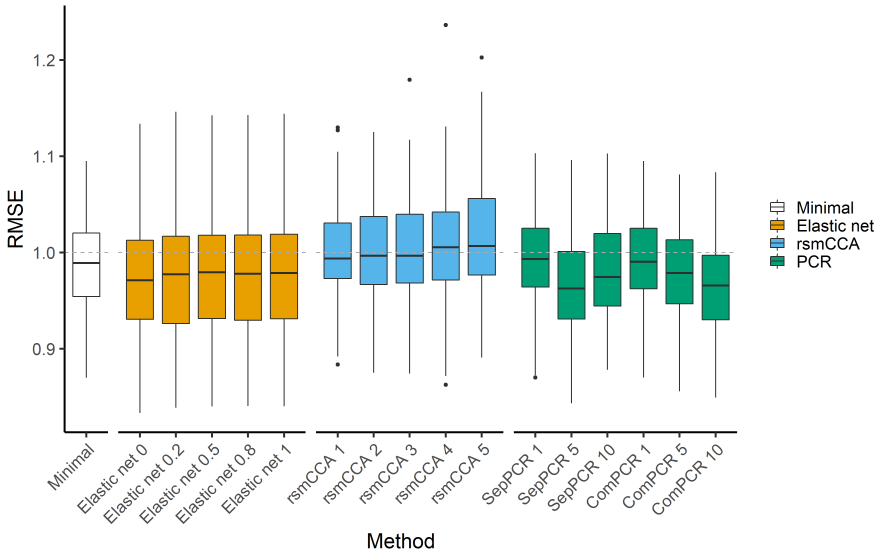

Figure 7.4: Boxplots for root mean squared error (RMSE) in the test set based on 100 repeated data splits with total ADHD symptoms as outcome. rsmCCA = relaxed sparse multiple canonical correlation analysis; PCR = principal component regression; SepPCR = PCR with principal components (PCs) calculated separately for genetic, epigenetic and NMR-quantified metabolic data; ComPCR = PCR with PCs calculated for combined genetic, epigenetic and NMR-quantified metabolic data.

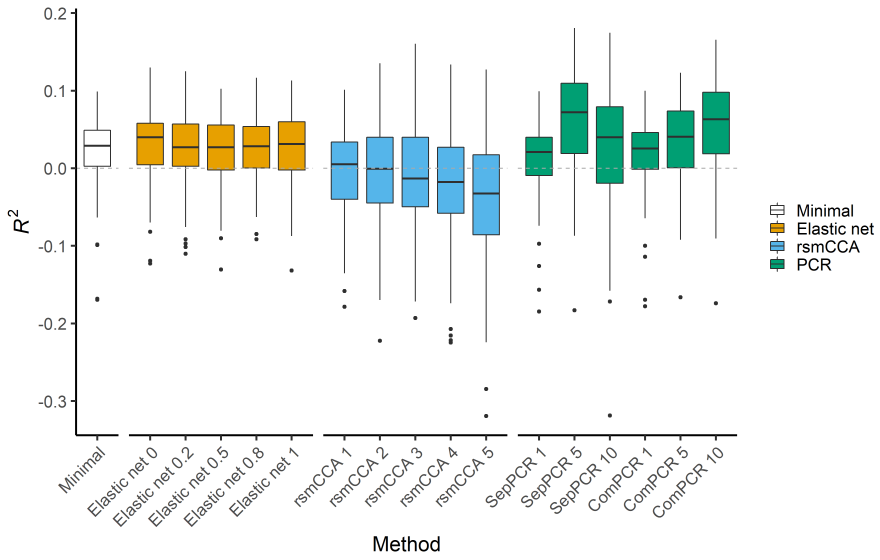

Figure 7.5: Boxplots for variance explained (R^2) in the test set based on 100 repeated data splits with total ADHD symptoms as outcome. rsmCCA = relaxed sparse multiple canonical correlation analysis; PCR = principal component regression; SepPCR = PCR with principal components (PCs) calculated separately for genetic, epigenetic and NMR-quantified metabolic data; ComPCR = PCR with PCs calculated for combined genetic, epigenetic and NMR-quantified metabolic data.

7.2.3 Multi-omics networks

For multivariate variable selection network model, the optimal sparsity penalties were $\tau_1 = 1$, $\tau_2 = 6$ and $\tau_3 = 1$ corresponding to penalisations for DNA methylation, genetic, and NMR metabolomics datasets, respectively (Figure 7.6).

Based on these tuning parameters and after applying 1,000 repetitions of subsampling, the similarity matrix was generated for the omics datasets. I further applied a hierarchical tree cut for the dissimilarities, after which there were two multi-omics networks remaining that were associated with the phenotype. One of them survived the weight cut-off $d < 0.5$, and this network included seven SNPs, one CpG site and nine metabolites (Figures 7.7 and 7.8).

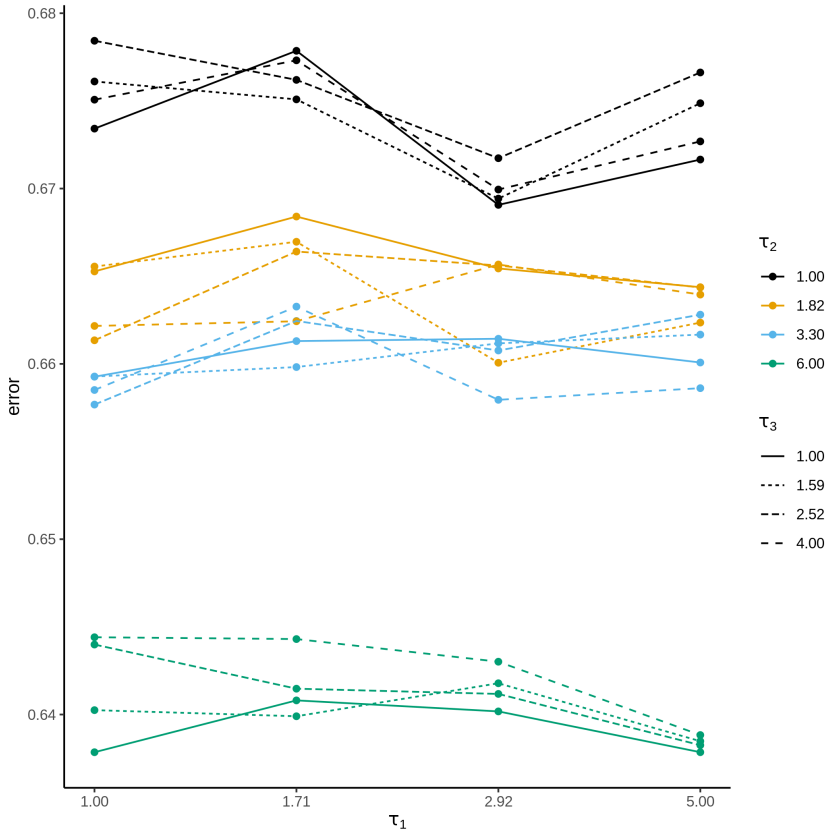

Figure 7.6: Cross-validation error for selecting sparsity penalties for SmCCNet.

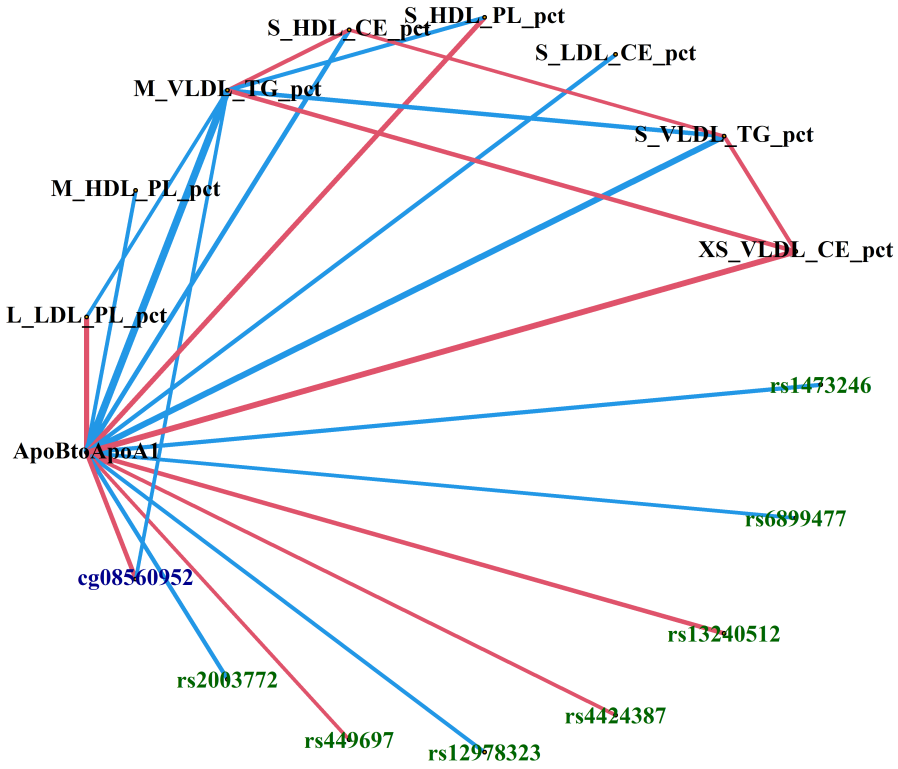

Figure 7.7: Multi-omics network related to ADHD symptoms based on the phenotype-specific network model. The nodes are the omics variables that were included in the network. Blue and red lines represent positive and negative correlations, respectively, between the variables, with line thickness representing the absolute value of the correlations.

The CpG included was cg08560952, located in an intergenic region in chromosome 16. The SNPs and their nearest genes were rs14743246 (in *SEC61A1* gene), rs6899477 (*LOC105378157*), rs13240512 (*POT1*), rs4424387 (*LOC158434*), rs12978323 (*AXL*), rs449697 (*ZNF544*) and rs2003772 (*LOC105369301*). No enriched genes were found in a look-up to KEGG database for over-representation gene analysis. The metabolites included the ratio of apolipoprotein B to apolipoprotein A1 (ApoBtoApoA1), and different cholesterol particle size proportions. ApoBtoApoA1 was found to be a central node in the network with connections to all other selected variables (Figure 7.7).

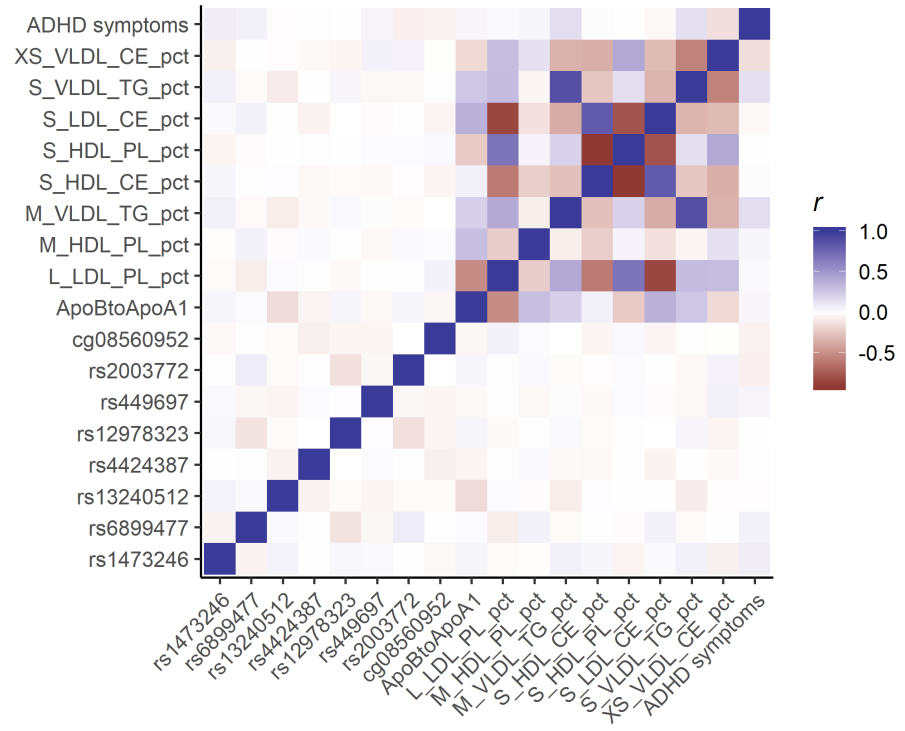

Figure 7.8: Correlation heatmap for ADHD symptoms and the related variables in the phenotype-specific multi-omics network model.

7.3 Discussion

To gain more insight of the molecular makeup of ADHD, I conducted association analysis across genomic, epigenomic and NMR-quantified metabolic datasets. From GWAS and EWAS analysis, there was no evidence for associations with ADHD symptoms.

Nine metabolic measures were found to be associated with ADHD symptoms. However, comorbidity of BMI and ADHD is well-known (Cortese & Tessari, 2017), and the metabolic profile is affected by adiposity (Würtz et al., 2014). Therefore, it is possible that BMI is driving the observed associations between ADHD and the metabolic profile. Sensitivity analysis was conducted with additional adjustment for BMI, and the results showed associations of ADHD symptoms and seven metabolic measures, independently of BMI. ADHD has both observational and genetic correlation with high-density lipoprotein cholesterol (Demontis et al., 2019). These results provide additional insight to the observational associations between different lipid particle sizes and ADHD symptoms.

In addition to different lipids, the NMR panel used here also quantifies fatty acids and related measures. Dietary supplementation of polyunsaturated fatty acids – such as omega-3 fatty acids – has emerged as a popular method for a nonpharmacologic intervention of ADHD, due to their importance in neural processes (J. P.-C. Chang et al., 2018; Sonuga-Barke et al., 2013). However, the benefits of the supplementations remain unclear (Banaschewski et al., 2018; Cooper et al., 2015). Here, the only fatty acid related measure showing association with ADHD symptoms was the ratio of monounsaturated fatty acids to total fatty acids.

For a further disentanglement of the role of maternal smoking-induced molecular perturbations in relation to ADHD, several high-dimensional prediction modelling methods were applied to a multi-omics dataset in predicting ADHD symptoms. There was no evidence for improved prediction over a minimal model, which itself was little better than predictions obtained by chance alone. These results demonstrate both the importance of large sample size, and the difficulty of prediction modelling across multi-omics datasets.

As an alternative strategy, I conducted variable selection across the multi-omics dataset by applying a modification to a previously suggested method to detect phenotype-specific multi-omics networks. The results provided suggestive associations for maternal smoking-related QTL, cholesterol particle size and apolipoprotein ratios with ADHD. While gene enrichment analysis did not show any enriched genes, the ratio of apolipoproteins B and A1 was found as a key node in the detected network.

While omics-wide association studies aim to find only univariate omics variables associated with the outcome, the network method applied here aims to detect cross-omics associations specific to the phenotype. The availability of multiple omics datasets from the same individuals allows for a more refined analysis, with a potential to reveal complex interplays across omics that would not be detectable in sinlge omics analysis.

Lipoproteins are involved in the transportation of cholesterol and triglycerides in the blood stream. ApoBtoApoA1 reflects cholesterol transportation balance and is linked to metabolic disorders (Jing et al., 2014; Jung et al., 2012) and cardiovascular events (McQueen et al., 2008; O'Donnell et al., 2016). This may suggest further investigation of the role of ApoBtoApoA1 in the associations of exposure to maternal smoking and ADHD with obesity and cardiovascular diseases (Cortese, 2019; Penninx & Lange, 2018; Power et al., 2010). These findings point out the potential in integrative analysis strategies for multi-omics data.

A key limitation for all models investigated here is the modest sample size. The omics-wide association studies reported here have low statistical power on their own, due to the expectation of small effects across the omics. Meta-analysing results from multiple studies with similar data available would yield more robust results. The inherent difficulty in power calculations for exploratory multi-omics studies is estimating a meaningful effect size *a priori*. Moreover, power calculations are made with a primary research question in mind, while integrative studies are often conducted as a secondary analysis. For predictive models, sample size considerations are developed for low-dimensional datasets only (Riley et al., 2019a,b, 2020).

The NMR metabolomics data had the largest sample size and was the best-powered analysis, however the observational nature of the analysis cannot rule out unmeasured confounding or reverse causation, therefore replication analyses are needed. In addition, the serum metabolites analysed here are only a small part of the full metabolome.

In summary, the extensive methods applied here for the first time on these data demonstrate the vast potential of integrative data analysis methods to improve the understanding of complex trait biology. The results give evidence for associations between ADHD symptoms and multiple metabolic measures, independently of BMI. Additionally, based on the multi-omics variable selection, ApoBtoApoA1 may have a role in the potential molecular interplay between maternal smoking and ADHD.

Chapter 8

Causality between ADHD and obesity-related traits

ADHD often co-occurs with obesity (Cortese et al., 2016; Cortese & Tessari, 2017), and maternal obesity at the start of pregnancy has been associated with offspring ADHD symptoms in observational studies (Andersen et al., 2018; Rodriguez et al., 2008; Rodriguez, 2010; Sanchez et al., 2018; E. L. Sullivan et al., 2015). However, there is known shared genetic aetiology between ADHD and BMI (Demontis et al., 2019; Du Rietz et al., 2018), and genetic familial confounding may account in part for the association between high maternal pre-pregnancy BMI and offspring ADHD (Q. Chen et al., 2013) (Figure 8.1). This familial co-aggregation could also account for the comorbidity of ADHD and obesity. Furthermore, it is not known if the shared genetic aetiology is driven by either of the core symptom subtypes, inattention or hyperactivity.

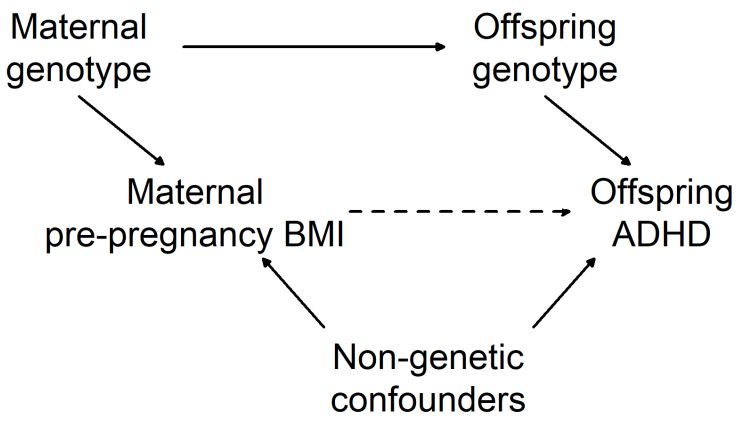

Figure 8.1: Potential relations between maternal pre-pregnancy BMI and offspring ADHD.

GWAS on both ADHD and BMI have shown considerable polygenicity (Demontis et al., 2019; Yengo, Sidorenko, et al., 2018). The effects of multiple common DNA variants from GWAS summary statistics can be aggregated to a polygenic risk score (PRS, Section 4.2.1.1) (Dudbridge, 2013), which can be used in epidemiological studies as a measure for the genetic liability to a trait.

The aim of this study was to examine the evidence for causality that may explain the ADHD-obesity association using genetically informed methods. I used bidirectional two-sample MR to investigate the potential causality between ADHD and BMI, as well as other obesity-related traits. To investigate the separate contributions of inattention and hyperactivity symptoms on the shared genetic aetiology between ADHD and BMI, I conducted association analysis of offspring BMI PRS and ADHD symptoms. Finally, to examine the effect of maternal obesity on offspring ADHD symptoms independent of the genetic liability, I conducted the association analysis between maternal pre-pregnancy BMI and offspring ADHD symptoms, adjusted for the inherited genetic liability to ADHD and BMI as measured by offspring PRS, using NFBC1986 dataset. Table 8.1 shows the conducted analyses and the datasets used in each step.

Table 8.1: Analyses and datasets used in this chapter.

	Analysis	Datasets
I	Bidirectional MR on ADHD and multiple obesity-related traits	GWAS summary statistics
II	Association analysis of offspring BMI PRS and offspring ADHD symptoms	NFBC1986
III	Association analysis of maternal BMI and offspring ADHD symptoms	NFBC1986

8.1 Methods

8.1.1 Bidirectional MR on ADHD and obesity-related traits

I conducted bidirectional two-sample MR on ADHD and obesity-related data using summarised data. The GWAS summary statistics were obtained from the largest and most recent GWAS available for ADHD and six obesity-related traits: BMI, waist circumference (WC), waist-hip-ratio (WHR), BMI-adjusted WHR, body fat percentage (BFP) and basal metabolic rate (BMR) (Table 8.2). Summary statistics for ADHD were obtained from Demontis et al. (2019). Summary statistics for BMI are combined from the GIANT consortium (Locke et al., 2015) and UKBiobank (Sudlow et al., 2015), comprising almost 700,000 individuals (Yengo, Sidorenko, et al., 2018). For WHR and BMI-adjusted WHR, I used summary statistics from GIANT consortium (Shungin et al., 2015). For WC, BFP and BMR, I used summary statistics from UK Biobank provided by the Neale Lab (the Neale Lab, 2018).

Table 8.2: Information of GWAS on ADHD and obesity-related traits.

Trait	Abbr	Source	N
ADHD	ADHD	Demontis et al.	55,374
Body mass index	BMI	Yengo et al.	688,566
Waist circumference	WC	UK Biobank	360,564
Waist-hip-ratio	WHR	GIANT	144,595
BMI-adjusted WHR	WHR (adj. BMI)	GIANT	626,892
Body fat percentage	BFP	UK Biobank	354,628
Basal metabolic rate	BMR	UK Biobank	354,825

Abbr = Abbreviation; N = Sample size in GWAS.

SNPs that were strongly associated with the exposure ($p < 10^{-8}$ for BMI as suggested in

(Yengo, Sidorenko, et al., 2018), $p < 5 \times 10^{-8}$ for other obesity-related traits and ADHD) were selected to proxy the exposure. The effect estimates of these SNPs were extracted from the GWAS summary statistics of both the exposure and the outcome and aligned to have the same effect allele. The SNPs were clumped (Privé et al., 2018) using a window of 1,000 kb and an r^2 threshold of 0.01. For those SNPs not available in the outcome summary statistics, proxies were sought using an r^2 cut-off of 0.8.

The IVW method based on summarised data was used for the main analysis. To combine the evidence from IVW MR results with ADHD as the exposure, I combined the p-values from IVW results using harmonic mean p-value (Good, 1958; Wilson, 2019). This method can be used to combine dependent p-values while controlling for FWER (Wilson, 2019).

To examine the robustness of the results related to the assumption of no horizontal pleiotropy (Section 4.3.3.3), I conducted additional MR analyses using weighted median method (Bowden et al., 2016) and MR-PRESSO (Verbanck et al., 2018), and tested for horizontal pleiotropy using MR-PRESSO Global test (Verbanck et al., 2018).

For the MR analyses with ADHD as the exposure, I used birth weight and hair colour as negative control outcomes. This was done to examine the validity of the genetic variants of ADHD as the instrumental variable for ADHD (Hemani, Bowden, & Davey Smith, 2018). No association between the ADHD instrumental variable and negative control outcomes is expected, and any detected association would indicate invalidity of the genetic variants of ADHD as the instrument for ADHD (Hemani, Bowden, & Davey Smith, 2018). For the MR analysis with ADHD as the outcome, the results are presented on the log-odds scale.

8.1.2 Phenotypic measures

NFBC1986 was used for individual-participant analysis. The phenotypic variables used were ADHD symptoms measured at eight years rated by parents and teachers using Rutter A and Rutter B scales, respectively, and at 16 years rated by parents using SWAN scale.

For data at eight years, inattention and hyperactivity symptom scores were constructed separately for each scale based on questions on the core symptoms as described in Rodriguez et al. (2007). For data at 16 years, the weakness sides of inattention and hyperactivity subscales were used, in order to make this measurement comparable with the variables at the earlier time point. To ensure the comparison of the scores between raters and time

points, all symptom score variables were scaled to range from zero (no symptoms) to two (maximum symptoms).

I also created three global scores to summarise the data by aggregating ADHD symptom scores across all raters and ages: global inattention, global hyperactivity, and combined global inattention-hyperactivity symptom scores. Both phenotypic measures and genetic data were available for 2,984 individuals.

8.1.3 Polygenic risk scores (PRS)

PRS for offspring ADHD and offspring BMI were calculated as a weighted sum of the number of risk-increasing alleles (equation (4.17)), with effect size estimates for each SNP obtained from GWAS summary statistics for the corresponding phenotype used as the weights. PRSice software (Euesden et al., 2014) was used to calculate PRS in NFBC1986. The number of SNPs was optimised by first clumping (Privé et al., 2018) the summary statistcs (clumping window of 250 kb and $r^2 = 0.01$) and then calculating the PRS with p-value thresholds of $1 \times 10^{-8}, 5 \times 10^{-8}, 1 \times 10^{-5}, 1 \times 10^{-4}, 0.001, 0.01, 0.1, 0.2, 0.3, 0.4$ and 0.5. Clinically measured BMI at 16 years and global combined ADHD symptom score were used as the target phenotypes for BMI and ADHD PRS, respectively. The PRS with the highest R^2 with the target phenotype was selected as the final PRS for the corresponding trait.

8.1.4 PRS association analysis for BMI and ADHD symptoms

To study the shared genetic aetiology between ADHD symptom subtypes and BMI, I conducted association analysis of offspring BMI PRS and inattention and hyperactivity symptoms in NFBC1986 by multiple outcomes ordinal regression (Hirk et al., 2019). In this method, the observed response $\mathbf{Y}^*$ is assumed to be a realisation of a latent variable $\tilde{\mathbf{Y}}$, which in turn depends on the covariates $\mathbf{X}$ by

$$\tilde{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is the error term which follows a multivariate logistic distribution (Hirk et al., 2019; O'Brien & Dunson, 2004). The link between the unobservable latent variable and the observed J categorical outcomes with K_j ordered categories is given by

$$Y_{ij} = r_{ij} \Leftrightarrow \delta_{j,r_{ij}-1} < \tilde{Y}_{ij} \leq \delta_{j,r_{ij}}, \quad r_{ij} \in \{1, \ldots, K_j\},$$

where $-\infty = \delta_{j,0} < \delta_{j,1} < \cdots < \delta_{j,K_j-1}, < \delta_{j,K_j} = \infty$ are threshold parameters for outcome $j,\ j = 1, \ldots, J$. (Hirk et al., 2019). The model included sex and the first ten genetic PCs as additional explanatory variables.

8.1.5 Association analysis for maternal pre-pregnancy BMI and offspring ADHD symptoms

I expanded the previously analysed association between maternal pre-pregnancy BMI and offspring ADHD symptoms rated by teachers (Rodriguez et al., 2008), by adjusting association analysis for the genetic liability of the phenotypes as measured by offspring PRS. Ordinal regression with a logit link function was used for the analysis. The additional explanatory variables in the model were parity, maternal education, smoking during pregnancy, age at delivery and offspring sex. As previous literature shows a J-shaped association between maternal BMI and offspring ADHD (Rodriguez, 2010), restricted cubic splines with three knots were used to allow non-linear effects for maternal BMI and other continuous explanatory variables.

8.2 Results

8.2.1 Bidirectional MR on ADHD and obesity-related traits

Depending on the trait, the number of independent SNPs available for MR with ADHD as the exposure was between nine and 12 (Table 8.3). The main MR analysis using IVW method showed evidence for genetically predicted liability to ADHD being associated with higher BMI (MR point estimate 0.053, 95 % CI [0.002, 0.103], p-value = 0.04), WC (0.045 [0.005, 0.086], $p = 0.03$), WHR (0.068 [0.009, 0.127], $p = 0.03$) and BMI-adjusted WHR (0.052 [0.020, 0.084], $p = 0.006$) (Figure 8.2). The point estimates for the other adiposity traits were also positive (BFP: 0.021 [-0.022, 0.064], $p = 0.31$; BMR: 0.019 [-0.001, 0.045], $p = 0.12$). Combining the evidence from all MR results using IVW method yielded a harmonic p-value of 0.025.

Table 8.3: Genetic variants used in MR analyses with ADHD as exposure and obesity-related traits as outcomes.

SNP	BMI	WC	WHR	WHR (adj. BMI)	BFP	BMR
rs11420276	rs12410155	Y	rs12410155	rs12410155	Y	Y
rs1222063	rs2391769	Y	rs2391769	rs2391769	Y	Y
rs9677504	Y*	Y	Y	Y	Y	Y
rs4858241	Y	Y			Y*	Y
rs28411770		Y*			Y*	Y
rs4916723	Y	Y	Y	Y	Y	Y
rs5886709	Y	rs10262192	Y	Y	rs10262192	rs10262192
rs74760947	Y*	Y			Y	Y
rs11591402	Y*	Y	Y	Y	Y	Y
rs1427829	Y*	Y*	Y	Y	Y*	Y*
rs281324	Y	Y	Y	Y	Y*	Y
rs212178	rs12596294	Y	rs12596294	rs12596294	Y	Y

'Y' indicates that the SNP was used as an instrumental variable. If a proxy SNP was used, this SNP is given.

* SNP was removed from MR-PRESSO analysis due to evidence for heterogeneity.

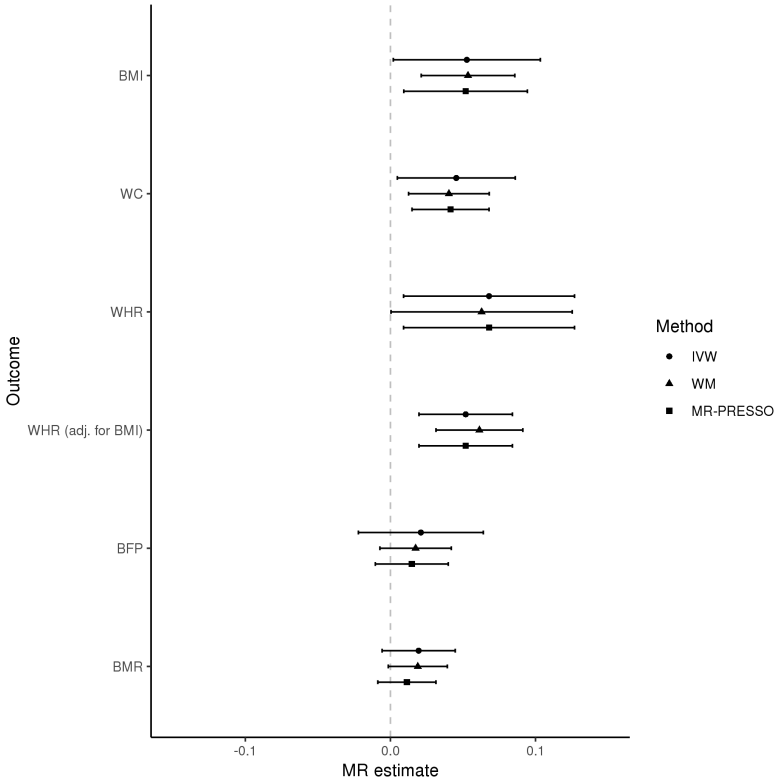

Figure 8.2: Forest plot of MR estimates and their 95% confidence intervals for the effect of genetically predicted liability to ADHD on obesity-related traits. WM = weighted median method.

Using weighted median and MR-PRESSO methods, which are more robust to violations of no horizontal pleiotropy assumption, showed similar results to the main analysis (Figure

8.2). All effect size estimates of genetically predicted liability to ADHD on obesity-related traits using weighted median method were similar to the IVW estimates. MR-PRESSO global test detected evidence for horizontal pleiotropy for all outcomes except waist-hip-ratio. Removing the potentially pleiotropic SNPs based on MR-PRESSO outlier test did not notably affect the effect size estimates, as demonstrated by the similarity of IVW and MR-PRESSO estimates (Figure 8.2).

In the negative control outcome analysis, there was no evidence for association between genetically predicted liability to ADHD and birth weight (IVW estimate -0.015 [-0.050, 0.020], $p = 0.36$) or hair colour (IVW estimate 0.002 [-0.010, 0.014], $p = 0.78$).

In MR analyses treating obesity-related traits as exposures and the risk of ADHD as the outcome, higher genetically predicted BMI, WC, BFP and BMR were associated with an increased risk of ADHD (Figure 8.3). Positive effect size estimates were detected for WHR and BMI-adjusted WHR. The results from sensitivity analyses accounting for horizontal pleiotropy had similar effect sizes as the IVW method (Figure 8.3).

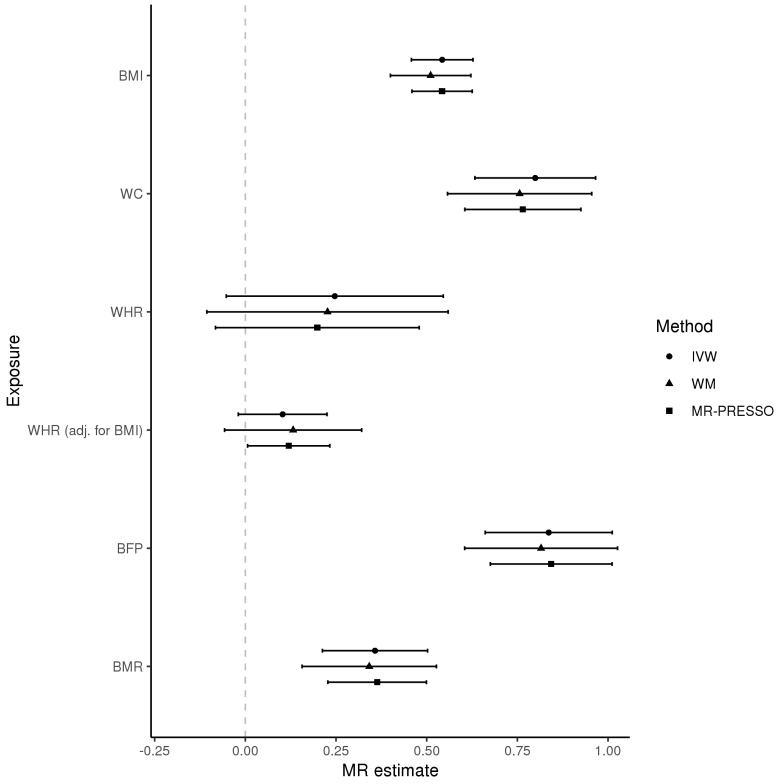

Figure 8.3: Forest plot of MR estimates and their 95% confidence intervals for the effect of genetically predicted obesity-related traits on the risk of ADHD on the log-odds scale. WM = weighted median method.

8.2.2 PRS associations with target phenotypes

The optimal p-value thresholds for offspring BMI and ADHD PRS were 0.01 and 0.2, respectively. Offspring BMI PRS explained 0.088 of the variation in the clinically measured BMI at 16 years. The distribution of the global ADHD symptom score was highly skewed (Figure 8.4), and offspring ADHD PRS explained 0.009 of the variation in the log-transformed global ADHD symptom score.

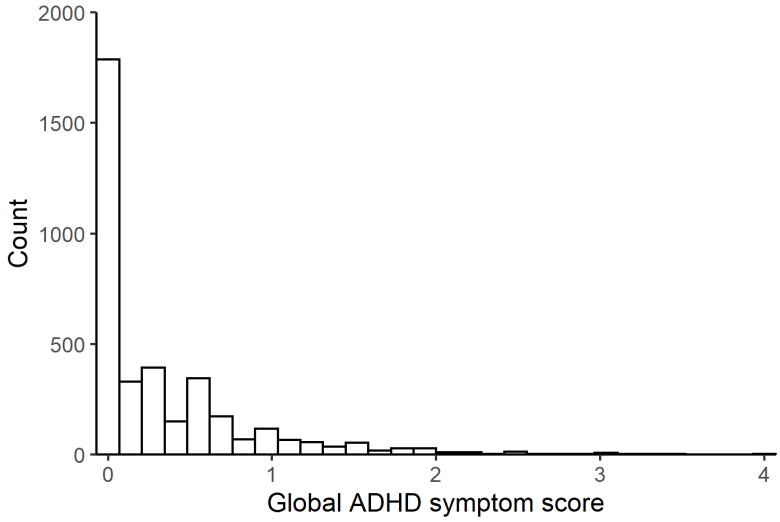

Figure 8.4: Histogram of the global ADHD symptom score.

8.2.3 PRS association analysis for BMI and ADHD symptoms

In the association analysis, offspring BMI PRS was associated with both global inattention and global hyperactivity. The odds ratios (OR) and their CIs for increasing number of symptoms per 1-standard-deviation (SD) increase in BMI PRS were identical for both inattention and hyperactivity (OR = 1.17, 95% CI [1.09, 1.25], p-value for difference of estimates = 0.99). Similar effect size estimates were observed when analysing each rater and time point separately (Figure 8.5). Offspring BMI PRS was also associated with global inattention-hyperactivity score (OR per 1-SD increase in BMI PRS 1.17, 95% CI [1.10, 1.24]). Table A.5 shows the descriptive statistics for the NFBC1986 observational data.

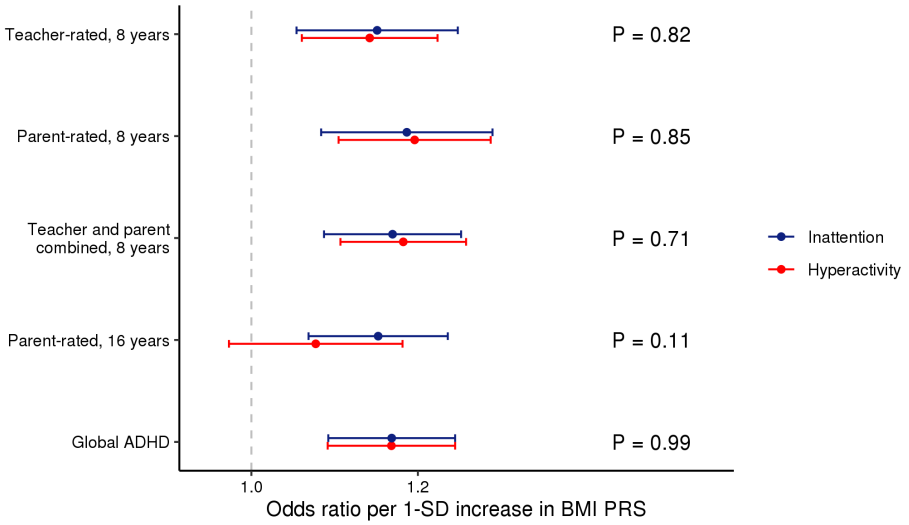

Figure 8.5: Effect size estimates and their 95% confidence intervals per 1-standard deviation increase in BMI PRS on increasing number of ADHD symptoms. *P*-values are for testing the null hypothesis of no difference between the effect sizes on inattention and hyperactivity.

8.2.4 Association analysis for maternal pre-pregnancy BMI and offspring ADHD symptom types

The associations between maternal pre-pregnancy BMI and the log-odds of the predicted offspring ADHD symptoms both before and after adjustment for offspring BMI PRS and ADHD PRS are characterised in Figures 8.6 and 8.7. Maternal pre-pregnancy BMI was associated with offspring ADHD symptoms rated by teachers. Some attenuation of the relationship was seen after adjusting for offspring BMI and ADHD PRS, however there was evidence for association even after the adjustment; *p*-values for testing the null hypothesis of no association in the PRS-adjusted model were 0.027 and 0.008 for inattention and hyperactivity, respectively.

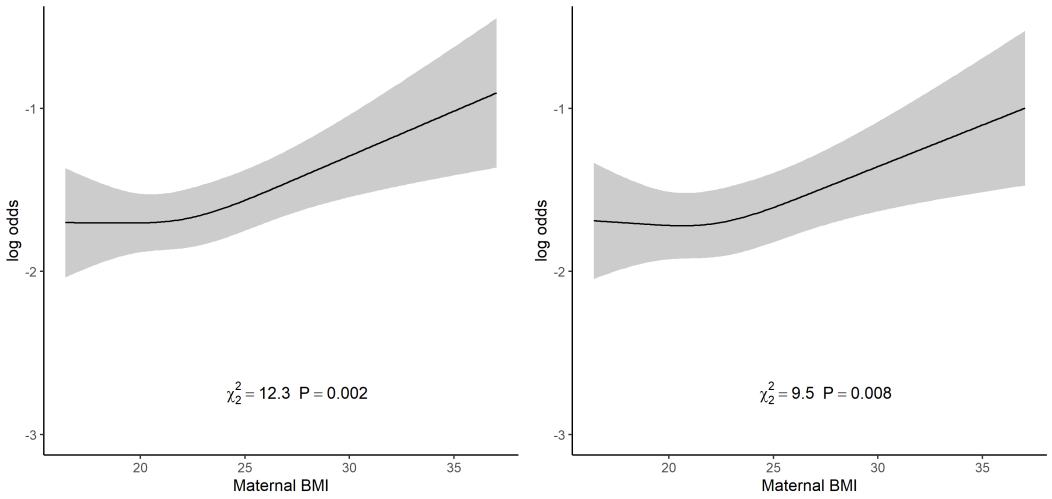

Figure 8.6: Association between maternal BMI (x-axis) and the log-odds of an increased number of hyperactivity symptoms at eight years, rated by teachers (y-axis). Left panel shows the association without adjustment for PRS of ADHD and BMI, and right panel shows the association with the adjustment. The predicted log-odds values are negative because the underlying probability for ADHD symptoms is low.

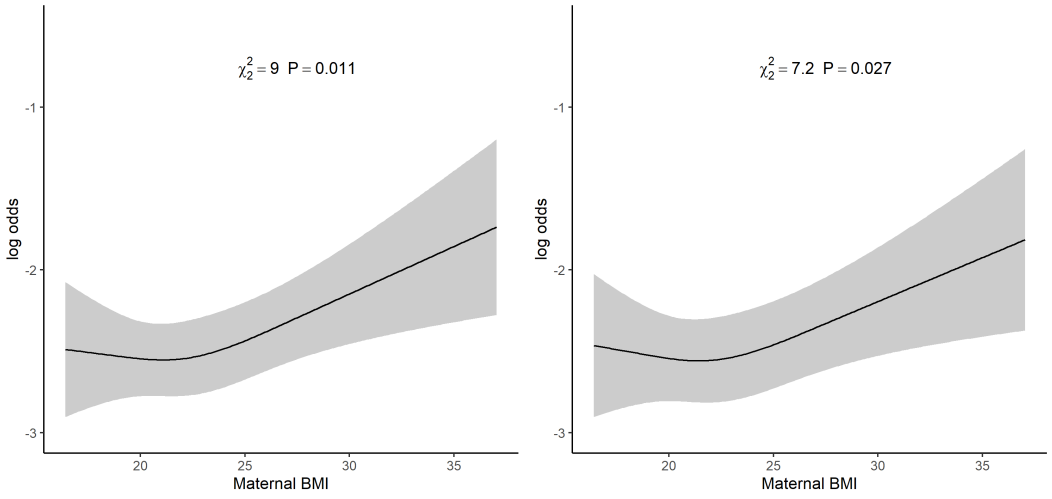

Figure 8.7: Association between maternal BMI (x-axis) and the log-odds of an increased number of inattention symptoms at eight years, rated by teachers (y-axis). Left panel shows the association without adjustment for PRS of ADHD and BMI, and right panel shows the association with the adjustment. The predicted log-odds values are negative because the underlying probability for ADHD symptoms is low.

8.3 Discussion

To explore potential causality between ADHD and obesity, I conducted bidirectional MR on ADHD and five obesity-related traits. The results showed consistent evidence for a bidirectional causal association. In addition, I studied the shared genetic architecture of BMI and ADHD symptoms and found the genetic liability to BMI being associated with both inattention and hyperactivity symptoms in childhood and adolescence. Finally, association analysis showed evidence that prenatal exposure to maternal overweight/obesity was associated with an increased risk of ADHD symptoms in children after accounting for the genetic liability as measured by offspring PRS.

These results are consistent with the suggested bidirectional association derived from evidence across observational studies (Cortese & Tessari, 2017). Different mechanisms are likely to underlie each possible direction. It is hypothesised that impulsivity and inattention symptoms may increase the liability to obesity via disrupted eating patterns (Ptacek et al., 2014). On the other hand, obesity might lead to higher risk of ADHD via e.g. obesity-induced sleep disruption or inflammatory mechanisms (Cortese, 2019; Cortese & Vincenzi, 2012). The results partially contrast with the only other MR analysis of ADHD and BMI (Martins-Silva et al., 2019), which reported no consistent evidence for the ADHD to BMI pathway. Here, I used five obesity-related traits, rather than relying solely on BMI. After combining the evidence across these five traits via harmonic p-value, there was MR evidence for ADHD being causal on obesity. The results were consistent in the sensitivity MR analysis more robust to horizontal pleiotropy, and negative control outcome analysis suggested that the variants used as instruments for ADHD worked as intended.

I studied the genetic overlap between inattention, hyperactivity and the combined symptoms with BMI using PRS. PRS offers the advantage of summarising genetic effects and providing a polygenic signal from a set of markers that individually explain only a small fraction of the trait's variance. These results are in line with the earlier evidence for genetic overlap between BMI and ADHD (Du Rietz et al., 2018). Here, I also assessed the ADHD symptoms separately and globally both in childhood and adolescence, using two informants (teachers and parents). The results were robust across ages and raters in the sample of 2,984 participants. There was no evidence for differences in the associations between BMI PRS and inattention and hyperactivity symptoms, suggesting that neither type of symptom is

driving the shared genetic aetiology with BMI.

An earlier study by Rodriguez et al. (2008), which included the same NFBC1986 dataset used here, had showed an association between maternal pre-pregnancy overweight/obesity and ADHD symptoms in children. The novelty in this work is that the associations were further adjusted for both BMI and ADHD PRS.

These results apply for the full continuum of ADHD symptoms in the population rather than extreme or diagnosed cases, which is subject to biases, including healthcare accessibility. ADHD symptoms were measured longitudinally in childhood and adolescence. The findings are consistent with shared genetic aetiology being comparable across development. Furthermore, the findings were similar across raters and in the composite global measure, suggesting that the results were not affected by rater bias.

The results should be interpreted in light of its limitations. One limitation of MR is the interpretation of the MR effect size estimate with a binary exposure. The genetic variants associated with ADHD were obtained from a GWAS where ADHD was considered as a binary trait, representing only the extreme of a continuous dimension of ADHD symptoms. When using a binary exposure, the MR point estimate does not have a clear interpretation (Burgess & Labrecque, 2018). However, testing for the causal null hypothesis is still valid (VanderWeele et al., 2014). The point estimates and their confidence intervals are reported here for the sake of completeness.

I used the GWAS summary statistics that are currently available which means that the genetic information on ADHD is based on diagnosed ADHD patients and includes just over 55,000 individuals in total (cases and controls), thus the precision for genetic instruments and PRS for ADHD contrasts with the GWAS data for the adiposity traits – the BMI PRS was based on summary statistics of almost 700,000 people. Despite this difference, the negative control analyses showed that the genetic instruments for ADHD functioned as aimed. As of writing, no summary statistics from substantially large GWAS are available for inattention and hyperactivity, and thus, sufficiently powered MR analysis or PRS for the separate ADHD symptom subtypes are not yet feasible.

For the association between maternal pre-pregnancy BMI and offspring ADHD symptoms, I was able to use the genotype of the offspring for the adjustment. As the offspring genotype is also influenced by father's genotype, adjusting for the offspring genotype might cause bias

in the results due to the fact that father's genotype may affect offspring ADHD symptoms via other pathways than through offspring PRS only. A more stringent adjustment for the genetic predisposition would include either maternal genotype or both parental genotypes, neither of which were available here.

Overall, this study adds to the previous literature of the comorbidity between ADHD and obesity by using genetically informative methods to infer potential causality between the traits. The results suggest a complex, bidirectional association between ADHD and obesity, and evidence for an intergenerational effect of maternal pre-pregnancy BMI on offspring ADHD symptoms. Further analyses using parental genotypes would provide a more comprehensive picture of the putative intergenerational causal association.

Chapter 9

Circulating cytokine levels and psychiatric outcomes

As discussed throughout this thesis, smoking and obesity are associated with ADHD, both intergenerationally and in the individual. Both smoking and obesity are known to induce systemic chronic inflammation (Section 2.4) (Furman et al., 2019; Gonçalves et al., 2011; Gregor & Hotamisligil, 2011). There is also accumulating evidence that supports a neuroinflammatory component in psychiatric illnesses in general (Najjar et al., 2013). Therefore, the link between chronic inflammation and neuropsychiatric outcomes warrants more detailed investigation.

Cytokines, chemokines, growth factors and interferons (hereafter cytokines) are circulating protein signalling molecules (Furman et al., 2019). They have a key role in a range of cellular processes and can act as biomarkers for chronic inflammation (Liu et al., 2017; Wahl et al., 1989). There is evidence for associations between inflammatory biomarkers and psychiatric outcomes (Yuan et al., 2019), such as ADHD (G. A. Dunn et al., 2019), autism spectrum disorder (ASD) (Xu et al., 2015), bipolar disorder (BPD) (Muneer, 2016), depression (Dowlati et al., 2010) and schizophrenia (Na et al., 2014).

Previous GWAS have identified genetic variants associated with circulating cytokine levels (Ahola-Olli et al., 2017; Sliz et al., 2019). For molecular exposures, such as proteins, the genetic variants used in MR can be selected from the proximity of the corresponding coding gene. Using variants with such biological relevance may improve the plausibility of the vari-

ants fulfilling the instrumental variable assumptions (Burgess et al., 2020), and consequently provide more robust causal evidence.

Here, I conducted GWAS meta-analysis on 47 circulating cytokines in up to 13,365 individuals to obtain genetic variants associated with the circulating cytokine levels. Further incorporating additional eQTL information, I identified genetic instruments with *a priori* high biological relevance to proxy the effect of varying circulating cytokine levels (Figure 9.1). These instruments were used in MR framework to examine the putative causal effects of circulating cytokine levels on five common neuropsychiatric outcomes: ADHD, ASD, BPD, major depressive disorder (MDD) and schizophrenia.

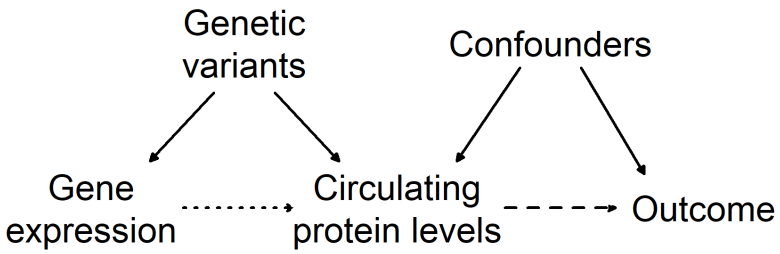

Figure 9.1: A DAG on instrumental variable selection for circulating protein levels. The interest is in the effect of circulating protein levels on an outcome (dashed line). Choosing genetic variants that are associated with both gene expression and circulating protein levels increase the plausibility that the variants used are on the biological signalling pathway from expression to protein levels (dotted line), and independent of confounders.

9.1 Materials and methods

9.1.1 Datasets

I conducted GWAS for 16 cytokines in NFBC1966 and obtained summary statistics for 41 cytokines in YFS and FINRISK (Table 9.1) (Ahola-Olli et al., 2017). The data preprocessing was done in a similar manner to previous GWAS analyses (Ahola-Olli et al., 2017; Sliz et al., 2019). Inverse normal rank transformation was first applied to the cytokine variables, before regressing the transformed measures on age, sex and the first 10 genetic PCs. In contrast to the previous analyses (Ahola-Olli et al., 2017; Sliz et al., 2019), BMI was not added in the regression model, as this could potentially introduce collider bias into consequent MR analyses (Day et al., 2016). The inverse normal rank transformation was again applied to the residuals of this regression, and these transformed residual estimates were used as response variables in the GWAS. The GWAS was conducted using an additive genetic model with SNPTEST2 software. The results for variants which showed poor imputation quality (model info < 0.7) or low MAF (<0.05) were discarded.

Table 9.1: Cytokines and their data sources.

Cytokine	Abbreviation	N_{NFBC1966}	N_{FINRISK}	N_{YFS}	N_{total}
Active plasminogen activator inhibitor-1	activePAI1	5,199			5,199
Beta nerve growth factor	βNGF		1,620	1,950	3,531
Cutaneous T-cell attracting chemokine	CTACK		1,651	2,019	3,631
Eotaxin	Eotaxin		6,186	2,011	8,153
Basic fibroblast growth factor	FGFBasic		5,592	2,017	7,565
Granulocyte colony-stimulating factor	GCSF		1,544	2,018	7,904
Growth regulated oncogene-alpha	GROα		1,541	2,003	3,505
Hepatocyte growth factor	HGF		6,317	2,019	8,292
Interferon-gamma	IFNγ		5,726	2,019	7,701
Interleukin-10	IL10		5,708	2,016	7,681
Interleukin-12p70	IL12p70		6,295	2,019	8,270
Interleukin-13	IL13		1,577	2,019	3,557
Interleukin-16	IL16		1,663	1,858	3,483
Interleukin-17	IL17	5,071	5,785	2,019	12,831
Interleukin-18	IL18		1,656	2,019	3,636
Interleukin-1-alpha	IL1α	5,014			5,014
Interleukin-1-beta	IL1β	5,067	1,330	2,018	8,376
Interleukin-1 receptor antagonist	IL1ra	4,957	1,658	2,019	8,595
Interleukin-2	IL2		1,498	2,016	3,475
Interleukin-2 receptor, alpha subunit	IL2rα		1,704	2,012	3,677
Interleukin-4	IL4	5,059	6,149	2,019	13,183
Interleukin-5	IL5		1,386	2,017	3,364

Table 9.1: Cytokines and their data sources. *(continued)*

Cytokine	Abbreviation	N_{NFBC1966}	N_{FINRISK}	N_{YFS}	N_{total}
Interleukin-6	IL6	5,063	6,215	2,018	13,252
Interleukin-7	IL7		1,429	2,019	3,409
Interleukin-8	IL8	5,071	1,546	2,019	8,597
Interleukin-9	IL9		1,656	2,017	3,634
Interferon gamma-induced protein 10	IP10	5,072	1,705	2,019	8,757
Monocyte chemotactic protein-1	MCP1	5,072	6,318	2,019	13,365
Monocyte specific chemokine 3	MCP3		843	256	843
Macrophage colony-stimulating factor	MCSF		1,632	866	840
Macrophage migration inhibitory factor	MIF		1,516	2,017	3,494
Monokine induced by interferon-gamma	MIG		1,705	2,019	3,685
Macrophage inflammatory protein-1a	MIP1α		1,542	2,019	3,522
Macrophage inflammatory protein-1b	MIP1β		6,268	2,019	8,243
Platelet derived growth factor BB	PDGFbb		6,318	2,019	8,293
Regulated on Activation, Normal T Cell Expressed and Secreted	RANTES		1,585	1,869	3,421
Soluble CD40 ligand	sCD40L	5,067			5,067
Stem cell factor	SCF		6,316	2,018	8,290
Stem cell growth factor beta	SCGFβ		1,704	2,017	3,682
Stromal cell-derived factor-1 alpha	SDF1α		6,003	1,826	5,998
Soluble E-selectin	sE-selectin	5,199			5,199
Soluble intercellular adhesion molecule-1	sICAM1	5,199			5,199
Soluble vascular cell adhesion molecule 1	sVCAM1	5,199			5,199
Tumor necrosis factor-alpha	TNFα	5,068	1,474	2,019	8,522
Tumor necrosis factor-beta	TNFβ		1,450	116	1,559
TNF-related apoptosis inducing ligand	TRAIL		6,218	2,012	8,186
Vascular endothelial growth factor	VEGF	5,037	5,143	2,019	12,155

N = sample size; NFBC1966 = Northern Finland Birth Cohort 1966; YFS = Young Finns Study; FINRISK = FINRISK Study; * = The total sample size with full genomic and cytokine data after quality control.

For the ten cytokines available in both NFBC1966 and YFS+FINRISK (Table 9.1), I meta-analysed the summary statistics by inverse-variance weighted fixed-effects model using Metal software (Willer et al., 2010). Gene expression GWAS summary statistics were obtained from GTEx project (release version 7) that relate to 10,361 samples from a multi-ethnic group of 635 individuals (GTEx Consortium, 2017). Results from 53 tissues were pooled using fixed-effects meta-analysis to produce cross-tissue estimates of association with gene expression (GTEx Consortium, 2017).

9.1.2 Psychiatric outcomes

GWAS summary statistics for psychiatric outcomes were obtained based on the largest GWAS conducted by Psychiatric Genomics Consortium on ADHD (Demontis et al., 2019), ASD (Grove et al., 2019), BPD (Stahl et al., 2019), MDD (Wray et al., 2018), and schizophrenia (Ripke et al., 2014). Effective sample sizes (calculated as $2/(1/N_{\text{cases}} + 1/N_{\text{controls}})$) ranged from 22,183 for ASD to 194,520 for MDD (Table 9.2). All studies were based primarily on populations of European ancestries and all summary statistics are publicly available.

Table 9.2: Table of psychiatric outcomes.

Phenotype	Abbr	N_{cases}	N_{controls}	N_{eff}	Source
Attention-deficit/hyperactivity disorder	ADHD	20,183	35,191	25,653	Demontis et al. 2019
Autism spectrum disorder	ASD	18,381	27,969	22,183	Grove et al. 2019
Bipolar disorder	BPD	20,352	31,358	24,684	Stahl et al. 2019
Major depressive disorder	MDD	135,458	344,901	194,520	Wray et al. 2019
Schizophrenia		36,989	113,075	55,743	Ripke et al. 2014

Abbr = abbreviation; N = sample size; N_{eff} = effective sample size.

9.1.3 MR analysis

I generated instrumental variables using two different criteria. In the first criterion, variants within 500kb of the gene locus corresponding to the cytokine that associated with circulating cytokine levels at $p < 10^{-4}$ were selected. These are referred to as *cis*-protein QTL (*cis*-pQTL) instruments. In the second criterion, variants within 500 kb of the corresponding cytokine gene locus that associated with both tissue expression at $p < 10^{-4}$ and circulating cytokine levels at $p < 0.05$ were selected, which is referred to as *cis*-eQTL instruments. A more relaxed p-value threshold was used for circulating cytokine levels as it was used only to verify that the eQTL also associates with the protein levels.

The gene locations were extracted per human genome build 19 (released February 2009) using UCSC Genome Browser (University of California Santa Cruz, accessed on 18th June 2019). For one cytokine, namely soluble CD40 ligand (sCD40L), the coding gene *CD40L* is in X chromosome, for which the genetic data were not available. Cytokines work in ligand-receptor pairs, and CD40 acts as a receptor for CD40L. *CD40* gene within chromosome 20 is therefore on the same signalling pathway, so the biological relevance of selecting instruments within *CD40* applies.

To maximise the number of available instruments, exposure and outcome data were first merged, and then clumped (with $r^2 < 0.01$), rather than using pre-clumped instruments for all outcomes. The main MR analysis were conducted for cis-pQTL and cis-eQTL instruments separately using Wald ratio. For cytokines with more than one instrumental variable available, the IVW method was used.

F statistic was used to measure the strength of the instrument. For the cytokines that showed evidence for association in the main MR analysis and which had more than one instrumental variable available, I conducted MR sensitivity analysis by weighted median (Bowden et al., 2016) and MR-PRESSO (Verbanck et al., 2018) methods to examine the robustness of the results to potential horizontal pleiotropy.

To correct for multiple testing, I applied Bonferroni correction for five outcomes, obtaining a threshold of $p = 0.01$ for statistical significance. The MR sensitivity analyses were only performed to explore the robustness of the main IVW analysis to potential horizontal pleiotropy, and thus no formal multiple testing correction was applied to these estimates.

9.2 Results

9.2.1 Instrumental variables

I conducted GWAS meta-analysis to obtain genetic variants that are robustly associated with the circulating cytokine levels. The *cis*-pQTL and *cis*-eQTL instruments were available for 23 and 11 cytokines, respectively, with both types of instruments available for ten cytokines (Table 9.3). F statistics (as a measure of instrument strength, based on circulating protein levels) for all instrument variants ranged from 15 to 928 for *cis*-pQTLs and from 5 to 178 for *cis*-eQTLs (Table 9.3).

Table 9.3: Cytokines with instrumental variables available.

Cytokine	Coding gene	Chr	Start	End	$SNPs_{pQTL}$	$SNPs_{eQTL}$	F_{pQTL}	F_{eQTL}
activePAI1	*SERPINE1*	7	100770370	100782547	1	0	17	
CTACK	*CCL27*	9	34661893	34662689	3	1	25 - 142	53
Eotaxin	*CCL11*	17	32612687	32615199	2	1	21 - 24	12
GROα	*CXCL1*	4	74735109	74737019	5	0	17 - 201	
IL12p70	*IL12A*	3	159706623	159713806	1	0	18	
IL16	*IL16*	15	81517640	81605104	1	0	31	
IL18	*IL18*	11	112013974	112034840	6	1	17 - 93	48
IL1α	*IL1A*	2	113531492	113542971	0	1		5
IL2rα	*IL2RA*	10	6052657	6104333	10	1	17 - 178	178
IL7	*IL7*	8	79645007	79717758	1	0	15	
IP10	*CXCL10*	4	76942269	76944689	4	1	26 - 32	17
MCP1	*CCL2*	17	32582296	32584220	1	0	25	
MIF	*MIF*	22	24236565	24237409	2	3	20 - 39	7 - 15
MIG	*CXCL9*	4	76922623	76928641	2	0	16 - 41	
MIP1β	*CCL4*	17	34431220	34433014	22	0	18 - 928	
RANTES	*CCL5*	17	34198496	34207377	1	0	27	
sCD40L	*CD40**	20	44746906	44758384	1	1	22	18
SCGFβ	*CLEC11A*	19	51226605	51228981	3	1	20 - 40	12
sE-selectin	*SELE*	1	169691781	169703220	3	0	16 - 29	
sICAM1	*ICAM1*	19	10381517	10397291	20	1	16 - 75	5
sVCAM1	*VCAM1*	1	101185196	101204601	1	1	19	17
TNFα	*TNF*	6	31543344	31546112	2	0	15 - 19	
TRAIL	*TNFSF10*	3	172223298	172241297	13	0	15 - 63	
VEGF	*VEGFA*	6	43737946	43754223	26	0	15 - 803	

Chr = chromosome; Start = gene start position (hg19); End = gene end position (hg19); SNPs = number of SNPs as instrumental variables; pQTL = protein quantitative trait loci; eQTL = expression quantitative trait loci.

* Receptor gene

9.2.2 MR results

Figure 9.2 shows all results from the main MR analysis. Using the *cis*-pQTL instruments, there was evidence for association between higher genetically predicted soluble CD40 ligand (sCD40L) levels and an increased risk of ADHD (OR per 1-standard deviation unit increase in the exposure 1.58, [95% CI 1.20; 2.12], $p = 0.001$), BPD (OR = 1.60 [1.20; 2.12], $p = 0.001$) and schizophrenia (OR = 1.57 [1.23; 1.97], $p = 0.0002$). Higher genetically predicted soluble vascular cell adhesion molecule-1 (sICAM1) levels were associated with higher risk of MDD (OR = 1.06 [1.02; 1.09], $p = 0.001$), and higher genetically predicted soluble E-selectin (sE-selectin) levels were associated with lower risk of schizophrenia (OR = 0.79 [0.66; 0.94], $p = 0.008$).

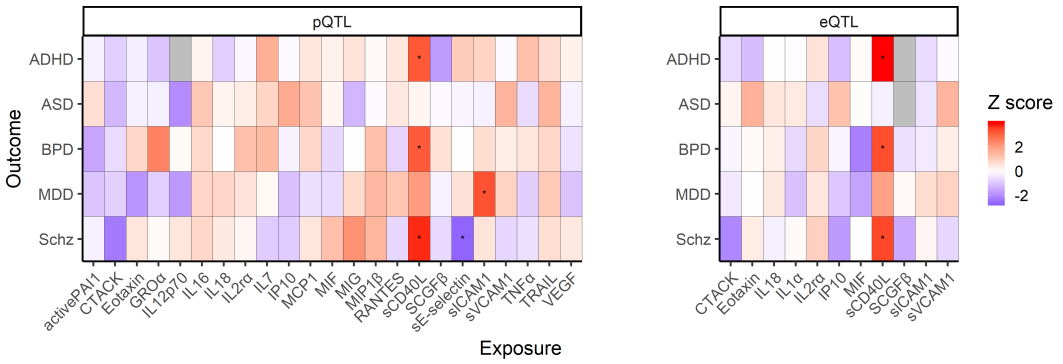

Figure 9.2: Mendelian Randomisation Z-scores for the effects of genetically predicted cytokine levels on the risk of psychiatric outcomes when considering cis-pQTL (left) and cis-eQTL (right) instruments. After performing a Bonferroni correction for testing of multiple disease outcomes, associations with $p < 0.01$ are denoted with an asterisk.

For the results with $p < 0.01$ in the main MR, sICAM1 was the only exposure that had three or more instrumental variables available (Table 9.3). Weighted median provided similar estimates to the main IVW analysis, and MR-PRESSO Global test did not find evidence for horizontal pleiotropy (Table 9.4). The associations for higher genetically predicted sCD40L and increased risk of ADHD, BPD and schizophrenia were all replicated using *cis*-eQTL instruments (Figure 9.2, Table 9.4).

Table 9.4: Results for top hits with p-value < 0.01 in the main MR analysis.

Instrument	Exposure	Outcome	Method	SNPs	OR (95% CI)	p-value
cis-pQTL	sCD40L	ADHD	Wald	1	1.58 (1.20; 2.12)	0.0013
		BPD	Wald	1	1.60 (1.20; 2.12)	0.0015
		Schz	Wald	1	1.57 (1.23; 1.97)	1.6e-04
	sE-selectin	Schz	IVW	2	0.79 (0.66; 0.94)	0.0076
	sICAM1	MDD	IVW	20	1.06 (1.02; 1.09)	9.3e-04
			Weighted Median	20	1.06 (1.01; 1.12)	0.015
			MR-PRESSO			0.86*
cis-eQTL	sCD40L	ADHD	Wald	1	1.82 (1.36; 2.46)	7.1e-05
		BPD	Wald	1	1.68 (1.25; 2.25)	7.3e-04
		Schz	Wald	1	1.54 (1.21; 1.95)	5.4e-04

pQTL = protein quantitative trait loci; eQTL = expression quantitative trait loci; SNPs = number of SNPs as instrumental variables; OR = odds ratio; CI = confidence interval; IVW = inverse-variance weighted.

* p-value for MR-PRESSO Global Test of horizontal pleiotropy

The results of the two types of instrumental variables for the effects of genetically predicted cytokine levels on the outcome traits were compared against each other. There was a positive correlation between the *cis*-pQTL main MR estimates and the *cis*-eQTL main MR estimates

(Pearson's correlation coefficient 0.79, p-value $= 2.1 \times 10^{-11}$, Figure 9.3). After removing the MR estimates with *cis*-pQTL effect size > 0.2, the Pearson's correlation coefficient was 0.48.

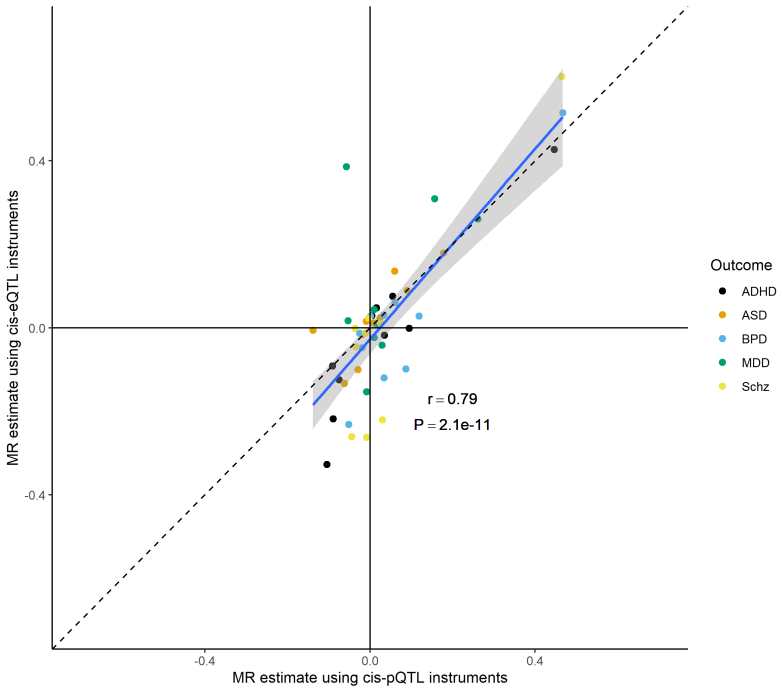

Figure 9.3: Scatterplot between *cis*-pQTL main MR estimates (x-axis, log-odds scale) and *cis*-eQTL main MR estimates (y-axis, log-odds scale). The blue line is the regression line, and the shaded area its 95% confidence interval. The dashed line is the reference line that indicates the equality of *cis*-pQTL and *cis*-eQTL MR estimates.

9.3 Discussion

I conducted a large GWAS meta-analysis on circulating cytokine levels that can act as biomarkers for chronic inflammation, and investigated their putative causal effects on neuropsychiatric outcomes. GWAS was conducted in up to 13,365 individuals, and the results were integrated with genome-wide eQTL data aggregated across 53 tissues in 635 individuals. These data were used to identify robust genetic instruments with plausible biological relevance. These instruments were then applied in MR setting to investigate the effects of circulating cytokines on the risk of psychiatric outcomes.

The MR results show a consistent pattern of higher genetically determined sCD40L levels being associated with increased risk of ADHD, BPD and schizophrenia. These results are consistent with an earlier study by Hope et al. (2015), who examined inflammatory markers

and cognitive function among schizophrenia and bipolar disorder patients and healthy controls, and found sCD40L being negatively associated with general cognitive function. The signalling pathway of CD40 and its ligand CD40L plays an important role in neuroinflammation (K. Chen et al., 2006).

The results showed evidence for protective effects of circulating sE-selectin levels on the risk of schizophrenia. Selectins are involved in chronic and acute inflammatory processes (Ley, 2003) and have a key role in the pathogenesis of inflammatory diseases (Impellizzeri & Cuzzocrea, 2014). In an earlier study, Iwata et al. (2007) found no evidence for differential sE-selectin levels between unmedicated schizophrenia patients and healthy controls, albeit with a small sample size.

There was also MR evidence for higher circulating sICAM1 levels increasing the risk of MDD. This is in line with the observational evidence of increased sICAM1 levels in psychiatric conditions (Müller, 2019). ICAM1 has a key function in regulating the movement of molecules into and out of the central nervous system (Müller, 2019). The soluble form sICAM1 is found in serum and may act as a biomarker for ICAM1 (Abe et al., 1998).

Cytokines are already targeted clinically for the prevention and treatment of a range of autoimmune and inflammatory processes, and all cytokines investigated here represent viable targets for pharmacological intervention (Rider et al., 2016). There is also evidence that cytokines are on the pathway of currently used pharmacological agents that are used to treat neuropsychiatric diseases (Robertson et al., 2019). Thus, the results may provide important clinical implications for the prevention of disease, which warrants further investigation and validation of this study's putative findings.

The methods that I applied to select genetic instrumental variables for the circulating cytokine levels aimed to maximise their biological validity by only considering variants that located at the corresponding gene locus or, in case of sCD40L, the corresponding signalling pathway. For *cis*-eQTL instruments, information from GWAS on gene-expression was used, combined with a more lenient p-value threshold for circulating protein levels to verify the associations. These two different methods provided distinct but consistent findings in MR.

There are some limitations to this work. The ability to unravel the inflammatory component of neuropsychiatric outcomes is limited by the availability of the measured proteins and their genetic instruments. There is a range of cytokines that has been associated with the

outcomes considered which were either not measured in our data (e.g. soluble interleukin-2 receptor and schizophrenia, Goldsmith et al. (2016)) or for which no instruments were found (e.g. interleukin-8 and ASD, Matta et al. (2019)). The genetic instrumental variables used in MR represent the cumulative effect of genetic predisposition, and these effects may be greater than those obtained at a discrete clinical intervention (Gill et al., 2019). Thus, the MR results obtained here cannot be directly extrapolated to infer the effect of clinical intervention. MR results may be biased due to violations to instrumental variable assumptions or inclusion of weak instruments. To minimise the risk of horizontal pleiotropy, genetic instrumental variables were selected from the corresponding gene locus to increase their biological validity (Swerdlow et al., 2016). Where possible, MR sensitivity analyses were conducted using methods that are more robust to horizontal pleiotropy, and obtained similar results. I examined the potential weak instrument bias via F-statistic. 29 out of 32 instruments included only variants with $F > 10$, the only exceptions being cis-eQTL instruments for IL1α, MIF and sICAM1. Finally, gene expression to detect eQTL was aggregated across tissues, rather than restricting to the most biologically relevant tissue, brain, for neuropsychiatric outcomes. This may have an impact on the generalisability of the findings.

In summary, I assessed the putative causal associations between chronic inflammation and the risk of psychiatric outcomes by incorporating data from genomic, transcriptomic and proteomic datasets. I identified valid and biologically plausible instruments to use in MR and investigated the effects of circulating cytokine levels on the risk of psychiatric outcomes. These results add to the evidence for the role of inflammation in the aetiology of neuropsychiatric outcomes and provides suggestive causal associations for further validation.

Chapter 10

General discussion and conclusions

This chapter summarises the results obtained in the study chapters. In addition to the more detailed discussion and limitations given in each of the study chapters, general limitations of the present work and future perspectives in molecular epidemiology are considered here.

10.1 Summary of the results

In this thesis, I have conducted extensive analyses applying different statistical modelling approaches in multi-omics datasets to elucidate the molecular background of ADHD. In Study I (Chapter 5), associations between exposure to maternal smoking during pregnancy and DNA methylation in the offspring adolescence and adulthood were examined. The role of differential DNA methylation as a potential mediator on offspring later life disease outcomes was also evaluated. The results showed evidence for persistent perturbations in blood DNA methylation levels up to offspring's middle age. Furthermore, there was some evidence of differential blood DNA methylation at *GNG12* mediating the association between exposure to maternal smoking during pregnancy and offspring personality scales.

In Study II (Chapter 6), the potential mediating role of DNA methylation specifically on offspring ADHD symptoms was examined. A DNA methylation based risk score on the exposure to maternal smoking during pregnancy was developed. This risk score was distinct

of paternal smoke exposure and not driven by offspring's own smoking status. The results did not provide strong evidence for a mediating effect of maternal smoking related DNA methylation on offspring ADHD symptoms. These results add to the evidence of a complex relationship between ADHD and smoking, both intergenerationally and within individuals.

In Study III (Chapter 7), I performed variable selection and prediction modelling methods for the association between molecular omics datasets and ADHD symptoms. The results gave insight to the potential role of apolipoproteins in the molecular interplay between DNA methylation and ADHD symptoms.

The well-known comorbidity of ADHD and obesity was examined in Study IV (Chapter 8) by using genetically informed methods. The results suggest a bidirectional causal relationship between ADHD and obesity-related traits, with further intergenerational studies needed to fully disentangle the role of prenatal environment and whether it affects offspring ADHD symptoms.

In Study V (Chapter 9), I conducted large-scale MR analysis on the effect of circulating cytokine levels on the risk of psychiatric outcomes. The results shed light into the suggested inflammatory component in the aetiology of ADHD as well as BPD, MDD and schizophrenia.

Overall, this thesis contributes to the literature on genetic and other molecular epidemiology for psychiatric outcomes by comprehensive analyses of multiple omics datasets in relation to ADHD. The complex relationship between ADHD and smoking has attracted a lot of attention in psychiatric research. The intergenerational association of maternal smoking during pregnancy and offspring ADHD symptoms is suggested to be due to shared genetic aetiology rather than causal intrauterine mechanisms due to smoke exposure (Thapar & Rice, 2020). DNA methylation has been suggested as a molecular mediator for the potential intrauterine effect, however, the results obtained in this work did not provide supporting evidence for this hypothesis.

A contested hypothesis is that smoking may also be used as self-medication for ADHD, as it is sometimes perceived to ameliorate ADHD symptoms (van Amsterdam et al., 2018). Therefore, those expecting mothers with higher liability to ADHD are suggested to be prone to smoke as a form of self-medication, even in the absence of a real alleviating effect. However, there is controversy regarding this theory, justifying further investigation of the self-medication hypothesis.

The reasons for the comorbidity of ADHD and obesity has also been a popular target for research (Cortese, 2019; Cortese et al., 2016; Cortese & Tessari, 2017). Due to the high cost of care of both conditions (Libutzki et al., 2019), the causes of the comorbidity is of great public health importance. The results obtained in this thesis are consistent with a bidirectional and intergenerational causal effects, however more sophisticated designs for parental effects are needed to further validate the results.

Both smoking and obesity are associated with systemic chronic inflammation that also associates with a large range of other diseases, as well as psychiatric outcomes (Furman et al., 2019; Yuan et al., 2019). Here, I have conducted a large-scale MR to examine the effect of inflammatory markers on the risk of common psychiatric outcomes. The results shed light to the role of inflammation in ADHD and other common psychiatric outcomes and provide promising avenues for further research.

The key datasets used in this thesis were two pregnancy birth cohorts, NFBC1986 and NFBC1966. These datasets are valuable sources for epidemiologic research due to their longitudinal design, deep phenotyping, availability for linking with registry data and the presence of multiple omics datasets. The longitudinal design gives a great opportunity to examine long-term exposures and lifecourse disease progression. In addition, the use of publicly available databases, such as ARIES, GTEx, and GWAS summary statistics (Section 3.7) complement the information obtained from individual-level data used, and enhance the statistical power to detect molecular associations that tend to have small effect sizes.

Finally, it should be noted that ADHD is not unlike other psychiatric outcomes in that it is highly heritable with a complex genetic and molecular make-up. The analytical approaches applied in this thesis are highly transferable for examining the aetiology of other psychiatric outcomes and complex traits in general.

10.2 General limitations and future perspectives

10.2.1 Confounding and different biases

The limitations of the work presented in this thesis must be fully appreciated. A large part of this work considered the aetiology of ADHD, which requires causal inference methods for disentangling correlation and causation. In the absence of randomisation, causal inference

is always "highly speculative" (Rothman, 2008) as it has to rely on unverifiable assumptions of the data generating process.

As discussed at length in Section 4.3.1, any causal conclusions based on non-randomised data are always limited by confounding, measurement error and different biases. Unmeasured confounding can never be reliably ruled out in observational studies. In this thesis, the potential effects of prenatal exposure to cigarette smoking (Chapters 5 and 6) and obesity (Chapter 8) were examined. As parental genetic factors are likely to influence both parental and offspring phenotype, parental genotype therefore acts as a confounder. However, in NFBC1986 dataset, there are no parental genotype data available. As also acknowledged in the discussions for the corresponding chapters, this limits the causal conclusions based on these results only.

Measurement error in the measured variables is also a potential source of bias in epidemiological studies, and the resulting bias in regression models with multiple error-prone variables can be to either direction (Keogh et al., 2020). Maternal smoking was a key exposure used in this work, collected in NFBC1986 via questionnaires (Chapter 3). Self-reported smoking is liable to measurement error due to recall bias or social or medical disapproval (Rebagliato, 2002). Validation of smoking status via biochemical measurements, such as cotinine, was not available. Nevertheless, a study by Vartiainen et al. (2002) on a Finnish population showed self-reported smoking to be a valid measure.

A characteristic feature in many biological measurements, such as metabolic measures, is their temporal fluctuation around an underlying 'usual' level (W. B. Dunn et al., 2011). Such inherent error of a biological measurement is therefore highly common in molecular epidemiology, albeit with a notable exception of genetic epidemiology as the genomic sequence is highly stable.

For selection bias, it is known that cohort studies are prone to the 'healthy volunteer' bias, that is, the cohort participants tend to be more healthy than the general population (Delgado-Rodriguez & Llorca, 2004; Fry et al., 2017). In particular, a study in NFBC1966 by Haapea et al. (2008) showed evidence that subjects with a psychiatric disorder are less likely to participate in the data collections. Any such biases may have a proportionally large effect on small effect sizes, which are particularly common in molecular epidemiology (Munafò et al., 2017). Methods to adjust for non-participation or other missing data, such as inverse probability weighting, imputation or propensity score methods (Perkins et al.,

2017; Rosenbaum & Rubin, 1983), were not applied in this thesis.

The use of instrumental variable methods attempt to circumvent the problem of confounding by using proxies for exposure that fulfill instrumental variable assumptions (Greenland, 2000; Hernán & Robins, 2019). Instrumental variable methods are also more robust to measurement errors in the exposure (Greenland, 2000). The use of genetic variants as such instrumental variables in the MR paradigm (Section 4.3.3) is increasingly popular in epidemiology (Burgess & Davey Smith, 2019).

However, causal inference based on MR is limited by other issues, such as potential weak instrument bias and horizontal pleiotropy, detailed in Section 4.3.3.3. Multiple different MR approaches that try to address the challenge of horizontal pleiotropy have been developed and were applied in this work (Section 4.3.3.4) (Slob & Burgess, 2020). However, even these methods are vulnerable to population-level gene–environment correlation of environmental confounders, where the genetic variants used as instrumental variables are correlated with confounders of the exposure–outcome association, and leading to bias in MR (Koellinger & de Vlaming, 2019).

The results with no strong evidence for associations for the mediating effect of DNA methylation on ADHD symptoms (Chapter 6) and the lack of strong predictive performance of multi-omics data (Chapter 7) might reflect low statistical power to detect meaningful effect sizes. It is important to incorporate further evidence from other datasets for more robust results.

Finally, all results obtained in this thesis are based on populations of European ancestries, and the results might not be applicable to other ethnic groups. In general, the large majority of genetic and other omics datasets are based on European ancestry populations, and the need for data from diverse populations for a better understanding of the aetiology of complex traits is recognised (Sirugo et al., 2019).

10.2.2 Data sources

The sample sizes available in genetic and other molecular epidemiology continue to increase. At the same time, the potential biases in selecting study participants should also be acknowledged and carefully addressed. While all possible selection biases cannot be completely eliminated, the aim should be to minimise them and to apply appropriate caution

to the conclusion that can be made.

The importance of family-based cohorts is also acknowledged. Genotyped parent-offspring trios would enable more accurate decomposition of familial effects. It is also worth noting that MR was originally proposed as a within-family design (Davey Smith & Ebrahim, 2003), and within-family MR methods are shown to be robust to biases due to dynastic effects and assortative mating (Section 4.3.1.2) (Brumpton et al., 2020; Davies et al., 2019).

An important aspect that is of equal importance with large sample sizes is the accurate measurement of the phenotypes. The measurement of behavioural traits such as ADHD is challenging and subject to heterogeneity due to different factors, such as rater effects and longitudinal changes (Brikell et al., 2015; Sibley et al., 2017). Therefore, longitudinal follow-ups with data obtained from multiple time points, potentially from different sources if possible (e.g. different raters, self-reports and registry data), enhance more accurate quantification of the phenome.

10.2.3 Biological knowledge

GWAS and other omics-wide association studies have given invaluable insight to the complex trait aetiology. GWAS have considerably improved upon the candidate gene studies (Section 2.1) which reported many false positive findings due to lack of statistical power (Duncan et al., 2019). Using genetic variants detected in GWAS for further analyses can be considered a 'second-generation' of genetic candidates that can be used to examine specific causal pathways in complex trait development and progression (Duncan et al., 2019; Schmidt et al., 2020). In particular, QTLs from a gene with a well-understood biological function are ideal 'candidates' to proxy a relevant molecular exposure in MR context (Section 4.3.3.1), as applied here in Chapter 9.

At the same time, unexpected phenomena such as the missing heritability (Section 4.2.1.1) (Maher, 2008) have also demonstrated the gaps in our biological knowledge of complex trait genetics. Based on the observation that SNPs contributing to the heritability of complex traits are spread across the genome, Boyle et al. (2017) proposed an *omnigenic model*, suggesting that most of the heritability of complex traits originates from indirect effects of multiple genes. In the omnigenic model, genes are divided to a limited number of core genes which, via complex and interconnected cell regulatory networks, affect gene expression in

a larger set of downstream peripheral genes. A large part of the observed SNP heritability are due to variants in these peripheral genes which are not necessarily near genes with trait-specific functions, but downstream from the core genes in gene regulatory pathways. Therefore, improving the biological knowledge of gene–gene interplays and efforts to disentangle these regulatory pathways are of great importance for a better understanding of complex trait aetiology.

10.2.4 Integrative statistical modelling

In the current surplus of omics data, the development of statistical techniques and integrative omics methods for detecting patterns that mirror true biological mechanisms across omics datasets is of great importance. Eventually, the novel statistical methodology will lead to an improved discovery and understanding of the underlying biology of complex traits.

In addition, as the optimal integrative modelling approach depends on the underlying biology, there is a synergist association between method development and biological knowledge, where advanced methods enhance biological discoveries, and the biological knowledge leads to the optimisation of adequate methods. Challenges to integrative modelling include data harmonisation and normalisation across the omics datasets and the scalability of the applied methods to large datasets (Ritchie et al., 2015).

10.3 Conclusion

The work presented here applies different modelling approaches and strategies to analyse multi-omics data in relation to the aetiology of ADHD. Innovative statistical methods were applied for integrative data analysis, and the findings add to the literature by improving the knowledge on not only the aetiology of ADHD, but also on maternal smoking related DNA methylation, and the inflammatory component of other common psychiatric outcomes. All statistical methods used here are highly applicable to other psychiatric outcomes or complex traits in general.

Overall, all science and the biological knowledge is highly dependent on the quality of the data, adequate statistical approaches, and combining results from multiple sources.

Molecular epidemiology will continue to progress through collection of high-quality data, application of integrative statistical methods and rigorous triangulation of the results.

Appendix A

Appendix

Table A.1: Metabolic measures quantified by NMR metabolomics panel.

Abbreviation	Metabolic measure
XXL_VLDL_P	Concentration of chylomicrons and extremely large VLDL particles
XXL_VLDL_L	Total lipids in chylomicrons and extremely large VLDL
XXL_VLDL_PL	Phospholipids in chylomicrons and extremely large VLDL
XXL_VLDL_C	Total cholesterol in chylomicrons and extremely large VLDL
XXL_VLDL_CE	Cholesterol esters in chylomicrons and extremely large VLDL
XXL_VLDL_FC	Free cholesterol in chylomicrons and extremely large VLDL
XXL_VLDL_TG	Triglycerides in chylomicrons and extremely large VLDL
XL_VLDL_P	Concentration of very large VLDL particles
XL_VLDL_L	Total lipids in very large VLDL
XL_VLDL_PL	Phospholipids in very large VLDL
XL_VLDL_C	Total cholesterol in very large VLDL
XL_VLDL_CE	Cholesterol esters in very large VLDL
XL_VLDL_FC	Free cholesterol in very large VLDL
XL_VLDL_TG	Triglycerides in very large VLDL
L_VLDL_P	Concentration of large VLDL particles
L_VLDL_L	Total lipids in large VLDL
L_VLDL_PL	Phospholipids in large VLDL
L_VLDL_C	Total cholesterol in large VLDL
L_VLDL_CE	Cholesterol esters in large VLDL
L_VLDL_FC	Free cholesterol in large VLDL
L_VLDL_TG	Triglycerides in large VLDL
M_VLDL_P	Concentration of medium VLDL particles
M_VLDL_L	Total lipids in medium VLDL
M_VLDL_PL	Phospholipids in medium VLDL
M_VLDL_C	Total cholesterol in medium VLDL
M_VLDL_CE	Cholesterol esters in medium VLDL
M_VLDL_FC	Free cholesterol in medium VLDL
M_VLDL_TG	Triglycerides in medium VLDL
S_VLDL_P	Concentration of small VLDL particles

Table A.1: Metabolic measures quantified by NMR metabolomics panel. *(continued)*

Abbreviation	Metabolic measure
S_VLDL_L	Total lipids in small VLDL
S_VLDL_PL	Phospholipids in small VLDL
S_VLDL_C	Total cholesterol in small VLDL
S_VLDL_CE	Cholesterol esters in small VLDL
S_VLDL_FC	Free cholesterol in small VLDL
S_VLDL_TG	Triglycerides in small VLDL
XS_VLDL_P	Concentration of very small VLDL particles
XS_VLDL_L	Total lipids in very small VLDL
XS_VLDL_PL	Phospholipids in very small VLDL
XS_VLDL_C	Total cholesterol in very small VLDL
XS_VLDL_CE	Cholesterol esters in very small VLDL
XS_VLDL_FC	Free cholesterol in very small VLDL
XS_VLDL_TG	Triglycerides in very small VLDL
IDL_P	Concentration of IDL particles
IDL_L	Total lipids in IDL
IDL_PL	Phospholipids in IDL
IDL_C	Total cholesterol in IDL
IDL_CE	Cholesterol esters in IDL
IDL_FC	Free cholesterol in IDL
IDL_TG	Triglycerides in IDL
L_LDL_P	Concentration of large LDL particles
L_LDL_L	Total lipids in large LDL
L_LDL_PL	Phospholipids in large LDL
L_LDL_C	Total cholesterol in large LDL
L_LDL_CE	Cholesterol esters in large LDL
L_LDL_FC	Free cholesterol in large LDL
L_LDL_TG	Triglycerides in large LDL
M_LDL_P	Concentration of medium LDL particles
M_LDL_L	Total lipids in medium LDL
M_LDL_PL	Phospholipids in medium LDL
M_LDL_C	Total cholesterol in medium LDL
M_LDL_CE	Cholesterol esters in medium LDL
M_LDL_FC	Free cholesterol in medium LDL
M_LDL_TG	Triglycerides in medium LDL
S_LDL_P	Concentration of small LDL particles
S_LDL_L	Total lipids in small LDL
S_LDL_PL	Phospholipids in small LDL
S_LDL_C	Total cholesterol in small LDL
S_LDL_CE	Cholesterol esters in small LDL
S_LDL_FC	Free cholesterol in small LDL
S_LDL_TG	Triglycerides in small LDL
XL_HDL_P	Concentration of very large HDL particles
XL_HDL_L	Total lipids in very large HDL
XL_HDL_PL	Phospholipids in very large HDL
XL_HDL_C	Total cholesterol in very large HDL
XL_HDL_CE	Cholesterol esters in very large HDL
XL_HDL_FC	Free cholesterol in very large HDL

Table A.1: Metabolic measures quantified by NMR metabolomics panel. *(continued)*

Abbreviation	Metabolic measure
XL_HDL_TG	Triglycerides in very large HDL
L_HDL_P	Concentration of large HDL particles
L_HDL_L	Total lipids in large HDL
L_HDL_PL	Phospholipids in large HDL
L_HDL_C	Total cholesterol in large HDL
L_HDL_CE	Cholesterol esters in large HDL
L_HDL_FC	Free cholesterol in large HDL
L_HDL_TG	Triglycerides in large HDL
M_HDL_P	Concentration of medium HDL particles
M_HDL_L	Total lipids in medium HDL
M_HDL_PL	Phospholipids in medium HDL
M_HDL_C	Total cholesterol in medium HDL
M_HDL_CE	Cholesterol esters in medium HDL
M_HDL_FC	Free cholesterol in medium HDL
M_HDL_TG	Triglycerides in medium HDL
S_HDL_P	Concentration of small HDL particles
S_HDL_L	Total lipids in small HDL
S_HDL_PL	Phospholipids in small HDL
S_HDL_C	Total cholesterol in small HDL
S_HDL_CE	Cholesterol esters in small HDL
S_HDL_FC	Free cholesterol in small HDL
S_HDL_TG	Triglycerides in small HDL
XXL_VLDL_PL_pct	Phospholipids to total lipds ratio in chylomicrons and extremely large VLDL
XXL_VLDL_C_pct	Total cholesterol to total lipids ratio in chylomicrons and extremely large VLDL
XXL_VLDL_CE_pct	Cholesterol esters to total lipids ratio in chylomicrons and extremely large VLDL
XXL_VLDL_FC_pct	Free cholesterol to total lipids ratio in chylomicrons and extremely large VLDL
XXL_VLDL_TG_pct	Triglycerides to total lipids ratio in chylomicrons and extremely large VLDL
XL_VLDL_PL_pct	Phospholipids to total lipds ratio in very large VLDL
XL_VLDL_C_pct	Total cholesterol to total lipids ratio in very large VLDL
XL_VLDL_CE_pct	Cholesterol esters to total lipids ratio in very large VLDL
XL_VLDL_FC_pct	Free cholesterol to total lipids ratio in very large VLDL
XL_VLDL_TG_pct	Triglycerides to total lipids ratio in very large VLDL
L_VLDL_PL_pct	Phospholipids to total lipds ratio in large VLDL
L_VLDL_C_pct	Total cholesterol to total lipids ratio in large VLDL
L_VLDL_CE_pct	Cholesterol esters to total lipids ratio in large VLDL
L_VLDL_FC_pct	Free cholesterol to total lipids ratio in large VLDL
L_VLDL_TG_pct	Triglycerides to total lipids ratio in large VLDL
M_VLDL_PL_pct	Phospholipids to total lipds ratio in medium VLDL
M_VLDL_C_pct	Total cholesterol to total lipids ratio in medium VLDL
M_VLDL_CE_pct	Cholesterol esters to total lipids ratio in medium VLDL
M_VLDL_FC_pct	Free cholesterol to total lipids ratio in medium VLDL
M_VLDL_TG_pct	Triglycerides to total lipids ratio in medium VLDL
S_VLDL_PL_pct	Phospholipids to total lipds ratio in small VLDL
S_VLDL_C_pct	Total cholesterol to total lipids ratio in small VLDL
S_VLDL_CE_pct	Cholesterol esters to total lipids ratio in small VLDL
S_VLDL_FC_pct	Free cholesterol to total lipids ratio in small VLDL
S_VLDL_TG_pct	Triglycerides to total lipids ratio in small VLDL

Table A.1: Metabolic measures quantified by NMR metabolomics panel. *(continued)*

Abbreviation	Metabolic measure
XS_VLDL_PL_pct	Phospholipids to total lipds ratio in very small VLDL
XS_VLDL_C_pct	Total cholesterol to total lipids ratio in very small VLDL
XS_VLDL_CE_pct	Cholesterol esters to total lipids ratio in very small VLDL
XS_VLDL_FC_pct	Free cholesterol to total lipids ratio in very small VLDL
XS_VLDL_TG_pct	Triglycerides to total lipids ratio in very small VLDL
IDL_PL_pct	Phospholipids to total lipds ratio in IDL
IDL_C_pct	Total cholesterol to total lipids ratio in IDL
IDL_CE_pct	Cholesterol esters to total lipids ratio in IDL
IDL_FC_pct	Free cholesterol to total lipids ratio in IDL
IDL_TG_pct	Triglycerides to total lipids ratio in IDL
L_LDL_PL_pct	Phospholipids to total lipds ratio in large LDL
L_LDL_C_pct	Total cholesterol to total lipids ratio in large LDL
L_LDL_CE_pct	Cholesterol esters to total lipids ratio in large LDL
L_LDL_FC_pct	Free cholesterol to total lipids ratio in large LDL
L_LDL_TG_pct	Triglycerides to total lipids ratio in large LDL
M_LDL_PL_pct	Phospholipids to total lipds ratio in medium LDL
M_LDL_C_pct	Total cholesterol to total lipids ratio in medium LDL
M_LDL_CE_pct	Cholesterol esters to total lipids ratio in medium LDL
M_LDL_FC_pct	Free cholesterol to total lipids ratio in medium LDL
M_LDL_TG_pct	Triglycerides to total lipids ratio in medium LDL
S_LDL_PL_pct	Phospholipids to total lipds ratio in small LDL
S_LDL_C_pct	Total cholesterol to total lipids ratio in small LDL
S_LDL_CE_pct	Cholesterol esters to total lipids ratio in small LDL
S_LDL_FC_pct	Free cholesterol to total lipids ratio in small LDL
S_LDL_TG_pct	Triglycerides to total lipids ratio in small LDL
XL_HDL_PL_pct	Phospholipids to total lipds ratio in very large HDL
XL_HDL_C_pct	Total cholesterol to total lipids ratio in very large HDL
XL_HDL_CE_pct	Cholesterol esters to total lipids ratio in very large HDL
XL_HDL_FC_pct	Free cholesterol to total lipids ratio in very large HDL
XL_HDL_TG_pct	Triglycerides to total lipids ratio in very large HDL
L_HDL_PL_pct	Phospholipids to total lipds ratio in large HDL
L_HDL_C_pct	Total cholesterol to total lipids ratio in large HDL
L_HDL_CE_pct	Cholesterol esters to total lipids ratio in large HDL
L_HDL_FC_pct	Free cholesterol to total lipids ratio in large HDL
L_HDL_TG_pct	Triglycerides to total lipids ratio in large HDL
M_HDL_PL_pct	Phospholipids to total lipds ratio in medium HDL
M_HDL_C_pct	Total cholesterol to total lipids ratio in medium HDL
M_HDL_CE_pct	Cholesterol esters to total lipids ratio in medium HDL
M_HDL_FC_pct	Free cholesterol to total lipids ratio in medium HDL
M_HDL_TG_pct	Triglycerides to total lipids ratio in medium HDL
S_HDL_PL_pct	Phospholipids to total lipds ratio in small HDL
S_HDL_C_pct	Total cholesterol to total lipids ratio in small HDL
S_HDL_CE_pct	Cholesterol esters to total lipids ratio in small HDL
S_HDL_FC_pct	Free cholesterol to total lipids ratio in small HDL
S_HDL_TG_pct	Triglycerides to total lipids ratio in small HDL
VLDL_D	Mean diameter for VLDL particles
LDL_D	Mean diameter for LDL particles

Table A.1: Metabolic measures quantified by NMR metabolomics panel. *(continued)*

Abbreviation	Metabolic measure
HDL_D	Mean diameter for HDL particles
Serum_C	Serum total cholesterol
VLDL_C	Total cholesterol in VLDL
Remnant_C	Remnant cholesterol (non-HDL, non-LDL -cholesterol)
LDL_C	Total cholesterol in LDL
HDL_C	Total cholesterol in HDL
HDL2_C	Total cholesterol in HDL2
HDL3_C	Total cholesterol in HDL3
EstC	Esterified cholesterol
FreeC	Free cholesterol
Serum_TG	Serum total triglycerides
VLDL_TG	Triglycerides in VLDL
LDL_TG	Triglycerides in LDL
HDL_TG	Triglycerides in HDL
TotPG	Total phosphoglycerides
TGtoPG	Ratio of triglycerides to phosphoglycerides
PC	Phosphatidylcholine and other cholines
SM	Sphingomyelins
TotCho	Total cholines
ApoA1	Apolipoprotein A-I
ApoB	Apolipoprotein B
ApoBtoApoA1	Ratio of apolipoprotein B to apolipoprotein A-I
TotFA	Total fatty acids
UnSat	Estimated degree of unsaturation
DHA	22:6, docosahexaenoic acid
LA	18:2, linoleic acid
FAw3	Omega-3 fatty acids
FAw6	Omega-6 fatty acids
PUFA	Polyunsaturated fatty acids
MUFA	Monounsaturated fatty acids; 16:1, 18:1
SFA	Saturated fatty acids
DHAtoFA	Ratio of 22:6 docosahexaenoic acid to total fatty acids
LAtoFA	Ratio of 18:2 linoleic acid to total fatty acids
FAw3toFA	Ratio of omega-3 fatty acids to total fatty acids
FAw6toFA	Ratio of omega-6 fatty acids to total fatty acids
PUFAtoFA	Ratio of polyunsaturated fatty acids to total fatty acids
MUFAtoFA	Ratio of monounsaturated fatty acids to total fatty acids
SFAtoFA	Ratio of saturated fatty acids to total fatty acids
Glc	Glucose
Lac	Lactate
Pyr	Pyruvate
Cit	Citrate
Glol	Glycerol
Ala	Alanine
Gln	Glutamine
Gly	Glycine
His	Histidine

Table A.1: Metabolic measures quantified by NMR metabolomics panel. *(continued)*

Abbreviation	Metabolic measure
Ile	Isoleucine
Leu	Leucine
Val	Valine
Phe	Phenylalanine
Tyr	Tyrosine
Ace	Acetate
AcAce	Acetoacetate
bOHBut	3-hydroxybutyrate
Crea	Creatinine
Alb	Albumin
Gp	Glycoprotein acetyls, mainly a1-acid glycoprotein

VLDL: very low-density lipoprotein; IDL: intermediate-density lipoprotein; LDL: low-density lipoprotein; HDL: high-density lipoprotein

Table A.2: Descriptive statistics for NFBC1966 in Study I.

	Unexposed, N=622 (87%)	Exposed, N=95 (13%)
Males	273 (44)	41 (43)
Females	349 (56)	54 (57)
BMI in kg/m2	24.5 (4.0)	25.0 (4.6)
Current smoker	229 (37)	35 (37)
Own SES at 31y		
Upper white-collar	160 (26)	19 (20)
Lower white-collar	218 (35)	30 (32)
Blue-collar	149 (24)	29 (31)
Farmer	11 (2)	1 (1)
Others	79 (13)	14 (15)
Maternal age in years	27.7 (6.5)	25.3 (6.9)
Pre-pregnancy BMI in kg/m2	23.4 (3.3)	22.2 (3.0)
Parental SES at birth		
I (highest)	32 (5)	2 (2)
II	111 (18)	18 (19)
III	250 (40)	34 (37)
IV (lowest)	100 (16)	23 (25)

Table A.3: Descriptive statistics for NFBC1986 in Study I.

	Unexposed, N=365 (83%)	Exposed, N=77 (17%)
Males	165 (45)	32 (42)
Females	200 (55)	45 (58)
BMI in kg/m2	21.3 (3.4)	22.1 (4.1)
Current smoker	83 (21)	33 (38)
Parental SES at 16y		
Professionals	293 (80)	48 (62)
Skilled workers	52 (14)	29 (38)
Unskilled workers	3 (1)	0 (0)
Farmers	17 (5)	0 (0)
Maternal age in years	28.3 (5.3)	26.3 (5.3)
Pre-pregnancy BMI in kg/m2	22.4 (3.3)	22.0 (3.0)
Parental SES at birth		
Professionals	109 (30)	11 (14)
Skilled workers	154 (43)	34 (45)
Unskilled workers	82 (23)	31 (41)
Farmers	13 (4)	0 (0)

Table A.4: Descriptive statistics for demographic variables in Study II.

	N (%) or Median (IQR)
Maternal smoking	76 (18)
Paternal smoking	139 (32)
Sex	
Male	193 (45)
Female	239 (55)
Never-smoker	321 (74)
Family socioeconomic status	
Professional	117 (27)
Skilled worker	190 (44)
Unskilled worker	112 (26)
Farmer	13 (3)
Maternal age	28 (24 - 31)

N = sample size; IQR = interquartile range.

Table A.5: Descriptive statistics for the observational data used in Study IV.

	N (%) or Median (IQR)
Sex	
Male	1448 (49)
Female	1536 (51)
Maternal smoking	
No	2381 (80)
Yes	603 (20)
Maternal education	
Basic	755 (25)
Upper secondary	1982 (66)
Tertiary	247 (8)
Maternal BMI	21.7 (20.1 - 23.8)
Parity	1 (0 - 2)
Mother's age at delivery	27.0 (24.0 - 31.0)

N = sample size; IQR = interquarile range.

References

Abe, Y., El-Masri, B., Kimball, K. T., Pownall, H., Reilly, C. F., Osmundsen, K., … Ballantyne, C. M. (1998). Soluble cell adhesion molecules in hypertriglyceridemia and potential significance on monocyte adhesion. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *18*(5), 723–731. doi: 10.1161/01.ATV.18.5.723

Aberg, K. A., Xie, L. Y., McClay, J. L., Nerella, S., Vunck, S., Snider, S., … van den Oord, E. J. (2013). Testing two models describing how methylome-wide studies in blood are informative for psychiatric conditions. *Epigenomics*, *5*(4), 367–377. doi: 10.2217/epi.13.36

Ahola-Olli, A. V., Würtz, P., Havulinna, A. S., Aalto, K., Pitkänen, N., Lehtimäki, T., … Raitakari, O. T. (2017). Genome-wide association study identifies 27 loci influencing concentrations of circulating cytokines and growth factors. *The American Journal of Human Genetics*, *100*(1), 40–50. doi: 10.1016/j.ajhg.2016.11.007

Alkam, T., Mamiya, T., Kimura, N., Yoshida, A., Kihara, D., Tsunoda, Y., … Nabeshima, T. (2017). Prenatal nicotine exposure decreases the release of dopamine in the medial frontal cortex and induces atomoxetine-responsive neurobehavioral deficits in mice. *Psychopharmacology*, *234*(12), 1853–1869. doi: 10.1007/s00213-017-4591-z

Andersen, C. H., Thomsen, P. H., Nohr, E. A., & Lemcke, S. (2018). Maternal body mass index before pregnancy as a risk factor for ADHD and autism in children. *European Child & Adolescent Psychiatry*, *27*(2), 139–148. doi: 10.1007/s00787-017-1027-6

Anttila, V., Bulik-Sullivan, B., Finucane, H. K., Walters, R. K., Bras, J., Duncan, L., … Neale, B. M. (2018). Analysis of shared heritability in common disorders of the brain. *Science*, *360*(6395). doi: 10.1126/science.aap8757

Arnold, K. F., Davies, V., de Kamps, M., Tennant, P. W. G., Mbotwa, J., & Gilthorpe, M. S. (2020). Reflections on modern methods: generalized linear models for prognosis and intervention—theory, practice and implications for machine learning. *International Journal of Epidemiology*. doi: 10.1093/ije/dyaa049

Arshad, S. H., Holloway, J. W., Karmaus, W., Zhang, H., Ewart, S., Mansfield, L., … Kurukulaaratchy, R. (2018). Cohort Profile: The Isle Of Wight Whole Population Birth Cohort (IOWBC). *International Journal of Epidemiology*, *47*(4), 1043–1044i. doi: 10.1093/ije/dyy023

Balázs, J., & Keresztény, A. (2014). Subthreshold attention deficit hyperactivity in children and adolescents: a systematic review. *European child & adolescent psychiatry*, *23*(6), 393–408. doi: 10.1007/s00787-013-0514-7

Banaschewski, T., Belsham, B., Bloch, M. H., Ferrin, M., Johnson, M., Kustow, J., … Zuddas, A. (2018). Supplementation with polyunsaturated fatty acids (PUFAs) in the management of attention deficit hyperactivity disorder (ADHD). *Nutrition and Health*, *24*(4), 279–284. doi: 10.1177/0260106018772170

Barker, E. D., Walton, E., & Cecil, C. A. (2018). Annual research review: DNA methylation as a mediator in the association between risk exposure and child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry*, *59*(4), 303–322. doi: 10.1111/jcpp.12782

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bernstein, B. E., Meissner, A., & Lander, E. S. (2007). The mammalian epigenome. *Cell*, *128*(4), 669–681. doi: 10.1016/j.cell.2007.01.033

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., … Shen, R. (2011). High density DNA methylation array with single CpG site resolution. *Genomics*, *98*(4), 288–295. doi: 10.1016/j.ygeno.2011.07.007

Biederman, J., Mick, E., & Faraone, S. V. (2000). Age-dependent decline of symptoms of attention deficit hyperactivity disorder: Impact of remission definition and symptom type. *American Journal of Psychiatry*, *157*(5), 816–818. doi: 10.1176/appi.ajp.157.5.816

Bojesen, S. E., Timpson, N., Relton, C., Davey Smith, G., & Nordestgaard, B. G. (2017). AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. *Thorax*, *72*(7), 646–653. doi: 10.1136/thoraxjnl-2016-208789

Borodulin, K., Tolonen, H., Jousilahti, P., Jula, A., Juolevi, A., Koskinen, S., … Vartiainen, E. (2017). Cohort Profile: The National FINRISK Study. *International Journal of Epidemiology*, *47*(3), 696–696i. doi: 10.1093/ije/dyx239

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, *44*(2), 512–525. doi: 10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., & Burgess, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, *40*(4), 304–314. doi: 10.1002/gepi.21965

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., … Davey Smith, G. (2012). Cohort Profile: The 'Children of the 90s'–the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology*, *42*(1), 111–127. doi: 10.1093/ije/dys064

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell*, *169*(7), 1177–1186. doi: 10.1016/j.cell.2017.05.038

Brikell, I., Kuja-Halkola, R., & Larsson, H. (2015). Heritability of attention-deficit hyperactivity disorder in adults. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *168*(6), 406–413. doi: 10.1002/ajmg.b.32335

Brumpton, B., Sanderson, E., Heilbron, K., Hartwig, F. P., Harrison, S., Vie, G. ., … Davies, N. M. (2020). Avoiding dynastic, assortative mating, and population stratification biases in mendelian randomization through within-family analyses. *Nature Communications*, *11*(1), 1–13. doi: 10.1038/s41467-020-17117-4

Buniello, A., MacArthur, J. A., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., … Parkinson, H. (2018). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. doi: 10.1093/nar/gky1120

Burgess, S., Butterworth, A., & Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, *37*(7), 658–665. doi: 10.1002/gepi.21758

Burgess, S., & Davey Smith, G. (2019). How humans can contribute to Mendelian randomization analyses. *International Journal of Epidemiology*, *48*(3), 661–664. doi: 10.1093/ije/dyz152

Burgess, S., Davey Smith, G., Davies, N. M., Dudbridge, F., Gill, D., Glymour, M. M., … Theodoratou, E. (2020). Guidelines for performing Mendelian randomization investigations [version 2; peer review: 2 approved]. *Wellcome Open Research*, *4*(186). doi: 10.12688/wellcomeopenres.15555.2

Burgess, S., Dudbridge, F., & Thompson, S. G. (2016). Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statistics in Medicine*, *35*(11), 1880–1906. doi: 10.1002/sim.6835

Burgess, S., & Labrecque, J. A. (2018). Mendelian randomization with a binary exposure variable: interpretation and presentation of causal estimates. *European Journal of Epidemiology*, *33*(10), 947–952. doi: 10.1007/s10654-018-0424-6

Burgess, S., Scott, R. A., Timpson, N. J., Smith, G. D., Thompson, S. G., & the EPIC-InterAct Consortium. (2015). Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *European journal of epidemiology*, *30*(7), 543–552. doi: 10.1007/s10654-015-0011-z

Burgess, S., & Thompson, S. G. (2011). Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology*, *40*(3), 755–764. doi: 10.1093/ije/dyr036

Burgess, S., & Thompson, S. G. (2013). Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology*, *42*(4), 1134–1144. doi: 10.1093/ije/dyt093

Burgess, S., & Thompson, S. G. (2015). *Mendelian randomization: methods for using genetic variants in causal estimation*. Chapman and Hall/CRC.

Burgess, S., & Thompson, S. G. (2017). Interpreting findings from Mendelian randomization using the MR-Egger method. *European journal of epidemiology*, *32*(5), 377–389. doi: 10.1007/s10654-017-0255-x

Cardon, L. R., & Palmer, L. J. (2003). Population stratification and spurious allelic association. *The Lancet*, *361*(9357), 598–604. doi: 10.1016/S0140-6736(03)12520-2

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1). doi: 10.1186/s13742-015-0047-8

Chang, J. P.-C., Su, K.-P., Mondelli, V., & Pariante, C. M. (2018). Omega-3 polyunsaturated fatty acids in youths with attention deficit hyperactivity disorder: a systematic review and meta-analysis of clinical trials and biological studies. *Neuropsychopharmacology*, *43*(3), 534–545. doi: 10.1038/npp.2017.160

Chen, K., Huang, J., Gong, W., Zhang, L., Yu, P., & Wang, J. M. (2006). Cd40/cd40l dyad in the inflammatory and immune responses in the central nervous system. *Cell Mol Immunol*, *3*(3), 163–169. (`https://pubmed.ncbi.nlm.nih.gov/16893496/`)

Chen, Q., Sjölander, A., Långström, N., Rodriguez, A., Serlachius, E., D'Onofrio, B. M., … Larsson, H. (2013). Maternal pre-pregnancy body mass index and offspring attention deficit hyperactivity disorder: a population-based cohort study using a sibling-comparison design. *International Journal of Epidemiology*, *43*(1), 83–90. doi: 10.1093/ije/dyt152

Chen, Y., Liang, X., Zheng, S., Wang, Y., & Lu, W. (2018). Association of body fat mass and fat distribution with the incidence of hypertension in a population-based chinese cohort: A 22-year follow-up. *Journal of the American Heart Association*, *7*(6), e007153. doi: 10.1161/JAHA.117.007153

Choi, S. W., Mak, T. S.-H., & O'Reilly, P. F. (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*, 1–14. doi: 10.1038/s41596-020-0353-1

Cole, J., Ball, H. A., Martin, N. C., Scourfield, J., & Mcguffin, P. (2009). Genetic overlap between measures of hyperactivity/inattention and mood in children and adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, *48*(11), 1094–1101. doi: 10.1097/CHI.0b013e3181b7666e

Cole, S. R., Chu, H., & Greenland, S. (2013). Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer. *American Journal of Epidemiology*, *179*(2), 252–260. doi: 10.1093/aje/kwt245

Cooper, R. E., Tye, C., Kuntsi, J., Vassos, E., & Asherson, P. (2015). Omega-3 polyunsaturated fatty acid supplementation and cognition: A systematic review and meta-analysis. *Journal of Psychopharmacology*, *29*(7), 753–763. doi: 10.1177/0269881115587958

Cortese, S. (2019). The Association between ADHD and Obesity: Intriguing, Progressively More Investigated, but Still Puzzling. *Brain Sciences*, *9*(10). doi: 10.3390/brainsci9100256

Cortese, S., Moreira-Maia, C. R., St. Fleur, D., Morcillo-Peñalver, C., Rohde, L. A., & Faraone, S. V. (2016). Association Between ADHD and Obesity: A Systematic Review and Meta-Analysis. *American Journal of Psychiatry*, *173*(1), 34–43. doi: 10.1176/appi.ajp.2015.15020266

Cortese, S., & Tessari, L. (2017). Attention-deficit/hyperactivity disorder (ADHD) and obesity: update 2016. *Current psychiatry reports*, *19*(1), 4. doi: 10.1007/s11920-017-0754-1

Cortese, S., & Vincenzi, B. (2012). Obesity and ADHD: Clinical and Neurobiological Implications. In C. Stanford & R. Tannock (Eds.), *Behavioral Neuroscience of Attention Deficit Hyperactivity Disorder and Its Treatment* (pp. 199–218). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/7854_2011_154

Cupul-Uicab, L. A., Skjaerven, R., Haug, K., Melve, K. K., Engel, S. M., & Longnecker, M. P. (2012). In Utero Exposure to Maternal Tobacco Smoke and Subsequent Obesity, Hypertension, and Gestational Diabetes Among Women in The MoBa Cohort. *Environmental Health Perspectives*, *120*(3), 355–360. doi: 10.1289/ehp.1103789

Dalsgaard, S., Østergaard, S. D., Leckman, J. F., Mortensen, P. B., & Pedersen, M. G. (2015). Mortality in children, adolescents, and adults with attention deficit hyperactivity disorder: a nationwide cohort study. *The Lancet*, *385*(9983), 2190–2196. doi: 10.1016/S0140-6736(14)61684-6

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., … Fuchsberger, C.

(2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. doi: 10.1038/ng.3656

Davey Smith, G. (2008). Assessing intrauterine influences on offspring health outcomes: Can epidemiological studies yield robust findings? *Basic & Clinical Pharmacology & Toxicology*, *102*(2), 245–256. doi: 10.1111/j.1742-7843.2007.00191.x

Davey Smith, G., & Ebrahim, S. (2002). Data dredging, bias, or confounding. *BMJ*, *325*(7378), 1437–1438. doi: 10.1136/bmj.325.7378.1437

Davey Smith, G., & Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, *32*(1), 1–22. doi: 10.1093/ije/dyg070

Davey Smith, G., & Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet*, *23*(R1), R89–98. doi: 10.1093/hmg/ddu328

Davey Smith, G., Lipsitch, M., Tchetgen Tchetgen, E., & Cohen, T. (2012). Negative control exposures in epidemiologic studies. *Epidemiology*, *23*(2), 350–352. doi: 10.1097/EDE.0b013e318245912c

Davies, N. M., Holmes, M. V., & Davey Smith, G. (2018). Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ*, *362*. doi: 10.1136/bmj.k601

Davies, N. M., Howe, L. J., Brumpton, B., Havdahl, A., Evans, D. M., & Davey Smith, G. (2019). Within family Mendelian randomization studies. *Human Molecular Genetics*, *28*(R2), R170–R179. doi: 10.1093/hmg/ddz204

Dawn Teare, M., & Barrett, J. H. (2005). Genetic linkage studies. *The Lancet*, *366*(9490), 1036–1044. doi: 10.1016/S0140-6736(05)67382-5

Day, F., Loh, P.-R., Scott, R., Ong, K., & Perry, J. (2016). A robust example of collider bias in a genetic association study. *The American Journal of Human Genetics*, *98*(2), 392–393. doi: 10.1016/j.ajhg.2015.12.019

De Biasi, M., & Dani, J. A. (2011). Reward, addiction, withdrawal to nicotine. *Annual Review of Neuroscience*, *34*(1), 105–130. doi: 10.1146/annurev-neuro-061010-113734

Delgado-Rodriguez, M., & Llorca, J. (2004). Bias. *Journal of Epidemiology & Community Health*, *58*(8), 635–641. doi: 10.1136/jech.2003.008466

Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., … Neale, B. M. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature Genetics*, *51*(1), 63–75. doi: 10.1038/s41588-018-0269-7

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*(4), 997–1004. doi: 10.1111/j.0006-341X.1999.00997.x

Diagnostic and statistical manual of mental disorders: DSM-5. (5th. ed.). (2013). Washington, D.C.: American Psychiatric Association.

Didelez, V., & Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, *16*(4), 309–330. doi: 10.1177/0962280206077743

Doherty, S. P., Grabowski, J., Hoffman, C., Ng, S. P., & Zelikoff, J. T. (2009). Early life insult from cigarette smoke may be predictive of chronic diseases later in life. *Biomarkers*, *14*(sup1), 97–101. doi: 10.1080/13547500902965898

Dowlati, Y., Herrmann, N., Swardfager, W., Liu, H., Sham, L., Reim, E. K., & Lanctôt, K. L. (2010). A meta-analysis of cytokines in major depression. *Biological Psychiatry*, *67*(5), 446–457. doi: 10.1016/j.biopsych.2009.09.033

Draisma, H. H. M., Morris, A. P., & Maegi, R. (2019). methylSCOPAv0.2.2. Zenodo. doi: 10.5281/zenodo.2540687

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., & Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, *11*(1), 587. doi: 10.1186/1471-2105-11-587

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLOS Genetics*, *9*(3), 1–17. doi: 10.1371/journal.pgen.1003348

Dudbridge, F., & Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, *32*(3), 227–234. doi: 10.1002/gepi.20297

Duncan, L. E., Ostacher, M., & Ballon, J. (2019). How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. *Neuropsychopharmacology*, *44*(9), 1518–1523. doi: 10.1038/s41386-019-0389-5

Dunn, G. A., Nigg, J. T., & Sullivan, E. L. (2019). Neuroinflammation as a risk factor for attention deficit hyperactivity disorder. *Pharmacology Biochemistry and Behavior*, *182*, 22–34. doi: 10.1016/j.pbb.2019.05.005

Dunn, W. B., Broadhurst, D. I., Atherton, H. J., Goodacre, R., & Griffin, J. L. (2011). Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem. Soc. Rev.*, *40*, 387–426. doi: 10.1039/B906712B

Du Rietz, E., Coleman, J., Glanville, K., Choi, S. W., O'Reilly, P. F., & Kuntsi, J. (2018). Association of Polygenic Risk for Attention-Deficit/Hyperactivity Disorder With Co-occurring Traits and Disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(7), 635–643. doi: 10.1016/j.bpsc.2017.11.013

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.*, *1*(1), 54–75. doi: 10.1214/ss/1177013815

Ekblad, M., Gissler, M., Lehtonen, L., & Korkeila, J. (2010). Prenatal Smoking Exposure and the Risk of Psychiatric Morbidity Into Young Adulthood. *Archives of General Psychiatry*, *67*(8), 841–849. doi: 10.1001/archgenpsychiatry.2010.92

Erskine, H. E., Ferrari, A. J., Polanczyk, G. V., Moffitt, T. E., Murray, C. J. L., Vos, T., … Scott, J. G. (2014). The global burden of conduct disorder and attention-deficit/hyperactivity disorder in 2010. *Journal of Child Psychology and Psychiatry*, *55*(4), 328–336. doi: 10.1111/jcpp.12186

Erskine, H. E., Norman, R. E., Ferrari, A. J., Chan, G. C., Copeland, W. E., Whiteford, H. A., & Scott, J. G. (2016). Long-term outcomes of attention-deficit/hyperactivity disorder and conduct disorder: A systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, *55*(10), 841–850. doi: 10.1016/j.jaac.2016.06.016

Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2014). PRSice: Polygenic Risk Score software. *Bioinformatics*, *31*(9), 1466–1468. doi: 10.1093/bioinformatics/btu848

Fallin, M. D., Duggal, P., & Beaty, T. H. (2016). Genetic Epidemiology and Public Health: The Evolution From Theory to Technology. *American Journal of Epidemiology*, *183*(5), 387–393. doi: 10.1093/aje/kww001

Faraone, S. V., Asherson, P., Banaschewski, T., Biederman, J., Buitelaar, J. K., Ramos-Quiroga, J. A., … Franke, B. (2015). Attention-deficit/hyperactivity disorder. *Nature Reviews Disease Primers*, *1*(1), 15020. doi: 10.1038/nrdp.2015.20

Faraone, S. V., Biederman, J., & Mick, E. (2006). The age-dependent decline of attention deficit hyperactivity disorder: a meta-analysis of follow-up studies. *Psychological Medicine*, *36*(2), 159–165. doi: 10.1017/S003329170500471X

Faraone, S. V., Biederman, J., & Wozniak, J. (2012). Examining the Comorbidity Between Attention Deficit Hyperactivity Disorder and Bipolar I Disorder: A Meta-Analysis of Family Genetic Studies. *American Journal of Psychiatry*, *169*(12), 1256–1266. doi: 10.1176/appi.ajp.2012.12010087

Faraone, S. V., & Larsson, H. (2019). Genetics of attention deficit hyperactivity disorder. *Molecular psychiatry*, *24*(4), 562–575. doi: 10.1038/s41380-018-0070-0

Faraone, S. V., Perlis, R. H., Doyle, A. E., Smoller, J. W., Goralnick, J. J., Holmgren, M. A., & Sklar, P. (2005). Molecular genetics of attention-deficit/hyperactivity disorder. *Biological Psychiatry*, *57*(11), 1313–1323. doi: 10.1016/j.biopsych.2004.11.024

Fedak, K. M., Bernal, A., Capshaw, Z. A., & Gross, S. (2015). Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerging themes in epidemiology*, *12*(1), 14. doi: 10.1186/s12982-015-0037-4

Forastiere, L., Mattei, A., & Ding, P. (2018). Principal ignorability in mediation analysis: through and beyond sequential ignorability. *Biometrika*, *105*(4), 979–986. doi: 10.1093/biomet/asy053

Franke, B., Faraone, S. V., Asherson, P., Buitelaar, J., Bau, C. H., Ramos-Quiroga, J. A., … Reif, A. (2012). The genetics of attention deficit/hyperactivity disorder in adults, a review. *Mol. Psychiatry*, *17*(10), 960–987. doi: 10.1038/mp.2011.138

Franke, B., Neale, B. M., & Faraone, S. V. (2009). Genome-wide association studies in ADHD. *Human genetics*, *126*(1), 13–50. doi: 10.1007/s00439-009-0663-4

Franz, A. P., Bolat, G. U., Bolat, H., Matijasevich, A., Santos, I. S., Silveira, R. C., … Moreira-Maia, C. R. (2018). Attention-deficit/hyperactivity disorder and very preterm/very low birth weight: A meta-analysis. *Pediatrics*, *141*(1). doi: 10.1542/peds.2017-1645

Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J., Davey Smith, G., … Lawlor, D. A. (2012). Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International Journal of Epidemiology*, *42*(1), 97–110. doi: 10.1093/ije/dys066

Friedman, J., Hastie, T., & Tibshirani, R. (2009). *The elements of statistical learning* (Second ed.). Springer series in statistics New York. doi: 10.1007/978-0-387-84858-7

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. doi: 10.18637/jss.v033.i01

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., … Allen, N. E. (2017). Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology*, *186*(9), 1026–1034. doi: 10.1093/aje/kwx246

Furman, D., Campisi, J., Verdin, E., Carrera-Bastos, P., Targ, S., Franceschi, C., … Slavich, G. M. (2019). Chronic inflammation in the etiology of disease across the life span. *Nature Medicine*, *25*(12), 1822–1832. doi: 10.1038/s41591-019-0675-0

Gage, S. H., Munafò, M. R., & Davey Smith, G. (2016). Causal Inference in Developmental Origins of Health and Disease (DOHaD) Research. *Annual Review of Psychology*, *67*(1), 567–585. doi: 10.1146/annurev-psych-122414-033352

Gallagher, M. D., & Chen-Plotkin, A. S. (2018). The Post-GWAS Era: From Association to Function. *The American Journal of Human Genetics*, *102*(5), 717–730. doi: 10.1016/j.ajhg.2018.04.002

Gaunt, T. R., Shihab, H. A., Hemani, G., Min, J. L., Woodward, G., Lyttleton, O., … Relton, C. L. (2016). Systematic identification of genetic influences on methylation across the human life course. *Genome Biology*, *17*(1), 61. doi: 10.1186/s13059-016-0926-z

Gelman, A. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gibney, E. R., & Nolan, C. M. (2010). Epigenetics and gene expression. *Heredity (Edinb)*, *105*(1), 4–13. doi: 10.1038/hdy.2010.54

Gill, D., Walker, V. M., Martin, R. M., Davies, N. M., & Tzoulaki, I. (2019). Comparison with randomized controlled trials as a strategy for evaluating instruments in Mendelian randomization. *International Journal of Epidemiology*. doi: 10.1093/ije/dyz236

Gizer, I. R., Ficks, C., & Waldman, I. D. (2009). Candidate gene studies of ADHD: a meta-analytic review. *Human genetics*, *126*(1), 51–90. doi: 10.1007/s00439-009-0694-x

Gkatzionis, A., & Burgess, S. (2018). Contextualizing selection bias in Mendelian randomization: how bad is it likely to be? *International Journal of Epidemiology*, *48*(3), 691–701. doi: 10.1093/ije/dyy202

Glass, T. A., Goodman, S. N., Hernán, M. A., & Samet, J. M. (2013). Causal inference in public health. *Annual Review of Public Health*, *34*(1), 61–75. doi: 10.1146/annurev-publhealth-031811-124606

Goldsmith, D., Rapaport, M., & Miller, B. (2016). A meta-analysis of blood cytokine network alterations in psychiatric patients: comparisons between schizophrenia, bipolar disorder and depression. *Molecular psychiatry*, *21*(12), 1696–1709. doi: 10.1038/mp.2016.3

Gonçalves, R. B., Coletta, R. D., Silvério, K. G., Benevides, L., Casati, M. Z., da Silva, J. S., & Nociti, F. H. (2011). Impact of smoking on inflammation: overview of molecular mechanisms. *Inflamm. Res.*, *60*(5), 409–424. doi: 10.1007/s00011-011-0308-7

Good, I. J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association*, *53*(284), 799–813. doi: 10.1080/01621459.1958.10501480

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *J Child Psychol Psychiatry*, *38*(5), 581–586. doi: 10.1111/j.1469-7610.1997.tb01545.x

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, *29*(4), 722–729. doi: 10.1093/ije/29.4.722

Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statist. Sci.*, *14*(1), 29–46. doi: 10.1214/ss/1009211805

Gregor, M. F., & Hotamisligil, G. S. (2011). Inflammatory mechanisms in obesity. *Annual Review of Immunology*, *29*(1), 415–445. doi: 10.1146/annurev-immunol-031210-101322

Grove, J., Ripke, S., Als, T. D., Mattheisen, M., Walters, R. K., Won, H., … Wilson, C. H. (2019). Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.*, *51*(3), 431–444. doi: 10.1038/s41588-019-0344-8

GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204–213. doi: 10.1038/nature24277

Guan, Y., & Stephens, M. (2008). Practical issues in imputation-based association mapping. *PLOS Genetics*, *4*(12), 1–11. doi: 10.1371/journal.pgen.1000279

Guintivano, J., & Kaminsky, Z. A. (2016). Role of epigenetic factors in the development of mental illness throughout life. *Neuroscience Research*, *102*, 56–66. (Trajectories in mental illness) doi: 10.1016/j.neures.2014.08.003

Haapea, M., Miettunen, J., Läärä, E., Joukamaa, M. I., Järvelin, M.-R., Isohanni, M. K., & Veijola, J. M. (2008). Non-participation in a field survey with respect to psychiatric disorders. *Scandinavian Journal of Public Health*, *36*(7), 728–736. doi: 10.1177/1403494808092250

Haas, R., Zelezniak, A., Iacovacci, J., Kamrad, S., Townsend, S., & Ralser, M. (2017). Designing and interpreting 'multi-omic' experiments that may change our understanding of biology. *Current Opinion in Systems Biology*, *6*, 37–45. doi: https://doi.org/10.1016/j.coisb.2017.08.009

Hamer, D. H. (2000). Beware the chopsticks gene. *Molecular psychiatry*, *5*(1), 11–13. doi: 10.1038/sj.mp.4000662

Hamza, M., Halayem, S., Bourgou, S., Daoud, M., Charfi, F., & Belhadj, A. (2019). Epigenetics and ADHD: Toward an Integrative Approach of the Disorder Pathogenesis. *J Atten Disord*, *23*(7), 655–664. doi: 10.1177/1087054717696769

Harrell, F. E. (2015). *Regression modeling strategies* (Second ed.). Springer.

Harrell, F. E., Lee, K. L., & Pollock, B. G. (1988). Regression Models in Clinical Studies: Determining Relationships Between Predictors and Response2. *JNCI: Journal of the National Cancer Institute*, *80*(15), 1198–1202. doi: 10.1093/jnci/80.15.1198

Hartwig, F. P., Davies, N. M., & Davey Smith, G. (2018). Bias in Mendelian randomization due to assortative mating. *Genetic Epidemiology*, *42*(7), 608–620. doi: 10.1002/gepi.22138

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol*, *18*(1), 83. doi: 10.1186/s13059-017-1215-1

Hawi, Z., Cummins, T. D., Tong, J., Johnson, B., Lau, R., Samarrai, W., & Bellgrove, M. A. (2015). The molecular genetic architecture of attention deficit hyperactivity disorder. *Mol. Psychiatry*, *20*(3), 289–297. doi: 10.1038/mp.2014.183

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504. doi: 10.1037/1082-989X.3.4.486

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection – A review and recommendations for the practicing statistician. *Biom J*, *60*(3), 431–449. doi: 10.1002/bimj.201700067

Hemani, G., Bowden, J., & Davey Smith, G. (2018). Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human Molecular Genetics*, *27*(R2), R195–R208. doi: 10.1093/hmg/ddy163

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., … Haycock, P. C. (2018). The MR-Base platform supports systematic causal inference across the human phenome. *eLife*, *7*, e34408. doi: 10.7554/eLife.34408

Hernán, M. A., & Robins, J. (2019). *Causal inference*. Taylor & Francis. (`https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/`)

Hernán, M. A. (2004). A definition of causal effect for epidemiological research. *J Epidemiol Community Health*, *58*(4), 265–271. doi: 10.1136/jech.2002.006361

Hernán, M. A., Hernández-Díaz, S., M., & Robins, J., M. (2004). A structural approach to selection bias. *Epidemiology*, *15*(5), 615–625. doi: 10.1097/01.ede.0000135174.63482.43

Hernán, M. A., Hsu, J., & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *CHANCE*, *32*(1), 42–49. doi: 10.1080/09332480.2019.1579578

Hernán, M. A., & Robins, J. M. (2006). Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, *17*(4), 360–372. doi: 10.1097/01.ede.0000222409.00878.37

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, *58*(5), 295–300. doi: 10.1177/003591576505800503

Hirk, R., Hornik, K., Vana, L., & Genz, A. (2019). *mvord: Multivariate Ordinal Regression Models.* (`https://CRAN.R-project.org/package=mvord`)

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, *6*(2), 95–108. doi: 10.1038/nrg1521

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. doi: 10.1080/00401706.1970.10488634

Hofhuis, W., de Jongste, J. C., & Merkus, P. J. (2003). Adverse health effects of prenatal and postnatal tobacco smoke exposure on children. *Arch. Dis. Child.*, *88*(12), 1086–1090. doi: 10.1136/adc.88.12.1086

Hope, S., Hoseth, E., Dieset, I., Mørch, R. H., Aas, M., Aukrust, P., … Andreassen, O. A. (2015). Inflammatory markers are associated with general cognitive abilities in schizophrenia and bipolar disorder patients and healthy controls. *Schizophrenia Research*, *165*(2), 188–194. doi: 10.1016/j.schres.2015.04.004

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*(3/4), 321–377. doi: 10.1093/biomet/28.3-4.321

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., … Kelsey, K. T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, *13*, 86. doi: 10.1186/1471-2105-13-86

Huang, L., Wang, Y., Zhang, L., Zheng, Z., Zhu, T., Qu, Y., & Mu, D. (2018). Maternal Smoking and Attention-Deficit/Hyperactivity Disorder in Offspring: A Meta-analysis. *Pediatrics*, *141*(1). doi: 10.1542/peds.2017-2465

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychol Methods*, *15*(4), 309–334. doi: 10.1037/a0020761

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science*, 51–71. doi: 10.1214/10-STS321

Impellizzeri, D., & Cuzzocrea, S. (2014). Targeting selectins for the treatment of inflammatory diseases. *Expert Opinion on Therapeutic Targets*, *18*(1), 55–67. doi: 10.1517/14728222.2013.841140

International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million snps. *Nature*, *449*(7164), 851. doi: 10.1038/nature06258

Iwata, Y., Suzuki, K., Nakamura, K., Matsuzaki, H., Sekine, Y., Tsuchiya, K. J., … Mori, N. (2007). Increased levels of serum soluble L-selectin in unmedicated patients with schizophrenia. *Schizophrenia Research*, *89*(1), 154–160. doi: 10.1016/j.schres.2006.08.026

Jaffe, A. E., & Irizarry, R. A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome biology*, *15*(2), 1–9. doi: 10.1186/gb-2014-15-2-r31

Jeon, S. W., Yoon, H.-k., & Kim, Y.-K. (2019). Role of inflammation in psychiatric disorders. In Y.-K. Kim (Ed.), *Frontiers in psychiatry: Artificial intelligence, precision medicine, and other paradigm shifts* (pp. 491–501). Singapore: Springer Singapore. doi: 10.1007/978-981-32-9721-0_24

Jing, F., Mao, Y., Guo, J., Zhang, Z., Li, Y., Ye, Z., … Chen, K. (2014). The value of Apolipoprotein B/Apolipoprotein A1 ratio for metabolic syndrome diagnosis in a Chinese population: a cross-sectional study. *Lipids in health and disease*, *13*(1), 81. doi: 10.1186/1476-511X-13-81

Joehanes, R., Just, A. C., Marioni, R. E., Pilling, L. C., Reynolds, L. M., Mandaviya, P. R., … London, S. J. (2016). Epigenetic Signatures of Cigarette Smoking. *Circ Cardiovasc Genet*, *9*(5), 436–447. doi: 10.1161/CIRCGENETICS.116.001506

Joubert, B. R., Felix, J. F., Yousefi, P., Bakulski, K. M., Just, A. C., Breton, C., … London, S. J. (2016). DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *Am J Hum Genet*, *98*(4), 680–96. doi: 10.1016/j.ajhg.2016.02.019

Jung, C. H., Hwang, J. Y., Yu, J. H., Shin, M. S., Bae, S. J., Park, J.-Y., … Lee, W. J. (2012). The value of apolipoprotein B/A1 ratio in the diagnosis of metabolic syndrome in a Korean population. *Clinical Endocrinology*, *77*(5), 699–706. doi: 10.1111/j.1365-2265.2012.04329.x

Järvelin, M.-R., Hartikainen-Sorri, A.-L., & Rantakallio, P. (1993). Labour induction policy in hospitals of different levels of specialisation. *Br J Obstet Gynaecol*, *100*(4), 310–5. doi: 10.1111/j.1471-0528.1993.tb12971.x

Katan, M. B. (1986). Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet*, *1*(8479), 507–508. doi: 10.1016/s0140-6736(86)92972-7

Katzman, M. A., Bilkey, T. S., Chokka, P. R., Fallu, A., & Klassen, L. J. (2017). Adult ADHD and comorbid disorders: clinical implications of a dimensional approach. *BMC Psychiatry*, *17*(1), 302. doi: 10.1186/s12888-017-1463-3

Keogh, R. H., Shaw, P. A., Gustafson, P., Carroll, R. J., Deffner, V., Dodd, K. W., … Freedman, L. S. (2020). Stratos guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—basic theory and simple methods of adjustment. *Statistics in Medicine*, *39*(16), 2197–2231. doi: 10.1002/sim.8532

Keyes, K. M., Davey Smith, G., & Susser, E. (2014). Associations of prenatal maternal smoking with offspring hyperactivity: causal or confounded? *Psychological Medicine*, *44*(4), 857–867. doi: 10.1017/S0033291713000986

Knopik, V. S., Maccani, M. A., Francazio, S., & McGeary, J. E. (2012). The epigenetics of maternal cigarette smoking during pregnancy and effects on child development. *Development and Psychopathology*, *24*(4), 1377–1390. doi: 10.1017/S0954579412000776

Koellinger, P. D., & de Vlaming, R. (2019). Mendelian randomization: the challenge of unobserved environmental confounds. *International Journal of Epidemiology*, *48*(3), 665–671. doi: 10.1093/ije/dyz138

Kong, A., Thorleifsson, G., Frigge, M. L., Vilhjalmsson, B. J., Young, A. I., Thorgeirsson, T. E., … Stefansson, K. (2018). The nature of nurture: Effects of parental genotypes. *Science*, *359*(6374), 424–428. doi: 10.1126/science.aan6877

Kotimaa, A. J., Moilanen, I., Taanila, A., Ebeling, H., Smalley, S. L., Mcgough, J. J., … Järvelin, M.-R. (2003). Maternal smoking and hyperactivity in 8-year-old children. *Journal of the American Academy of Child & Adolescent Psychiatry*, *42*(7), 826–833. doi: 10.1097/01.CHI.0000046866.56865.A2

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer New York.

Küpers, L. K., Xu, X., Jankipersadsing, S. A., Vaez, A., la Bastide-van Gemert, S., Scholtens, S., … Snieder, H. (2015). DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. *Int J Epidemiol*, *44*(4), 1224–1237. doi: 10.1093/ije/dyv048

Lahti, J., Räikkönen, K., Sovio, U., Miettunen, J., Hartikainen, A.-L., Pouta, A., … Veijola, J. (2009). Early-life origins of schizotypal traits in adulthood. *British Journal of Psychiatry*, *195*(2), 132–137. doi: 10.1192/bjp.bp.108.054387

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. doi: 10.1038/35057062

Larson, K. C., Draper, M. P., Lipko, M., & Dabrowski, M. (2010). Gng12 is a novel negative regulator of LPS-induced inflammation in the microglial cell line BV-2. *Inflammation research*, *59*(1), 15–22. doi: 10.1007/s00011-009-0062-2

Larsson, H., Anckarsater, H., Råstam, M., Chang, Z., & Lichtenstein, P. (2012). Childhood attention-deficit hyperactivity disorder as an extreme of a continuous trait: a quantitative genetic study of 8,500 twin pairs. *J Child Psychol Psychiatry*, *53*(1), 73–80. doi: 10.1111/j.1469-7610.2011.02467.x

Larsson, H., Rydén, E., Boman, M., Långström, N., Lichtenstein, P., & Landén, M. (2013). Risk of bipolar disorder and schizophrenia in relatives of people with attention-deficit hyperactivity disorder. *British Journal of Psychiatry*, *203*(2), 103–106. doi: 10.1192/bjp.bp.112.120808

Lawlor, D. A. (2016). Commentary: Two-sample Mendelian randomization: opportunities and challenges. *Int J Epidemiol*, *45*(3), 908–915. doi: 10.1093/ije/dyw127

Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., & Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med*, *27*(8), 1133–1163. doi: 10.1002/sim.3034

Lawlor, D. A., Tilling, K., & Davey Smith, G. (2017). Triangulation in aetiological epidemiology. *International Journal of Epidemiology*, *45*(6), 1866–1886. doi: 10.1093/ije/dyw314

Lawson, D. J., Davies, N. M., Haworth, S., Ashraf, B., Howe, L., Crawford, A., … Timpson, N. J. (2020). Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Hum. Genet.*, *139*(1), 23–41. doi: 10.1007/s00439-019-02014-8

Le, H. H., Hodgkins, P., Postma, M. J., Kahle, J., Sikirica, V., Setyawan, J., … Doshi, J. A. (2014). Economic impact of childhood/adolescent ADHD in a European setting: the Netherlands as a reference case. *Eur Child Adolesc Psychiatry*, *23*(7), 587–598. doi: 10.1007/s00787-013-0477-8

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., … Cesarini, D. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics*, *50*(8), 1112–1121. doi: 10.1038/s41588-018-0147-3

Lee, K. W., Richmond, R. C., Hu, P., French, L., Shin, J., Bourdon, C., … Pausova, Z. (2015). Prenatal Exposure to Maternal Cigarette Smoking and DNA Methylation: Epigenome-Wide Association in a Discovery Sample of Adolescents and Replication in an Independent Cohort at Birth through 17 Years of Age. *Environmental Health Perspectives*, *123*(2), 193–199. doi: 10.1289/ehp.1408614

Lee, P. H., Anttila, V., Won, H., Feng, Y.-C. A., Rosenthal, J., Zhu, Z., … Smoller, J. W. (2019). Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell*, *179*(7), 1469–1482.e11. doi: 10.1016/j.cell.2019.11.020

Lehne, B., Drong, A. W., Loh, M., Zhang, W., Scott, W. R., Tan, S. T., … Chambers, J. C. (2015). A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. *Genome Biol*, *16*, 37. doi: 10.1186/s13059-015-0600-x

Ley, K. (2003). The role of selectins in inflammation and disease. *Trends in Molecular Medicine*, *9*(6), 263–268. doi: 10.1016/S1471-4914(03)00071-6

Li, L., Lagerberg, T., Chang, Z., Cortese, S., Rosenqvist, M. A., Almqvist, C., … Larsson, H. (2020). Maternal pre-pregnancy overweight/obesity and the risk of attention-deficit/hyperactivity disorder in offspring: a systematic review, meta-analysis and quasi-experimental family-based study. *International Journal of Epidemiology*. doi: 10.1093/ije/dyaa040

Libutzki, B., Ludwig, S., May, M., Jacobsen, R. H., Reif, A., & Hartman, C. A. (2019). Direct medical costs of ADHD and its comorbid conditions on basis of a claims data analysis. *European Psychiatry*, *58*, 38–44. doi: 10.1016/j.eurpsy.2019.01.019

Lichtenstein, P., Carlström, E., Råstam, M., Gillberg, C., & Anckarsäter, H. (2010). The genetics of autism spectrum disorders and related neuropsychiatric disorders in childhood. *American Journal of Psychiatry*, *167*(11), 1357–1363. doi: 10.1176/appi.ajp.2010.10020223

Lipsitch, M., Tchetgen Tchetgen, E., & Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, *21*(3), 383–388. doi: 10.1097/EDE.0b013e3181d61eeb

Liu, C. H., Abrams, N. D., Carrick, D. M., Chander, P., Dwyer, J., Hamlet, M. R., … Vedamony, M. M. (2017). Biomarkers of chronic inflammation in disease development and prevention: challenges and opportunities. *Nature immunology*, *18*(11), 1175–1180. doi: 10.1038/ni.3828

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., … Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, *518*(7538), 197–206. doi: 10.1038/nature14177

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., … Price, A. L. (2016). Reference-based phasing using the haplotype reference consortium panel. *Nature genetics*, *48*(11), 1443. doi: 10.1038/ng.3679

Lokk, K., Modhukur, V., Rajashekar, B., Märtens, K., Mägi, R., Kolde, R., … Tõnisson, N. (2014). DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol.*, *15*(4), r54. doi: 10.1186/gb-2014-15-4-r54

Madley-Dowd, P., Rai, D., Zammit, S., & Heron, J. (2020). Simulations and directed acyclic graphs explained why assortative mating biases the prenatal negative control design. *Journal of Clinical Epidemiology*, *118*, 9–17. doi: 10.1016/j.jclinepi.2019.10.008

Mägi, R., Suleimanov, Y. V., Clarke, G. M., Kaakinen, M., Fischer, K., Prokopenko, I., & Morris, A. P. (2017). SCOPA and META-SCOPA: software for the analysis and aggregation of genome-wide association studies of multiple correlated phenotypes. *BMC bioinformatics*, *18*(1), 25. doi: 10.1186/s12859-016-1437-3

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, *456*(7218), 18–21. doi: 10.1038/456018a

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., & Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, *41*(6), 469-480. doi: 10.1002/gepi.22050

Manzari, N., Matvienko-Sikar, K., Baldoni, F., O'Keeffe, G. W., & Khashan, A. S. (2019). Prenatal maternal stress and risk of neurodevelopmental disorders in the offspring: a systematic review and meta-analysis. *Soc Psychiatry Psychiatr Epidemiol*, *54*(11), 1299–1309. doi: 10.1007/s00127-019-01745-3

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–511. doi: 10.1038/nrg2796

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, *39*(7), 906–13. doi: 10.1038/ng2088

Martins-Silva, T., Vaz, J. d. S., Hutz, M. H., Salatino-Oliveira, A., Genro, J. P., Hartwig, F. P., … Tovo-Rodrigues, L. (2019). Assessing causality in the association between attention-deficit/hyperactivity disorder and obesity: a Mendelian randomization study. *International Journal of Obesity*, *43*(12), 2500–2508. doi: 10.1038/s41366-019-0346-8

Matta, S. M., Hill-Yardin, E. L., & Crack, P. J. (2019). The influence of neuroinflammation in autism spectrum disorder. *Brain, Behavior, and Immunity*, *79*, 75–90. doi: 10.1016/j.bbi.2019.04.037

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*(5), 356–369. doi: 10.1038/nrg2344

McCaw, Z. R., Lane, J. M., Saxena, R., Redline, S., & Lin, X. (2019). Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics*. doi: 10.1111/biom.13214

McCullagh, P. (1989). *Generalized linear models* (2nd ed.). London: Chapman and Hall.

McQueen, M. J., Hawken, S., Wang, X., Ounpuu, S., Sniderman, A., Probstfield, J., … Yusuf, S. (2008). Lipids, lipoproteins, and apolipoproteins as risk markers of myocardial infarction in 52 countries (the INTERHEART study): a case-control study. *The Lancet*, *372*(9634), 224–233. doi: 10.1016/S0140-6736(08)61076-4

Menting, M. D., van de Beek, C., Mintjens, S., Wever, K. E., Korosi, A., Ozanne, S. E., … Painter, R. C. (2019). The link between maternal obesity and offspring neurobehavior: A systematic review of animal experiments. *Neuroscience & Biobehavioral Reviews*, *98*, 107–121. doi: 10.1016/j.neubiorev.2018.12.023

Miettunen, J., Haapea, M., Björnholm, L., Huhtaniska, S., Juola, T., Kinnunen, L., … Nordström, T. (2019). Psychiatric research in the Northern Finland Birth Cohort 1986 – a systematic review. *International Journal of Circumpolar Health*, *78*(1), 1571382. doi: 10.1080/22423982.2019.1571382

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *BMJ*, *339*. doi: 10.1136/bmj.b2535

Mooney, M. A., Ryabinin, P., Wilmot, B., Bhatt, P., Mill, J., & Nigg, J. T. (2020). Large epigenome-wide association study of childhood ADHD identifies peripheral DNA methylation associated with disease and polygenic risk burden. *Transl Psychiatry*, *10*(1), 8. doi: 10.1038/s41398-020-0710-4

Morris, T. T., Davies, N. M., Hemani, G., & Smith, G. D. (2020). Population phenomena inflate genetic associations of complex social traits. *Science Advances*, *6*(16). doi: 10.1126/sciadv.aay0328

Morton, N. E. (2006). Fifty years of genetic epidemiology, with special reference to Japan. *J. Hum. Genet.*, *51*(4), 269–277. doi: 10.1007/s10038-006-0366-9

Mowlem, F. D., Rosenqvist, M. A., Martin, J., Lichtenstein, P., Asherson, P., & Larsson, H. (2019). Sex differences in predicting ADHD clinical diagnosis and pharmacological treatment. *Eur Child Adolesc Psychiatry*, *28*(4), 481–489. doi: 10.1007/s00787-018-1211-3

Munafò, M. R. (2006). Candidate gene studies in the 21st century: meta-analysis, me-

diation, moderation. *Genes, Brain and Behavior*, *5*(S1), 3–8. doi: 10.1111/j.1601-183X.2006.00188.x

Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M., & Davey Smith, G. (2017). Collider scope: when selection bias can substantially influence observed associations. *International Journal of Epidemiology*, *47*(1), 226–235. doi: 10.1093/ije/dyx206

Muneer, A. (2016). Bipolar disorder: Role of inflammation and the development of disease biomarkers. *Psychiatry Investig*, *13*(1), 18–33. doi: 10.4306/pi.2016.13.1.18

Müller, N. (2019). The role of intercellular adhesion molecule-1 in the pathogenesis of psychiatric disorders. *Frontiers in Pharmacology*, *10*(1251). doi: 10.3389/fphar.2019.01251

Müller, N., Weidinger, E., Leitner, B., & Schwarz, M. J. (2015). The role of inflammation in schizophrenia. *Frontiers in Neuroscience*, *9*, 372. doi: 10.3389/fnins.2015.00372

Na, K.-S., Jung, H.-Y., & Kim, Y.-K. (2014). The role of pro-inflammatory cytokines in the neuroinflammation and neurogenesis of schizophrenia. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *48*, 277–286. doi: 10.1016/j.pnpbp.2012.10.022

Najjar, S., Pearlman, D. M., Alper, K., Najjar, A., & Devinsky, O. (2013). Neuroinflammation and psychiatric illness. *Journal of neuroinflammation*, *10*(1), 816. doi: 10.1186/1742-2094-10-43

Nathan, C., & Ding, A. (2010). Nonresolving inflammation. *Cell*, *140*(6), 871–882. doi: 10.1016/j.cell.2010.02.029

Ng, S. P., & Zelikoff, J. T. (2007). Smoking during pregnancy: subsequent effects on offspring immune competence and disease vulnerability in later life. *Reprod. Toxicol.*, *23*(3), 428–437. doi: 10.1016/j.reprotox.2006.11.008

Nielsen, C. H., Larsen, A., & Nielsen, A. L. (2016). DNA methylation alterations in response to prenatal exposure of maternal cigarette smoking: A persistent epigenetic impact on health from maternal lifestyle? *Archives of toxicology*, *90*(2), 231–245. doi: 10.1007/s00204-014-1426-0

Niemelä, S., Sourander, A., Surcel, H.-M., Hinkka-Yli-Salomäki, S., McKeague, I. W., Cheslack-Postava, K., & Brown, A. S. (2016). Prenatal Nicotine Exposure and Risk

of Schizophrenia Among Offspring in a National Birth Cohort. *American Journal of Psychiatry*, *173*(8), 799–806. doi: 10.1176/appi.ajp.2016.15060800

Northern Finland Cohorts. (2020). Northern Finland Cohorts. (`http://www.oulu.fi/nfbc` [Accessed: 2020-07-06])

Obel, C., Linnet, K. M., Henriksen, T. B., Rodriguez, A., Järvelin, M.-R., Kotimaa, A., … Olsen, J. (2008). Smoking during pregnancy and hyperactivity-inattention in the offspring—comparing results from three Nordic cohorts. *International Journal of Epidemiology*, *38*(3), 698–705. doi: 10.1093/ije/dym290

O'Brien, S. M., & Dunson, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics*, *60*(3), 739–746. doi: 10.1111/j.0006-341X.2004.00224.x

O'Donnell, M. J., Chin, S. L., Rangarajan, S., Xavier, D., Liu, L., Zhang, H., … Yusuf, S. (2016). Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. *The Lancet*, *388*(10046), 761–775. doi: 10.1016/S0140-6736(16)30506-2

O'Donnell, K. J., & Meaney, M. J. (2017). Fetal origins of mental health: The developmental origins of health and disease hypothesis. *American Journal of Psychiatry*, *174*(4), 319–328. doi: 10.1176/appi.ajp.2016.16020138

Paaby, A. B., & Rockman, M. V. (2013). The many faces of pleiotropy. *Trends in Genetics*, *29*(2), 66–73. doi: 10.1016/j.tig.2012.10.010

Parikh, N., & Boyd, S. (2014). Proximal algorithms. *Found. Trends Optim.*, *1*(3), 127–239. doi: 10.1561/2400000003

Parmar, P., Lowry, E., Cugliari, G., Suderman, M., Wilson, R., Karhunen, V., … Sebert, S. (2018). Association of maternal prenatal smoking GFI1-locus and cardio-metabolic phenotypes in 18,212 adults. *EBioMedicine*, *38*, 206–216. doi: 10.1016/j.ebiom.2018.10.066

Patti, G. J., Yanes, O., & Siuzdak, G. (2012). Innovation: Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.*, *13*(4), 263–269. doi: 10.1038/nrm3314

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688. doi: 10.1093/biomet/82.4.669

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (p. 411–420). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. (`https://arxiv.org/abs/1301.2300`)

Penninx, B. W., & Lange, S. M. (2018). Metabolic syndrome in psychiatric patients: overview, mechanisms, and implications. *Dialogues in clinical neuroscience*, *20*(1), 63. (`https://pubmed.ncbi.nlm.nih.gov/29946213/`)

Perkins, N. J., Cole, S. R., Harel, O., Tchetgen Tchetgen, E. J., Sun, B., Mitchell, E. M., & Schisterman, E. F. (2017). Principled Approaches to Missing Data in Epidemiologic Studies. *American Journal of Epidemiology*, *187*(3), 568-575. doi: 10.1093/aje/kwx348

Pingault, J. B., O'Reilly, P. F., Schoeler, T., Ploubidis, G. B., Rijsdijk, F., & Dudbridge, F. (2018). Using genetic data to strengthen causal inference in observational research. *Nat. Rev. Genet.*, *19*(9), 566–580. doi: 10.1038/s41576-018-0020-3

Pinheiro, J., & Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. New York, NY: Springer New York.

Polanczyk, G. V., de Lima, M. S., Horta, B. L., Biederman, J., & Rohde, L. A. (2007). The Worldwide Prevalence of ADHD: A Systematic Review and Metaregression Analysis. *American Journal of Psychiatry*, *164*(6), 942–948. doi: 10.1176/ajp.2007.164.6.942

Polanczyk, G. V., Willcutt, E. G., Salum, G. A., Kieling, C., & Rohde, L. A. (2014). ADHD prevalence estimates across three decades: an updated systematic review and meta-regression analysis. *Int J Epidemiol*, *43*(2), 434–442. doi: 10.1093/ije/dyt261

Porta, M. (2014). *A dictionary of epidemiology*. Oxford University Press.

Potter, A. S., Newhouse, P. A., & Bucci, D. J. (2006). Central nicotinic cholinergic systems: A role in the cognitive dysfunction in attention-deficit/hyperactivity disorder? *Behavioural Brain Research*, *175*(2), 201–211. doi: 10.1016/j.bbr.2006.09.015

Power, C., Atherton, K., & Thomas, C. (2010). Maternal smoking in pregnancy, adult adiposity and other risk factors for cardiovascular disease. *Atherosclerosis*, *211*(2), 643–648. doi: 10.1016/j.atherosclerosis.2010.03.015

Preacher, K. J. (2015). Advances in mediation analysis: a survey and synthesis of new developments. *Annu Rev Psychol*, *66*, 825–852. doi: 10.1146/annurev-psych-010814-015258

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, *38*(8), 904–909. doi: 10.1038/ng1847

Privé, F., Aschard, H., Ziyatdinov, A., & Blum, M. G. B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics*, *34*(16), 2781–2787. doi: 10.1093/bioinformatics/bty185

Ptacek, R., Kuzelova, H., Stefano, G. B., Raboch, J., Sadkova, T., Goetz, M., & Kream, R. M. (2014). Disruptive patterns of eating behaviors and associated lifestyles in males with adhd. *Medical science monitor : international medical journal of experimental and clinical research*, *20*, 608–613. doi: 10.12659/msm.890495

Purcell, S. M., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., … Sham, P. C. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, *81*(3), 559–575. doi: 10.1086/519795

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., … the International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748–752. doi: 10.1038/nature08185

Qi, T., Wu, Y., Zeng, J., Zhang, F., Xue, A., Jiang, L., … Zeng, B. (2018). Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat Commun*, *9*(1), 2282. doi: 10.1038/s41467-018-04558-1

Raitakari, O. T., Juonala, M., Rönnemaa, T., Keltikangas-Järvinen, L., Räsänen, L., Pietikäinen, M., … Viikari, J. S. (2008). Cohort Profile: The Cardiovascular Risk in Young Finns Study. *International Journal of Epidemiology*, *37*(6), 1220-1226. doi: 10.1093/ije/dym225

Rakyan, V. K., Down, T. A., Balding, D. J., & Beck, S. (2011). Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.*, *12*(8), 529–541. doi: 10.1038/nrg3000

Rantakallio, P. (1969). Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatrica Scandinavica*, *193*, Suppl–193. (`https://pubmed.ncbi.nlm.nih.gov/4911003/`)

Rebagliato, M. (2002). Validation of self reported smoking. *Journal of Epidemiology & Community Health*, *56*(3), 163–164. doi: 10.1136/jech.56.3.163

Reese, S. E., Zhao, S., Wu, M. C., Joubert, B. R., Parr, C. L., Håberg, S. E., … London, S. J. (2017). DNA Methylation Score as a Biomarker in Newborns for Sustained Maternal Smoking during Pregnancy. *Environmental Health Perspectives*, *125*(4), 760–766. doi: 10.1289/EHP333

Relton, C. L., & Davey Smith, G. (2010). Epigenetic epidemiology of common complex disease: prospects for prediction, prevention, and treatment. *PLoS Med.*, *7*(10), e1000356. doi: 10.1371/journal.pmed.1000356

Relton, C. L., Gaunt, T., McArdle, W., Ho, K., Duggirala, A., Shihab, H., … Davey Smith, G. (2015). Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *Int J Epidemiol*, *44*(4), 1181–1190. doi: 10.1093/ije/dyv072

Rice, F., Langley, K., Woodford, C., Davey Smith, G., & Thapar, A. (2018). Identifying the contribution of prenatal risk factors to offspring development and psychopathology: What designs to use and a critique of literature on maternal smoking and stress in pregnancy. *Development and Psychopathology*, *30*(3), 1107–1128. doi: 10.1017/S0954579418000421

Richardson, S., Tseng, G. C., & Sun, W. (2016). Statistical methods in integrative genomics. *Annual Review of Statistics and Its Application*, *3*(1), 181–209. doi: 10.1146/annurev-statistics-041715-033506

Richmond, R. C., Al-Amin, A., Davey Smith, G., & Relton, C. L. (2014). Approaches for drawing causal inferences from epidemiological birth cohorts: a review. *Early Hum. Dev.*, *90*(11), 769–780. doi: 10.1016/j.earlhumdev.2014.08.023

Richmond, R. C., Simpkin, A. J., Woodward, G., Gaunt, T. R., Lyttleton, O., McArdle, W. L., … Relton, C. L. (2014). Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Human Molecular Genetics*, *24*(8), 2201–2217. doi: 10.1093/hmg/ddu739

Richmond, R. C., Suderman, M., Langdon, R., Relton, C. L., & Davey Smith, G. (2018). DNA methylation as a marker for prenatal smoke exposure in adults. *Int J Epidemiol*, *47*(4), 1120–1130. doi: 10.1093/ije/dyy091

Rider, P., Carmi, Y., & Cohen, I. (2016). Biologics for targeting inflammatory cytokines, clinical uses, and limitations. *International journal of cell biology*, *2016*. doi: 10.1155/2016/9259646

Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., … van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, *368*. doi: 10.1136/bmj.m441

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell Jr, F. E., Moons, K. G., & Collins, G. S. (2019a). Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine*, *38*(7), 1276–1296. doi: 10.1002/sim.7992

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell Jr, F. E., Moons, K. G., & Collins, G. S. (2019b). Minimum sample size for developing a multivariable prediction model: Part I – Continuous outcomes. *Statistics in Medicine*, *38*(7), 1262–1275. doi: 10.1002/sim.7993

Ripke, S., Neale, B., Corvin, A., Walters, J., Farh, K., Holmans, P., … the Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*(7510), 421–427. doi: 10.1038/nature13595

Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, *273*(5281), 1516–1517. doi: 10.1126/science.273.5281.1516

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.*, *16*(2), 85–97. doi: 10.1038/nrg3868

Robertson, O. D., Coronado, N. G., Sethi, R., Berk, M., & Dodd, S. (2019). Putative neuroprotective pharmacotherapies to target the staged progression of mental illness. *Early intervention in psychiatry*, *13*(5), 1032–1049. doi: 10.1111/eip.12775

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, *3*(2), 143–155. doi: 10.1097/00001648-199203000-00013

Robinson, M. R., Kleinman, A., Graff, M., Vinkhuyzen, A. A. E., Couper, D., Miller, M. B., … Visscher, P. M. (2017). Genetic evidence of assortative mating in humans. *Nature Human Behaviour*, *1*(1), 0016. doi: 10.1038/s41562-016-0016

Robinson, S. L., Ghassabian, A., Sundaram, R., Trinh, M.-H., Lin, T.-C., Bell, E. M., & Yeung, E. (2020). Parental weight status and offspring behavioral problems and psychiatric symptoms. *The Journal of Pediatrics*, *220*, 227–236.e1. doi: 10.1016/j.jpeds.2020.01.016

Robinson, W. R., Renson, A., & Naimi, A. I. (2019). Teaching yourself about structural racism will improve your machine learning. *Biostatistics*, *21*(2), 339–344. doi: 10.1093/biostatistics/kxz040

Rodosthenous, T., Shahrezaei, V., & Evangelou, M. (2020). Integrating multi-OMICS data through sparse Canonical Correlation Analysis for the prediction of complex traits: A comparison study. *Bioinformatics*. doi: 10.1093/bioinformatics/btaa530

Rodriguez, A. (2010). Maternal pre-pregnancy obesity and risk for inattention and negative emotionality in children. *Journal of Child Psychology and Psychiatry*, *51*(2), 134-143. doi: 10.1111/j.1469-7610.2009.02133.x

Rodriguez, A., Järvelin, M.-R., Obel, C., Taanila, A., Miettunen, J., Moilanen, I., … Olsen, J. (2007). Do inattention and hyperactivity symptoms equal scholastic impairment? evidence from three european cohorts. *BMC Public Health*, *7*(1), 327. doi: 10.1186/1471-2458-7-327

Rodriguez, A., Miettunen, J., Henriksen, T. B., Olsen, J., Obel, C., Taanila, A., … Järvelin, M.-R. (2008). Maternal adiposity prior to pregnancy is associated with ADHD symptoms in offspring: evidence from three prospective pregnancy cohorts. *Int J Obes (Lond)*, *32*(3), 550–7. doi: 10.1038/sj.ijo.0803741

Rommelse, N. N., Franke, B., Geurts, H. M., Hartman, C. A., & Buitelaar, J. K. (2010). Shared heritability of attention-deficit/hyperactivity disorder and autism spectrum disorder. *Eur Child Adolesc Psychiatry*, *19*(3), 281–295. doi: 10.1007/s00787-010-0092-x

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. doi: 10.1093/biomet/70.1.41

Rothman, K. J. (2008). *Modern epidemiology* (3rd ed.). Philadelphia, Pa. ; London: Lippincott Williams & Wilkins.

Rovira, P., Sánchez-Mora, C., Pagerols, M., Richarte, V., Corrales, M., Fadeuilhe, C., … Ribasés, M. (2020). Epigenome-wide association study of attention-deficit/hyperactivity disorder in adults. *Translational psychiatry*, *10*(1), 1–12. doi: 10.1038/s41398-020-0860-4

Royston, P., & Parmar, M. K. B. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, *21*(15), 2175–2197. doi: 10.1002/sim.1203

Rucklidge, J. J. (2010). Gender differences in attention-deficit/hyperactivity disorder. *Psychiatric Clinics of North America*, *33*(2), 357–373. doi: 10.1016/j.psc.2010.01.006

Russell, A. E., Ford, T., Williams, R., & Russell, G. (2016). The Association Between Socioeconomic Disadvantage and Attention Deficit/Hyperactivity Disorder (ADHD): A Systematic Review. *Child Psychiatry Hum Dev*, *47*(3), 440–458. doi: 10.1007/s10578-015-0578-3

Rutter, M. (1967). A children's behaviour questionnaire for completion by teachers: preliminary findings. *J Child Psychol Psychiatry*, *8*(1), 1–11. doi: 10.1111/j.1469-7610.1967.tb02175.x

Sabatti, C., Service, S. K., Hartikainen, A.-L., Pouta, A., Ripatti, S., Brodsky, J., … Peltonen, L. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics*, *41*(1), 35–46. doi: 10.1038/ng.271

Sanchez, C. E., Barry, C., Sabhlok, A., Russell, K., Majors, A., Kollins, S. H., & Fuemmeler, B. F. (2018). Maternal pre-pregnancy obesity and child neurodevelopmental outcomes: a meta-analysis. *Obesity Reviews*, *19*(4), 464-484. doi: 10.1111/obr.12643

Sanderson, E., Davey Smith, G., Bowden, J., & Munafò, M. R. (2019). Mendelian randomisation analysis of the effect of educational attainment and cognitive ability on smoking behaviour. *Nature Communications*, *10*(1), 2949. doi: 10.1038/s41467-019-10679-y

Sanderson, E., Macdonald-Wallis, C., & Davey Smith, G. (2017). Negative control exposure studies in the presence of measurement error: implications for attempted effect estimate calibration. *International Journal of Epidemiology*, *47*(2), 587–596. doi: 10.1093/ije/dyx213

Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M., & Esteller, M. (2011). Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*, *6*(6), 692–702. doi: 10.4161/epi.6.6.16196

Santalahti, K., Havulinna, A., Maksimow, M., Zeller, T., Blankenberg, S., Vehtari, A., … Salmi, M. (2017). Plasma levels of hepatocyte growth factor and placental growth factor predict mortality in a general population: a prospective cohort study. *Journal of Internal Medicine*, *282*(4), 340–352. doi: 10.1111/joim.12648

Santalahti, K., Maksimow, M., Airola, A., Pahikkala, T., Hutri-Kähönen, N., Jalkanen, S., … Salmi, M. (2016). Circulating cytokines predict the development of insulin resistance in a prospective finnish population cohort. *The Journal of Clinical Endocrinology & Metabolism*, *101*(9), 3361–3369. doi: 10.1210/jc.2016-2081

Sauerbrei, W., Perperoglou, A., Schmid, M., Abrahamowicz, M., Becher, H., Binder, H., … for TG2 of the STRATOS initiative (2020). State of the art in selection of variables and functional forms in multivariable analysis–outstanding issues. *Diagnostic and prognostic research*, *4*, 1–18. doi: 10.1186/s41512-020-00074-3

Saukkonen, T., Mutt, S. J., Jokelainen, J., Saukkonen, A. M., Raza, G. S., Karhu, T., … Keinanen-Kiukaanniemi, S. (2018). Adipokines and inflammatory markers in elderly subjects with high risk of type 2 diabetes and cardiovascular disease. *Sci Rep*, *8*(1), 12816. doi: 10.1038/s41598-018-31144-8

Savva, S. C., Tornaritis, M., Savva, M. E., Kourides, Y., Panagi, A., Silikiotou, N., … Kafatos, A. (2000). Waist circumference and waist-to-height ratio are better predictors of cardiovascular disease risk factors in children than body mass index. *Int. J. Obes. Relat. Metab. Disord.*, *24*(11), 1453–1458. doi: 10.1038/sj.ijo.0801401

Sayal, K., Prasad, V., Daley, D., Ford, T., & Coghill, D. (2018). ADHD in children and young people: prevalence, care pathways, and service provision. *The Lancet Psychiatry*, *5*(2), 175–186. doi: 10.1016/S2215-0366(17)30167-0

Schmidt, A. F., Finan, C., Gordillo-Marañón, M., Asselbergs, F. W., Freitag, D. F., Patel, R. S., … Hingorani, A. D. (2020). Genetic drug target validation using Mendelian randomisation. *Nature communications*, *11*(1), 1–12. doi: 10.1038/s41467-020-16969-0

Serati, M., Barkin, J. L., Orsenigo, G., Altamura, A. C., & Buoli, M. (2017). Research review: The role of obstetric and neonatal complications in childhood attention deficit and hyperactivity disorder – a systematic review. *Journal of Child Psychology and Psychiatry*, *58*(12), 1290–1300. doi: 10.1111/jcpp.12779

Shi, W. J., Zhuang, Y., Russell, P. H., Hobbs, B. D., Parker, M. M., Castaldi, P. J., … Kechris, K. (2019). Unsupervised discovery of phenotype-specific multi-omics networks. *Bioinformatics*, *35*(21), 4336–4343. doi: 10.1093/bioinformatics/btz226

Shmueli, G. (2010). To explain or to predict? *Statist. Sci.*, *25*(3), 289–310. doi: 10.1214/10-STS330

Shungin, D., Winkler, T. W., Croteau-Chonka, D. C., Ferreira, T., Locke, A. E., Mägi, R., … Mohlke, K. L. (2015). New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, *518*, 187. doi: 10.1038/nature14132

Sibley, M. H., Swanson, J. M., Arnold, L. E., Hechtman, L. T., Owens, E. B., Stehli, A., … the MTA Cooperative Group (2017). Defining ADHD symptom persistence in adulthood: optimizing sensitivity and specificity. *Journal of Child Psychology and Psychiatry*, *58*(6), 655–662. doi: 10.1111/jcpp.12620

Sikdar, S., Joehanes, R., Joubert, B. R., Xu, C.-J., Vives-Usano, M., Rezwan, F. I., … London, S. J. (2019). Comparison of smoking-related DNA methylation between newborns from prenatal exposure and adults from personal smoking. *Epigenomics*, *11*(13), 1487–1500. doi: 10.2217/epi-2019-0066

Simon, V., Czobor, P., Bálint, S., Mészáros, ., & Bitter, I. (2009). Prevalence and correlates of adult attention-deficit hyperactivity disorder: meta-analysis. *British Journal of Psychiatry*, *194*(3), 204–211. doi: 10.1192/bjp.bp.107.048827

Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell*, *177*(1), 26–31. doi: 10.1016/j.cell.2019.02.048

Slatkin, M. (2008). Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.*, *9*(6), 477–485. doi: 10.1038/nrg2361

Sliz, E., Kalaoja, M., Ahola-Olli, A., Raitakari, O., Perola, M., Salomaa, V., … Kettunen, J. (2019). Genome-wide association study identifies seven novel loci associating with

circulating cytokines and cell adhesion molecules in Finns. *Journal of Medical Genetics*, *56*(9), 607–616. doi: 10.1136/jmedgenet-2018-105965

Slob, E. A. W., & Burgess, S. (2020). A comparison of robust Mendelian randomization methods using summary data. *Genetic Epidemiology*, *44*(4), 313–329. doi: 10.1002/gepi.22295

Smalley, S. L., McGough, J. J., Moilanen, I. K., Loo, S. K., Taanila, A., Ebeling, H., … Järvelin, M.-R. (2007). Prevalence and Psychiatric Comorbidity of Attention-Deficit/Hyperactivity Disorder in an Adolescent Finnish Population. *Journal of the American Academy of Child & Adolescent Psychiatry*, *46*(12), 1575–1583. doi: 10.1097/chi.0b013e3181573137

Soininen, P., Kangas, A. J., Würtz, P., Suna, T., & Ala-Korpela, M. (2015). Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circulation: Cardiovascular Genetics*, *8*(1), 192–206. doi: 10.1161/CIRCGENETICS.114.000216

Sokolova, E., Groot, P., Claassen, T., van Hulzen, K. J., Glennon, J. C., Franke, B., … Buitelaar, J. (2016). Statistical evidence suggests that inattention drives hyperactivity/impulsivity in attention deficit-hyperactivity disorder. *PLOS ONE*, *11*(10), 1-17. doi: 10.1371/journal.pone.0165120

Sonuga-Barke, E. J., Brandeis, D., Cortese, S., Daley, D., Ferrin, M., Holtmann, M., … Sergeant, J. a. (2013). Nonpharmacological Interventions for ADHD: Systematic Review and Meta-Analyses of Randomized Controlled Trials of Dietary and Psychological Treatments. *American Journal of Psychiatry*, *170*(3), 275–289. doi: 10.1176/appi.ajp.2012.12070991

Sourander, A., Sucksdorff, M., Chudal, R., Surcel, H. M., Hinkka-Yli-Salomäki, S., Gyllenberg, D., … Brown, A. S. (2019). Prenatal Cotinine Levels and ADHD Among Offspring. *Pediatrics*, *143*(3). doi: 10.1542/peds.2018-3144

Sousa, N. O., Grevet, E. H., Salgado, C. A., Silva, K. L., Victor, M. M., Karam, R. G., … Bau, C. H. (2011). Smoking and adhd: An evaluation of self medication and behavioral disinhibition models based on comorbidity and personality patterns. *Journal of Psychiatric Research*, *45*(6), 829–834. doi: 10.1016/j.jpsychires.2010.10.012

Sovio, U., Bennett, A. J., Millwood, I. Y., Molitor, J., O'Reilly, P. F., Timpson, N. J., … Järvelin, M.-R. (2009). Genetic Determinants of Height Growth Assessed Longitudinally from Infancy to Adulthood in the Northern Finland Birth Cohort 1966. *PLoS Genet*, *5*(3), e1000409. doi: 10.1371/journal.pgen.1000409

Stahl, E. A., Breen, G., Forstner, A. J., McQuillin, A., Ripke, S., Trubetskoy, V., … the Bipolar Disorder Working Group of the Psychiatric Genomics Consortium (2019). Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature genetics*, *51*(5), 793–803. doi: 10.1038/s41588-019-0397-8

Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating*. Cham: Springer International Publishing.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., … Collins, R. (2015). UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, *12*(3), e1001779. doi: 10.1371/journal.pmed.1001779

Sullivan, E. L., Riper, K. M., Lockard, R., & Valleau, J. C. (2015). Maternal high-fat diet programming of the neuroendocrine system and behavior. *Hormones and Behavior*, *76*, 153–161. doi: 10.1016/j.yhbeh.2015.04.008

Sullivan, P. F., Agrawal, A., Bulik, C. M., Andreassen, O. A., Børglum, A. D., Breen, G., … the Psychiatric Genomics Consortium (2018). Psychiatric genomics: An update and an agenda. *American Journal of Psychiatry*, *175*(1), 15–27. doi: 10.1176/appi.ajp.2017.17030283

Sun, Y. V., & Hu, Y. J. (2016). Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases. *Adv. Genet.*, *93*, 147–190. doi: 10.1016/bs.adgen.2015.11.004

Suo, X., Minden, V., Nelson, B., Tibshirani, R., & Saunders, M. (2017). *Sparse canonical correlation analysis.* (`https://arxiv.org/abs/1705.10865v2`)

Swanson, J. M., Schuck, S., Porter, M. M., Carlson, C., Hartman, C. A., Sergeant, J. A., … Wigal, T. (2012). Categorical and dimensional definitions and evaluations of symptoms of ADHD: History of the SNAP and the SWAN rating scales. *Int J Educ Psychol Assess*, *10*(1), 51–70. (`https://europepmc.org/articles/PMC4618695`)

Swerdlow, D. I., Kuchenbaecker, K. B., Shah, S., Sofat, R., Holmes, M. V., White, J., … Hingorani, A. D. (2016). Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *Int J Epidemiol*, *45*(5), 1600–1616. doi: 10.1093/ije/dyw088

Tabor, H. K., Risch, N. J., & Myers, R. M. (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews Genetics*, *3*(5), 391–397. doi: 10.1038/nrg796

Talati, A., Bao, Y., Kaufman, J., Shen, L., Schaefer, C. A., & Brown, A. S. (2013). Maternal smoking during pregnancy and bipolar disorder in offspring. *American Journal of Psychiatry*, *170*(10), 1178–1185. doi: 10.1176/appi.ajp.2013.12121500

Taylor, A. E., Davey Smith, G., Bares, C. B., Edwards, A. C., & Munafò, M. R. (2014). Partner smoking and maternal cotinine during pregnancy: Implications for negative control methods. *Drug and Alcohol Dependence*, *139*, 159–163. doi: 10.1016/j.drugalcdep.2014.03.012

Taylor, A. E., Howe, L. D., Heron, J. E., Ware, J. J., Hickman, M., & Munafò, M. R. (2014). Maternal smoking during pregnancy and offspring smoking initiation: assessing the role of intrauterine exposure. *Addiction*, *109*(6), 1013–1021. doi: 10.1111/add.12514

Taylor, G. M. J., & Munafò, M. R. (2019). Does smoking cause poor mental health? *The Lancet Psychiatry*, *6*(1), 2–3. doi: 10.1016/S2215-0366(18)30459-0

Tehranifar, P., Wu, H.-C., McDonald, J. A., Jasmine, F., Santella, R. M., Gurvich, I., … Terry, M. B. (2018). Maternal cigarette smoking during pregnancy and offspring dna methylation in midlife. *Epigenetics*, *13*(2), 129–134. doi: 10.1080/15592294.2017.1325065

Teixeira-Pinto, A., Siddique, J., Gibbons, R., & Normand, S. L. (2009). Statistical approaches to modeling multiple outcomes in psychiatric studies. *Psychiatric annals*, *39*(7), 729–735. doi: 10.3928/00485713-20090625-08

Teschendorff, A. E., & Relton, C. L. (2018). Statistical and integrative system-level analysis of DNA methylation data. *Nat. Rev. Genet.*, *19*(3), 129–147. doi: 10.1038/nrg.2017.86

Thapar, A., & Cooper, M. (2016). Attention deficit hyperactivity disorder. *The Lancet*, *387*(10024), 1240–1250. doi: 10.1016/S0140-6736(15)00238-X

Thapar, A., & Rice, F. (2020). Family-Based Designs that Disentangle Inherited Factors from Pre- and Postnatal Environmental Exposures: In Vitro Fertilization, Discordant Sibling Pairs, Maternal versus Paternal Comparisons, and Adoption Designs. *Cold Spring Harb Perspect Med*. doi: 10.1101/cshperspect.a038877

the Neale Lab. (2018). Updated GWAS Analysis of the UK Biobank. (`http://www.nealelab.is/uk-biobank` [Accessed: 2019-05-16])

Thomas, D. C., & Conti, D. V. (2004). Commentary: The concept of 'Mendelian Randomization'. *International Journal of Epidemiology*, *33*(1), 21–25. doi: 10.1093/ije/dyh048

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *58*(1), 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, *59*(5), 1–38. doi: 10.18637/jss.v059.i05

Treur, J. L., Demontis, D., Smith, G. D., Sallis, H., Richardson, T. G., Wiers, R. W., … Munafò, M. R. (2019). Investigating causality between liability to ADHD and substance use, and liability to substance use and ADHD risk, using Mendelian randomization. *Addiction biology*, e12849. doi: 10.1111/adb.12849

van Dongen, J., Zilhão, N. R., Sugden, K., Heijmans, B. T., 't Hoen, P. A., van Meurs, J., … Boomsma, D. I. (2019). Epigenome-wide Association Study of Attention-Deficit/Hyperactivity Disorder Symptoms in Adults. *Biological Psychiatry*, *86*(8), 599–607. doi: 10.1016/j.biopsych.2019.02.016

van Amsterdam, J., van der Velde, B., Schulte, M., & van den Brink, W. (2018). Causal Factors of Increased Smoking in ADHD: A Systematic Review. *Substance Use & Misuse*, *53*(3), 432–445. doi: 10.1080/10826084.2017.1334066

VanderWeele, T. J. (2016). Mediation Analysis: A Practitioner's Guide. *Annu Rev Public Health*, *37*, 17–32. doi: 10.1146/annurev-publhealth-032315-021402

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, *34*(3), 211–219. doi: 10.1007/s10654-019-00494-6

VanderWeele, T. J., Tchetgen Tchetgen, E. J., Cornelis, M., & Kraft, P. (2014). Methodological Challenges in Mendelian Randomization. *Epidemiology*, *25*(3), 427–435. doi: 10.1097/EDE.0000000000000081

van Iterson, M., van Zwet, E. W., & Heijmans, B. T. (2017). Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol.*, *18*(1), 19. doi: 10.1186/s13059-016-1131-9

Vartiainen, E., Seppälä, T., Lillsunde, P., & Puska, P. (2002). Validation of self reported smoking by serum cotinine measurement in a community-based study. *Journal of Epidemiology & Community Health*, *56*(3), 167–170. doi: 10.1136/jech.56.3.167

Verbanck, M., Chen, C. Y., Neale, B., & Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.*, *50*(5), 693–698. doi: 10.1038/s41588-018-0099-7

Vermeulen, J., Schirmbeck, F., Blankers, M., van Tricht, M., van den Brink, W., de Haan, L., … van Winkel, R. (2019). Smoking, symptoms, and quality of life in patients with psychosis, siblings, and healthy controls: a prospective, longitudinal cohort study. *The Lancet Psychiatry*, *6*(1), 25–34. doi: 10.1016/S2215-0366(18)30424-3

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. doi: 10.18637/jss.v036.i03

Vigo, D., Thornicroft, G., & Atun, R. (2016). Estimating the true global burden of mental illness. *The Lancet Psychiatry*, *3*(2), 171–178. doi: 10.1016/S2215-0366(15)00505-2

Vilhjálmsson, B., Yang, J., Finucane, H., Gusev, A., Lindström, S., Ripke, S., … Price, A. L. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*, *97*(4), 576–592. doi: 10.1016/j.ajhg.2015.09.001

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.*, *101*(1), 5–22. doi: 10.1016/j.ajhg.2017.06.005

Visscher, P. M., & Yang, J. (2016). A plethora of pleiotropy across complex traits. *Nat. Genet.*, *48*(7), 707–708. doi: 10.1038/ng.3604

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). Strengthening the reporting of observational studies in epidemiology (strobe) statement: guidelines for reporting observational studies. *BMJ*, *335*(7624), 806–808. doi: 10.1136/bmj.39335.541782.AD

Vrijheid, M. (2014). The exposome: a new paradigm to study the impact of environment on health. *Thorax*, *69*(9), 876–878. doi: 10.1136/thoraxjnl-2013-204949

Wahl, S., Drong, A., Lehne, B., Loh, M., Scott, W. R., Kunze, S., … Chambers, J. C. (2017). Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature*, *541*(7635), 81–86. doi: 10.1038/nature20784

Wahl, S., Wong, H., & McCartney-Francis, N. (1989). Role of growth factors in inflammation and repair. *Journal of cellular biochemistry*, *40*(2), 193–199. doi: 10.1002/jcb.240400208

Walton, E., Hass, J., Liu, J., Roffman, J. L., Bernardoni, F., Roessner, V., … Ehrlich, S. (2015). Correspondence of DNA Methylation Between Blood and Brain Tissue and Its Application to Schizophrenia Research. *Schizophrenia Bulletin*, *42*(2), 406–414. doi: 10.1093/schbul/sbv074

Walton, E., Pingault, J. B., Cecil, C. A., Gaunt, T. R., Relton, C. L., Mill, J., & Barker, E. D. (2017). Epigenetic profiling of ADHD symptoms trajectories: a prospective, methylome-wide study. *Mol. Psychiatry*, *22*(2), 250–256. doi: 10.1038/mp.2016.85

Wang, Q., Würtz, P., Auro, K., Morin-Papunen, L., Kangas, A. J., Soininen, P., … Ala-Korpela, M. (2016). Effects of hormonal contraception on systemic metabolism: cross-sectional and longitudinal evidence. *International Journal of Epidemiology*, *45*(5), 1445–1457. doi: 10.1093/ije/dyw147

Wang, W. Y. S., Barratt, B. J., Clayton, D. G., & Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, *6*(2), 109–118. doi: 10.1038/nrg1522

Wang, Y. R., Jiang, K., Feldman, L. J., Bickel, P. J., & Huang, H. (2015). Inferring gene–gene interactions and functional modules using sparse canonical correlation analysis. *The Annals of Applied Statistics*, *9*(1), 300–323. doi: 10.1214/14-AOAS792

Westreich, D., & Greenland, S. (2013). The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients. *American Journal of Epidemiology*, *177*(4), 292–298. doi: 10.1093/aje/kws412

Wiklund, P., Karhunen, V., Richmond, R. C., Parmar, P., Rodriguez, A., De Silva, M., … Järvelin, M.-R. (2019). DNA methylation links prenatal smoking exposure to later life health outcomes in offspring. *Clin Epigenetics*, *11*(1), 97. doi: 10.1186/s13148-019-0683-4

Wild, C. P. (2005). Complementing the genome with an exposome: the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol. Biomarkers Prev.*, *14*(8), 1847–1850. doi: 10.1158/1055-9965.EPI-05-0456

Wilens, T. E., Adamson, J., Sgambati, S., Whitley, J., Santry, A., Monuteaux, M. C., & Biederman, J. (2007). Do individuals with adhd self-medicate with cigarettes and substances of abuse? results from a controlled family study of adhd. *The American Journal on Addictions*, *16*(s1), 14–23. doi: 10.1080/10550490601082742

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, *26*(17), 2190–1. doi: 10.1093/bioinformatics/btq340

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, *116*(4), 1195–1200. doi: 10.1073/pnas.1814092116

Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, *10*(3), 515–534. doi: 10.1093/biostatistics/kxp008

Wood, S. N. (2006). *Generalized additive models : an introduction with R*. Boca Raton, FL: Chapman & Hall/CRC.

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., … the Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature genetics*, *50*(5), 668–681. doi: 10.1038/s41588-018-0090-3

Wynants, L., van Smeden, M., McLernon, D. J., Timmerman, D., Steyerberg, E. W., Van Calster, B., & the STRATOS initiative. (2019). Three myths about risk thresholds for prediction models. *BMC medicine*, *17*(1), 192. doi: 10.1186/s12916-019-1425-3

Würtz, P., Wang, Q., Kangas, A. J., Richmond, R. C., Skarp, J., Tiainen, M., … Ala-Korpela, M. (2014). Metabolic Signatures of Adiposity in Young Adults: Mendelian Randomization Analysis and Effects of Weight Change. *PLOS Medicine*, *11*(12), 1–18. doi: 10.1371/journal.pmed.1001765

Xu, N., Li, X., & Zhong, Y. (2015). Inflammatory cytokines: potential biomarkers of immunologic dysfunction in autism spectrum disorders. *Mediators of inflammation*, *2015*. doi: 10.1155/2015/531518

Yan, J., Risacher, S. L., Shen, L., & Saykin, A. J. (2017). Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in Bioinformatics*, *19*(6), 1370–1381. doi: 10.1093/bib/bbx066

Ye, Y., Zhang, Z., Liu, Y., Diao, L., & Han, L. (2020). A multi-omics perspective of quantitative trait loci in precision medicine. *Trends in Genetics*, *36*(5), 318–336. doi: 10.1016/j.tig.2020.01.009

Yengo, L., Robinson, M. R., Keller, M. C., Kemper, K. E., Yang, Y., Trzaskowski, M., … Visscher, P. M. (2018). Imprint of assortative mating on the human genome. *Nature Human Behaviour*, *2*(12), 948–954. doi: 10.1038/s41562-018-0476-3

Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., … the GIANT Consortium (2018). Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of European ancestry. *Human Molecular Genetics*, *27*(20), 3641–3649. doi: 10.1093/hmg/ddy271

Yuan, N., Chen, Y., Xia, Y., Dai, J., & Liu, C. (2019). Inflammation-related biomarkers in major psychiatric disorders: a cross-disorder assessment of reproducibility and specificity in 43 meta-analyses. *Translational psychiatry*, *9*(1), 1–13. doi: 10.1038/s41398-019-0570-y

Zhu, J., Zhang, X., Xu, Y., Spencer, T. J., Biederman, J., & Bhide, P. G. (2012). Prenatal nicotine exposure mouse model showing hyperactivity, reduced cingulate cortex volume,

reduced dopamine turnover, and responsiveness to oral methylphenidate treatment. *Journal of Neuroscience*, *32*(27), 9410–9418. doi: 10.1523/JNEUROSCI.1041-12.2012

Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., & Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, *50*, 71–91. doi: 10.1016/j.inffus.2018.09.012

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. doi: 10.1111/j.1467-9868.2005.00503.x